NORTHWESTERN UNIVERSITY

Essays on "Small" Sample Problems in "Large" Datasets

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Yong Cai

EVANSTON, ILLINOIS

June 2023

ABSTRACT

Essays on "Small" Sample Problems in "Large" Datasets

Yong Cai

Many estimation and inference procedures rely on asymptotic approximations for quantities that are unknown to researchers. While often convenient, such approximations can be poor in practice, even when the number of observations is ostensibly large. One response is to eschew asymptotics in favor of finite sample bounds. While remarkable progress has been made in this regard, bounds are often wide, or involve unknown parameters that limit their use. This dissertation takes a different approach. Our view is that the failure of asymptotics can often be attributed to certain pathological features of the data that reduce effective sample size, so that data sets may be small for the purpose of asymptotic approximation, even when they are nominally large. Our solution is to develop alternative asymptotic theories that explicitly incorporate said features, so that "small" data problems persist in the limiting approximations, which we expect to be more accurate as a result. We pursue such an approach in three different settings.

Chapter 1 studies the properties of linear regression on centrality measures when network data is sparse – that is, when there are many more agents than links per agent – and

when they are constructed by proxy. Network data contains little information when they are sparse, since the adjacency matrices are mostly zeroes. Conventional analyses, based on taking the number of nodes to infinity, ignore the fact that centrality measures may have no variation when networks are sparse. Instead, we study the theoretical properties of OLS under sequences of increasingly sparse networks, making three contributions: (1) We show that OLS estimators can become inconsistent under sparsity and characterize the threshold at which this occurs, with and without proxy error. This threshold depends on the centrality measure used. Specifically, regression on eigenvector is less robust to sparsity than on degree and diffusion. (2) We develop distributional theory for OLS estimators under proxy error and sparsity, finding that OLS estimators are subject to asymptotic bias even when they are consistent. Moreover, bias can be large relative to the variances, so that bias correction is necessary for inference. (3) We propose novel bias correction and inference methods for OLS with sparse proxy networks. Simulation evidence suggests that our theory and methods perform well, particularly in settings where the usual OLS estimators and heteroskedasticity-consistent/robust t-tests are deficient. Finally, we demonstrate the utility of our results in an application inspired by De Weerdt and Dercon (2006), in which we study the relationship between consumption smoothing and informal insurance in Nyakatoke, Tanzania.

Chapters 2 and 3 consider the issue of inference with cluster-dependent data. When researchers are concerned about dependence between observations in their datasets, they typically group observations into independent clusters in order to facilitate inference using approximate randomization tests (ART) or tests based on the clustered-covariance estimator (CCE). Because researchers are often willing to make only minimal assumptions

about the dependence structure within each cluster, cluster-dependent methods typically have effective sample size equal to the number of clusters, which is low in many empirical settings, even if the total number of observations is a large. To better understand the challenges posed by few clusters, Chapters 2 and 3 study issues in inference with cluster-dependent data in asymptotic frameworks in which the number of clusters are finite in the limit.

Chapter 2 proposes a test for the level of clustering. CCE and ART require the cluster structure of the data to be known ex ante. However, researchers often have some choice in clustering their data. As such, a researcher who has chosen to cluster their data at a finer or more disaggregated level may be unsure about their decision, especially given knowledge that observations are independent when clustered at a coarser, or more aggregated level. Chapter 2 proposes a modified randomization test as a robustness check for the chosen level of clustering in a linear regression setting. Existing tests require either the number of coarse clusters or number of fine clusters to be large. Our method is designed for settings with few coarse and fine clusters. While the method is conservative, it has competitive power in settings that may be relevant to empirical work.

Chapter 3 (joint with Ivan A. Canay, Deborah Kim and Azeem M. Shaikh) considers issues in the implementation of approximate randomization tests, an inference method explicitly designed for settings with few clusters. We show that the ARTs admit an equivalent implementation based on weighted scores and that the test and confidence intervals are invariant to whether the test statistic is studentized or not. When the test involves scalar parameters, we prove that the confidence intervals formed via test inversion are convex. We also present a novel, exact algorithm for test inversion with

scalar parameters, which reliably outperforms grid and bisection search. This chapter is written as a user's guide: we articulate the main requirements underlying the test, emphasizing in particular common pitfalls that researchers may encounter and provide two empirical demonstrations based on Munyo and Rossi (2015) and Meng et al. (2015).

Finally, Chapter 4 (joint with Ahnaf Rafi) considers the issue of experiment design with the Neyman Allocation, which is used in many papers on experiment design. These papers typically assume that researchers have access to large pilot studies, which may not be realistic. To understand the properties of the Neyman Allocation with small pilots, we study its behavior in a novel asymptotic framework for two-wave experiments in which the pilot size is assumed to be fixed even as the main wave sample size grows. Our analysis shows that the Neyman Allocation can lead to estimates of the ATE with higher asymptotic variance than with (non-adaptive) balanced randomization. That is, even with a large main-wave experiment, the reduction in asymptotic variance that results from the Neyman Allocation depends on the size of the pilot study used for its estimation. We find that the method performs especially poorly compared to balanced randomization when the outcome variable is relatively homoskedastic with respect to treatment status or when it exhibits high kurtosis. We also provide a series of empirical examples showing that these situations arise frequently in practice. Our results therefore suggest that researchers should not use the Neyman Allocation with small pilots, especially in such instances.

Acknowledgements

The chance to work under the guidance of my committee has been an enormous gift. Ivan Canay has been a dedicated mentor and has indelibly shaped the way I think about econometrics. This dissertation is organized around an approach to asymptotic theory that I came to appreciate only through working with him. Eric Auerbach introduced me to the wondrous field of network econometrics and has been unreasonably generous with his time. Joel Horowitz has been an invaluable font of wisdom; nothing escapes his eagle eyes.

I have benefited immensely from conversations with Deborah Kim, Ahnaf Rafi and the other econometrics students at Northwestern. My PhD journey would not be the same without the friendship of many colourful classmates, especially Santiago Camara and Jose Flor-Toro. Grant Goehring provided tireless support through the most trying of times.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Linear Regression with Centrality Measures

1.1. Introduction

A large and rapidly growing body of work documents the influence of networks in a wide range of economic outcomes: peer effects drive academic achievement, production networks shape shock propagation in the macroeconomy, social networks affect information- and risk-sharing with important implications for development (see Sacerdote 2011, Carvalho and Tahbaz-Salehi 2019 and Breza et al. 2019 for recent reviews). Many other examples abound.

One particular strand of research has explored the relationship between an agent's network position and their economic outcomes. For example, Hochberg et al. (2007) considers the network of venture capital firms and finds that better-networked firms successfully exit a greater proportion of their investments. Meanwhile, Cruz et al. (2017) examines the social networks in the Philippines and shows that more central families are disproportionately represented in political offices. Similarly, Banerjee et al. (2013) studies the problem of diffusing microfinance in India and establishes that seeding information to more central agents led to greater participation in the program.

In these papers, researchers often estimate linear models by ordinary least squares (OLS), using centrality measures as explanatory variables. Centrality measures are node-level statistics that capture notions of importance in a network. Since nodes can be

important for many reasons, a variety of centrality measures exist, each capturing a particular aspect of network position. For example, the degree centrality of an agent reflects the number or intensity of their direct links, while eigenvector centrality is designed so that the influence of agents is proportional to that of their connections. The correlation between an outcome variable and a particular centrality measure may be revealing about the types of interactions that drive a given economic phenomenon: an outcome that is well-predicted solely by degree is likely to be determined in an extremely local manner, whereas one that is more strongly associated with eigenvector centrality may involve non-linear interactions between agents that are further apart. As such, when researchers estimate these correlations and test their statistical significance, they frequently do so with the goal of drawing conclusions about the economic significance of various centrality measures and the implied mechanisms for outcome determination. Such an exercise is credible only if the OLS estimator is close to the estimand, and if the chosen test statistic (typically the heteroskedasticity-consistent/robust t-statistics) is well described by its asymptotic distribution (standard normal for t-statistics) in finite sample.

However, network data have two features that may threaten the statistical validity of OLS. First, networks may be sparse, with many more agents than links per agent. This could happen because interactions are observed with low frequency, or because the interactions in question are rare. Chandrasekhar (2016) argues that many economic networks are sparse, providing evidence from commonly used social network data (e.g. AddHealth; Karnataka Villages (Banerjee et al. 2013); Harvard social network (Leider et al. 2009)). Sparsity poses a challenge to estimation and inference: if networks are largely empty, there might not be enough variation in centrality measures to identify the parameters of

interest. Despite its importance, sparsity has received relatively little attention in the network econometrics literature.

Second, the observed network may differ from the true network of interest. Centrality measures are often calculated on data which are obtained by survey or constructed using some proxy for interaction between agents, though subsequent analysis would frequently treat the true network as known. Ignoring proxy error may thus lead to estimates that perform poorly. A growing literature works with proxy networks, though they generally do not consider sparse settings. This is important since sparsity and proxy error are mutually reinforcing: sparser networks contain weaker signals, which are in turn more difficult to pick out from noisy proxies. The upshot is that OLS estimators computed on sparse, proxy networks may have particularly poor properties. Asymptotic theory that ignores these features will provide similarly poor approximations to their finite sample behavior. Consequently, estimation and inference procedures based on these theories may lead to invalid conclusions about the economic significance of centrality measures.

This chapter studies the statistical properties of OLS on centrality measures in an asymptotic framework which features both proxy error and sparsity. Our analysis focuses on degree, diffusion and eigenvector centralities, which are among the most popular measures. Our contribution is threefold: (1) We characterize the amount of sparsity at which OLS estimators become inconsistent with and without proxy error, finding that this threshold varies depending on the centrality measure used. Specifically, regression on eigenvector centrality is less robust to sparsity than that on degree and diffusion. This suggests that researchers should be cautious about comparing regressions on different centrality measures, since they may differ in statistical properties in addition to economic

significance. (2) We develop distributional theory for OLS estimators under proxy error and sparsity. We restrict ourselves to sparsity ranges under which OLS is consistent, but we find that asymptotic bias can be large even in this case. Furthermore, the bias may be of larger order than variance, in which case bias correction would be necessary for obtaining non-degenerate asymptotic distributions. Additionally, we find that under sparsity, the estimator converges at a slower rate than is reflected by the usual heteroskedasticity-consistent(hc)/robust standard errors, requiring a different estimator. (3) In view of the distributional theory, we propose novel bias-corrected estimators and inference methods for OLS with sparse, proxy networks. We also clarify the settings under which hc/robust t-statistics are appropriate for testing.

Our theoretical results are derived in an asymptotic framework where networks are modeled as realizations of sparse random graphs. As the number of agents, n, tends to infinity. the expected number of links per agent grows much more slowly than n. Because our statistical model captures important features of real world data, we expect our methods to be reliable for estimation and inference with sparse, proxy networks. We provide simulation evidence supporting this view. The utility of our results is also evident from an application inspired by De Weerdt and Dercon (2006), where we conduct a stylized study of consumption smoothing and social insurance in Nyakatoke, Tanzania.

Our choice of asymptotic framework poses technical challenges. First, the eigenvectors and eigenvalues of sparse random graphs are difficult to characterize. We draw on recent advances in random matrix theory (Alt et al. 2021a,b; Benaych-Georges et al. 2019, 2020) to overcome this challenge. Second, spectral norms of random matrices concentrate slowly in sparse regimes. To obtain our results, we bound the moments of noisy adjacency

matrices by relating them to counts of particular graphs, in the spirit of Wigner (1957) (see Chapter 2 of Tao 2012 more generally). Finally, in order for bias correction to improve mean-squared error, the bias needs to be estimated at a sufficiently fast rate. Because the variance is of a lower order than the bias, a naive plug-in approach does not work for estimating higher-order bias terms, although it is sufficient for the first-order term. We leverage this fact to recursively construct good estimators for higher order terms.

Related Literature

Our work is most closely related to papers that study linear regression with centrality statistics. To our best knowledge, we are the first to study linear regression with diffusion centrality, though there exists prior work on eigenvector centrality. Le and Li (2020) studies linear regression on multiple eigenvectors of a network assuming the same type of proxy error as this chapter. They focus on denser settings than we do and provide inference method only for the null hypothesis that the slope coefficient is 0. We are concerned only with eigenvector centrality, which is the leading eigenvector, but our results cover the sparse case as well as tests of non-zero null hypotheses (more details in Remark 1.6). This chapter is also related to Cai et al. (2021), which proposes penalized regressions on the leading left and right singular vectors of a network. They consider networks that are as sparse as the ones we study, but their networks are observed with an additive, normally distributed error (more details in Remark 1.7). Auerbach (2022) also considers linear regressions with network positions as explanatory variables. However, their approach is nonparametric and results are provided only in the dense case. Outside of the linear regression setting, Cheng et al. (2021) considers inference on deterministic linear functionals

of eigenvectors. They study symmetric matrices with asymmetric noise, proposing novel estimators that leverage asymmetry to improve performance when eigengaps are small. We focus on symmetric matrices with symmetric noise and study the plug-in estimator in which eigenvector is estimated using the noisy adjacency matrix in place of the true matrix.

This chapter also relates to a growing body of work that considers proxy error in centrality statistics. Early work provided simulation evidence that centrality measures on proxy networks become less accurate as sparsity increases (Costenbader and Valente (2003); Borgatti et al. (2006)). Segarra and Ribeiro (2015) theoretically studies the stability of network statistics under perturbations, finding that degree and eigenvector centralities are stable, while betweenness is not. Avella-Medina et al. (2020) and Dasaratha (2020) consider settings similar to ours, with additive proxy errors that are "classical" in that they have mean zero and are independent across edges. These authors provide concentration results for degree and eigenvector centralities among others, but not for diffusion centrality. Additionally, they accommodate less sparsity than us, in part because we are not concerned with estimation of centrality measures, only their use in subsequent regression.

A separate literature has focused on non-classical error in network data. Chandrasekhar and Lewis (2016) examines settings in which researchers have access to a panel of networks, but which are constructed using only a partial sample of nodes or edges. Thirkettle (2019) studies a similar missing data problem, but in a cross-sectional setting with only one network. The author is concerned with forming bounds on centrality statistics and does not consider subsequent linear regression. Griffith (2022) considers

censoring in network data, which arises when agents are only allowed to list a fixed number of relationships during the sampling process. The above papers study missing data problems under the assumption that the observed network is without error. We assume that the entirety of one network is observed but with error. Lewbel et al. (2021) studies more general forms of proxy error in peer effects regression, finding that 2SLS with friends-of-friends instruments is valid as long as the proxy error is small. All of the above papers do not discuss centrality regressions.

This chapter is also connected to the nascent literature on the statistical properties of sparse networks. A strand of this literature is concerned with network formation models that can give rise to sparsity in the observed data. Dong et al. (2020) and Motalebi et al. (2021) consider modifications to the stochastic block model. A more general model takes the form of inhomogeneous Erdos-Renyi graph, which are generated by a graphon with a sparsity parameter that tends to zero in the limit (see for instacne Bollobás et al. 2007 and Bickel and Chen 2009). This chapter takes such an approach. Yet another model for sparse graphs is based on graphex processes, which generalizes graphons by generating vertices through Poisson point processes (see Borgs et al. 2018, Veitch and Roy 2019 and references therein). Our choice of inhomogeneous Erdos-Renyi graphs is motivated by their prevalence in econometrics (Section 3 of De Paula 2017 and Section 6 of Graham 2020a provide many examples), as well as tractability considerations. To our best knowledge, few papers have tackled the challenges that sparse networks pose for regression. Two notable exceptions study network formation models, which take the form of edge-level logistic regressions (Jochmans 2018; Graham 2020b). A separate literature considers estimation of peer effects regressions involving sparse networks using panel data

(Manresa 2016; Rose 2016; De Paula et al. 2020). Here, sparsity is an assumption used to justify regularization methods. We consider a node-level regression in a cross-sectional setting with one large network.

The rest of this chapter is organized as follows. Section 1.2 describes the set-up. Section 1.3 presents the theoretical results. Simulation results are contained in Section 1.4. In Section 1.5, we apply our results to the social insurance network in Nyakatoke, Tanzania. Section 1.6 concludes the chapter with our recommendations for empirical work. All proofs are contained in Appendix A.4.

1.2. Set-Up and Notation

In this section, we introduce notation before describing our econometric model and the asymptotic framework.

We use the following notation. When X is a vector or matrix, X_i and X_{ij} refer the i^{th} and $(i,j)^{\text{th}}$ component of X respectively. Similarly, if X_i or X_{ij} are defined, we use X to denote the full vector or matrix respectively. X' is the transpose of X. When X is a square matrix, $\lambda_j(X)$ denotes the j^{th} eigenvalue of X while $v_j(X)$ denotes the corresponding eigenvector. When $f \in L^2([0,1]^2)$ is a symmetric real function, $\lambda_j(f)$ denotes the j^{th} eigenvalue of the corresponding Hillbert-Schmidt integral operator, $T(g) = \int f(x,y)g(y)dy$, while ϕ_j is the corresponding eigenfunction. For deterministic, monotone sequences x_n and y_n, we write $x_n \succ z_n$ if $x_n/z_n \to \infty$ and $x_n \prec z_n$ if $x_n/z_n \to 0$. $x_n \approx z_n$ indicates that $x_n/z_n \to k$, where $0 < k < \infty$. We write $x_n \succcurlyeq z_n$ to mean $\neg(x_n \prec z_n)$ and similarly for $x_n \preccurlyeq z_n$. Let ι_n be the $n \times 1$ vector of 1's. For two $m \times n$ matrices X and Z, let $X \circ Z$ denote their entrywise (Hadamard) product. Finally, $[n]$ denotes the set of integers from 1 to n.

1.2.1. Econometric Framework

For simplicity, suppose that there are no covariates besides centrality. Consider the regression:

$$Y_i = \beta^{(d)} C_i^{(d)} + \varepsilon_i^{(d)}$$

where i indexes the agents on the network. Y_i is the outcome of interest and $C_i^{(d)}$ is a network centrality measure of type d. We assume that researchers observe $\{Y_i\}_{i=1}^n$ and

either an adjacency matrix A, or a proxy for it, denoted $\hat{A}$. A is an $n \times n$ matrix whose $(i,j)^{\text{th}}$, A_{ij}, records the link intensities between agents i and j. $\hat{A}$ is some estimate of A.

While we do not observe $C_i^{(d)}$, it can be exactly computed using A, or estimated using $\hat{A}$. The parameter of interest is $\beta^{(d)}$. After defining the data-generating process for the true and observed networks, Assumption 1.3 will provide conditions allowing us interpret $\beta^{(d)}$ as the slope coefficient in the linear conditional expectation function of Y_i on $C_i^{(d)}$.

We assume that the data-generating process yields $\{(\varepsilon_i, U_i)\}_{i=1}^n$ which are independent and identically distributed. ε_i is the linear regression residual and U_i is an unobserved latent type that will be used to construct the network.

In the following, we describe (i) the data-generating process for A and $\hat{A}$ via the U_i's and (ii) the use of A and $\hat{A}$ in computing/estimating centrality statistics for OLS estimation. Throughout our discussion, we motivate the econometric framework through the example of consumption smoothing via informal insurance:

Example 1.1. Suppose we are interested in the relationship between informal insurance and consumption smoothing. This is a question that has been studied by De Weerdt and Dercon (2006); Udry (1994); Kinnan and Townsend (2012) and Bourlès et al. (2021) among many others. Here, we might posit that agents which are more central in the informal insurance network can better smooth consumption. To test this hypothesis, we are interested in the regression where Y_i is variance in i's consumption and $C_i^{(d)}$ is centrality in the informal insurance network. $\beta^{(d)}$ is then the reduction in consumption variance associated with being more central. In the informal insurance network, A_{ij} records the probability that i lends money to j or vice versa in the event of an adverse income shock.

However, A is hard to obtain by surveys. Instead, we observe the matrix of actual loans $\hat{A}$, which is a proxy measure of A.

Data-Generating Process for A and $\hat{A}$. Let A be an $n \times n$ symmetric adjacency matrix. We assume that the relationship between two agents in a network is solely determined by their unobserved latent types U_i through the graphon f:

Assumption 1.1 (Graphon)**.** Suppose $U_i \sim U[0,1]$ and $f : [0,1]^2 \mapsto [0,1]$ is such that:

$$\int_{[0,1]^2} f(u,v)\, du\, dv > 0.$$

Let $p_n \in (0,1]$ and $j > i$, define:

$$A_{ij} = p_n f(U_i, U_j) \ .$$

We set $A_{ji} = A_{ij}$ for $j < i$ and normalize $A_{ii} = 0$ for all $i \in [n]$.

In this model, any two agents have a relationship that is between 0 and 1. We can think of this as a measure of intensity, reflecting factors such as duration of friendship, frequency of interaction, or similarity in personalities. Alternatively, it could be the probability with which a relationship is observed. p_n is a parameter that we will let go to 0 at various rates. As we will explain in Section 1.2.2, this is a theoretical device that will help us understand the behavior of OLS estimators when the network is sparse. We restrict attention to symmetric matrices because eigenvector centrality, one of the most popular network centrality measures, may not be well-defined when the adjacency matrix is not symmetric. We also ignore the trivial case when $f = 0$, in which case the network

is always empty. Finally, note that defining $U_i \sim U[0,1]$ is without loss of generality since we have placed no functional form restrictions on f.

Example 1.1 (continued)**.** Suppose that $U_i \in [0,1]$ indexes the riskiness of a villager's income as a result of the crops they choose to cultivate. Assumption 1.1 posits that the relationship between two villagers depends only on their respective income risks. For example, if $f(U_i, U_j) = 1 - (U_i - U_j)^2$, then farmers with similar income risks have higher link intensities between them. U_i can also incorporate other observed or unobserved farmer characteristics, such as place of residence, farming skills or gregariousness. Together with the choice of f, the graphon is a rich model of linking behavior.

When A is observed, we say that there is no proxy error. This setting provides a useful benchmark. When A is not observed, we assume that we have access to the the proxy network, $\hat{A}$, generated as follows.

Assumption 1.2 (Proxy Network)**.** The adjacency matrix of the proxy network is the $n \times n$ matrix symmetric $\hat{A}$, where for $j > i$,

$$\hat{A}_{ij} \,|\, U \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}\left(A_{ij}\right) \ .$$

Set $\hat{A}_{ji} = \hat{A}_{ij}$ for $j < i$. $\hat{A}_{ii} = 0$ since $A_{ii} = 0$. Furthermore, suppose for $d \in \{1, T, \infty\}$ that

$$\hat{A}_{ij} \perp\!\!\!\perp \varepsilon_i^{(d)} \mid U \ .$$

The form of proxy error we consider randomly rounds A_{ij} into zero or one in proportion to the intensity of the true relationship. Conditional on U, this is an additive error with

a mean of 0, and which is independent across agent pairs. Formally, we are assuming a conditionally independent dyad (CID) model for $\hat{A}$. This model is commonly used in econometrics (see for instance Section 3 of De Paula 2017 or Section 6 of Graham 2020a) and is fairly general.

A key motivation for our choice of framework is analytical tractability. Our definitions imply that conditional on U, $\hat{A}$ is a sparse inhomogeneous Erdos-Renyi graph, allowing us to borrow results from the random graph literature. Nonetheless, the model is a reasonable description of network data. Proxy errors of this form often arise due to limitations in data collection or survey methods. Below, we discuss how Example 1.1 fits our proxy error model. More examples can be found in Appendix A.2, where we also discuss our econometric framework in the context of the "Weak Ties" theory of social networks (Granovetter 1973).

Finally, we assume that proxy error is independent of ε_i conditional on U. Together with the CID assumption, proxy error on the network is additive white noise, akin to classical measurement error. It should be distinguished from misclassification error, in which 1's in the adjacency matrices may be observed as 0's and vice versa. In our setting, the key econometric challenge arises because U is unobserved. This is exacerbated by the fact that additive, white noise errors in the network translate into non-linear error in centrality statistics, introducing complications into the analysis.

Example 1.1 (continued)**.** Assumption 1.2 is reasonable in the context of our leading example. Here, each entry of the unobserved A_{ij} represents the probability of loans. However, $\hat{A}_{ij}$ records actual loans, which are realizations of Bernoulli(A_{ij}). The conditional independence assumption means that conditional on friendship, the decision of i to lend

to j is independent of the decision of k to lend to i. This might be the case if the loan amounts are small relative to the income shortfall, so that any agent's decision to lend to i does not significantly reduce their need to borrow. Alternatively, such a condition might be satisfied if borrowing is private, so that friends of i cannot coordinate their lending decisions.

Centrality Statistics and OLS Estimation. Given our adjacency matrices A and $\hat{A}$, we now define centrality statistics and the OLS estimators that are based on them.

Centrality measures are agent-level measures of importance in a network. Many centrality measures exist, each capturing a different aspect of network position. However, they are all functions of A and can be exactly computed when A is observed. We focus on three popular measures: degree, diffusion and eigenvector centralities. While they are most intuitive when A is binary, centrality measures should be understood as functions of general weighted (symmetric) adjacency matrices. Except where noted, our definitions are standard (see e.g. Jackson 2010; Bloch et al. 2021).

Definition 1.1 (Degree Centrality)**.** Degree centrality computed on the $n \times n$ adjacency matrix A is the $n \times 1$ vector:

$$C^{(1)} = A\iota_n \ .$$

Agent i's degree centrality is simply the sum of row i in A. If A is binary, degree centrality is the number of agents with whom i has a relationship.

Definition 1.2 (Diffusion Centrality)**.** For a given $\delta \in [0,1]$ and $T \in \mathbf{N}$, diffusion centrality computed on the $n \times n$ adjacency matrix A is the $n \times 1$ vector:

$$C^{(T)} = \left(\sum_{t=1}^{T} \delta^t A^t \right) \iota_n \ .$$

Proposed by Banerjee et al. (2013), diffusion centrality captures the influence of agent i in terms of how many agents they can reach over T periods. Consider again the case of binary A. Then the $(i,j)^{\text{th}}$ of A^t is the number of walks from i to j that are of length t, which can be thought of as the influence of i on j in period t. Diffusion centrality for agent i is simply sum of their influence on all other agents in the network over time up to period T, with a decay of δ per period. Bramoullé and Genicot (2018) provides further discussion on the theoretical foundations of diffusion centrality. In practice, researchers often choose δ to be $1/\lambda_1(\hat{A})$, so that effectively $\delta \to 0$ as $n \to \infty$. An extension of our results to this case is in preparation.

Definition 1.3 (Eigenvector Centrality)**.** For a given $a_n > 0$, eigenvector centrality computed on the $n \times n$ adjacency matrix A is the $n \times 1$ vector:

$$C^{(\infty)} = a_n v_1(A) \ ,$$

where $v_1(A)$ is the eigenvector corresponding to the eigenvalue of A with the largest absolute value (leading eigenvalue).

Eigenvector centrality is based on the idea that an individual's influence is proportional to the influence of their friends. That is, for some $k > 0$, we seek the following property:

$$C_i^{(\infty)} = k \sum_{j \neq i} A_{ij} C_j^{(\infty)} \quad \text{for all } i \in [n] \ . \tag{1.1}$$

The eigenvectors of A solve the above equations, with k being the corresponding eigenvalue. By the Perron-Frobenius Theorem, the leading eigenvector is the unique eigenvector that can be chosen so that every entry is non-negative, motivating its use as a centrality measure. The leading eigenvector of related matrices also emerge as measures of influence in popular models of social learning (e.g. DeGroot 1974)

The leading eigenvector is well-defined only if the largest eigenvalue of A has multiplicity 1, that is, if $\lambda_1(A) \neq \lambda_2(A)$. To ensure that this occurs with high probability, we will make the following assumption when analyzing eigenvector centrality:

Assumption E1. Suppose $\lambda_1(f) \neq \lambda_2(f)$.

Note also that eigenvectors are defined only up to scale: if C satisfies Equation 1.1, so will $a_n C$ for any $a_n \in \mathbf{R}$. Eigenvector centrality is commonly defined to have length 1 (e.g. Banerjee et al. 2013; Cruz et al. 2017), although researchers sometimes scale eigenvectors so that its standard deviation is 1 (Chandrasekhar 2016; Banerjee et al. 2019). Of the two papers that have considered the statistical properties of regression on eigenvector centrality, Cai et al. (2021) sets the length to $\sqrt{n}$, claiming it to be a normalization. Le and Li (2020) does not fix the length, though their goal is essentially to recover the projection $C^{(d)}\beta^{(d)}$ and not $\beta^{(d)}$ itself. We depart from the literature by leaving a_n as a free parameter. We will analyze the properties of regression on eigenvector centrality

making explicit their dependence on a_n. As we explain in Section 1.3.1, the choice of a_n is not innocuous and can have implications for estimation and inference.

This paper focuses on the above three centrality measures, which are intimately related (Bloch et al. 2021). When $T = 1$, $C^{(1)} \propto C^{(T)}$. Furthermore, as shown by Banerjee et al. (2019), if $\delta \geq 1/\lambda_1(A)$,

$$\lim_{T\to\infty} C^{(T)} \propto C^{(\infty)} .$$

We can thus understand the centrality measures as lying on a line, motivating our notational choice. Notably, we do not discuss betweenness and closeness centralities. These are path-based measures of centrality,which do not have clearly defined counterparts in the graphon. We conjecture that their analysis require a different statistical framework and defer it to future work.

Example 1.1 (continued)**.** In the context of risk sharing and social insurance, we can interpret

- $C_i^{(1)}$ as the probability-weighted number of friends who will lend to or borrow from i.
- $C_i^{(T)}$ as the probability-weighted number of friends who will lend to or borrow from i directly or through their friends. T is the maximum length of the borrowing chain. For example if T is 2, i can borrow from friends of friends but not friends of friends of friends. δ is the increased difficulty of borrowing from a person that is one step further, e.g. of borrowing from friends of friends relative to borrowing from a friend directly.

- $C_i^{(\infty)}$ as requiring the borrowing ability of i to be proportional to the borrowing ability of their friends. Implicitly, this means agents can form borrowing chains that are arbitrarily long.

When A is observed,we have access to the following infeasible estimators.

Definition 1.4 (OLS Estimators without Proxy Error)**.** Suppose A is observed. For $d \in \{1, T, \infty\}$, define the OLS estimators for $\beta^{(d)}$ to be

$$\tilde{\beta}^{(d)} = \frac{Y' C^{(d)}}{\left(C^{(d)}\right)' C^{(d)}} \ .$$

When networks are observed with errors, we assume that network centralities are estimated using $\hat{A}$ in place of A:

Definition 1.5 (Centralities with Proxy Error)**.** Suppose $\hat{A}$ is observed but not A. Define:

$$\begin{aligned}
\hat{C}^{(1)} &= \hat{A}\iota_n \ , \\
\hat{C}^{(T)} &= \left(\sum_{t=1}^{T} \delta^t \hat{A}^t \right) \iota_n \ , \\
\hat{C}^{(\infty)} &= a_n v_1(\hat{A}) \ .
\end{aligned}$$

The corresponding OLS estimators are thus defined using the noisy centrality measures.

Definition 1.6 (OLS Estimators with Proxy Error)**.** Suppose $\hat{A}$ is observed but not A. For $d \in \{1, T, \infty\}$, define the OLS estimators for $\beta^{(d)}$ to be

$$\hat{\beta}^{(d)} = \frac{Y'\hat{C}^{(d)}}{\left(\hat{C}^{(d)}\right)'\hat{C}^{(d)}} \ .$$

Next, define the regression residuals.

Definition 1.7 (Regression Residuals)**.** For $d \in \{1, T, \infty\}$, define:

$$\tilde{\varepsilon}_i^{(d)} := Y_i - \tilde{\beta}^{(d)} C_i^{(d)} \ , \tag{1.2}$$

$$\hat{\varepsilon}_i^{(d)} := Y_i - \hat{\beta}^{(d)} \hat{C}_i^{(d)} \ . \tag{1.3}$$

We conclude this section with an assumption on the moments of ε_i conditional on U_i:

Assumption 1.3. Suppose for $d \in \{1, T, \infty\}$ that:

(a) $E\left[\varepsilon_i^{(d)} | U_i\right] = 0$

(b) $0 < \underline{\sigma}^2 \leq E\left[\left(\varepsilon_i^{(d)}\right)^2 | U_i\right] \leq \bar{\sigma}^2 < \infty.$

(c) $E\left[\left|\varepsilon_i^{(d)}\right|^3 | U_i\right] \leq \bar{\kappa}_3.$

In the above, (a) justifies linear regression since it implies that

$$E\left[\varepsilon_i^{(d)} \,\middle|\, C_i^{(d)}\right] = 0 \ .$$

Meanwhile, (b) and (c) control the amount of heterogeneity across different U_i's. (c) implies the upper bound in (b). We introduce $\bar{\sigma}^2$ for notational convenience.

1.2.2. Sparse Network Asymptotics

To better capture the behavior of estimators when agents in the networks have few relationships with one another, we study their properties under *sparse* network asymptotics. Following Bollobás et al. (2007) and Bickel and Chen (2009), we want to consider settings in which $p_n \to 0$ as $n \to \infty$. p_n is not an empirical quantity. It is a theoretical device to ensure that the sequence of models we consider remains sparse even as $n \to \infty$.

In many settings, a vector or matrix is said to be sparse if many of the entries are 0. In our setting, we say that A and $\hat{A}$ are sparse if their row sums – that is, the actual or observed degrees of agents respectively – are small. Because the entries of $\hat{A}$ are restricted to be binary, having low degrees is the same as having many entries which are 0. We do not place such a restriction on the entries of A, so that row sums could be small even if no entry takes value 0, as long as each non-zero entry is small. Sparsity of A should therefore be understood as referring to low intensities of relationship between agents, but which gives rise to observed networks, $\hat{A}$, that are sparse in the conventional sense.

To see how $p_n \to 0$ gives rise to sparsity, suppose for example that $p_n \to c > 0$. Then the network is dense and each agent has total relationships that are roughly of order n in expectation. That is,

$$E\left[C_i^{(1)}\right] \approx n,$$

corresponding to a situation in which each agent is linked to many others. In practice, however, researchers may face sparse networks, in which each agent has few or weak relationships. Choosing $p_n \to 0$ leads to networks that remain sparse as n increases. For

example, if we set $p_n = k/n$ for some $k > 0$, then,

$$E\left[C_i^{(1)}\right] \approx 1 \ .$$

That is, each agent has a bounded number of relationships in expectation. A sequence of p_n that goes to 0 more quickly corresponds to data which is more sparse.

To understand the effect of sparsity on OLS estimation, we study how the statistical properties of $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ change as we vary the rate at which $p_n \to 0$. Our goal is to obtain theoretical results that better describe the properties of estimators under sparsity by explicitly incorporating it into the asymptotic framework. Using p_n to model sparse networks is standard in statistics literature (see e.g. Bickel and Chen 2009; Bickel et al. 2011, Borgs et al. 2018 Avella-Medina et al. 2020 among many more). Within econometrics, related approaches have been used to study network formation models (Jochmans 2018; Graham 2020b).

As a theoretical device, p_n bears semblance to drifting alternatives in local power analysis (also known as Pitman drift; see Rothenberg 1984). Suppose we want to compare the power of tests for the hypothesis $H_0 : \beta = \beta_0$ against $H_1 : \beta = \beta_1$. Asymptotic analysis with a fixed β_1 is not useful since consistent tests have power that converges to 1 in probability under all alternatives, so that we cannot meaningfully differentiate between these tests. One interpretation of such a failure is that the asymptotic model fails to capture reality: in the limit, $|\beta_1 - \beta_0|$ is large relative to the sampling noise which is of order $1/\sqrt{n}$. In practice, sampling noise can be large relative to the parameter of interest. Local power analysis employs the alternative hypothesis $\beta_1 = \beta_0 + k/\sqrt{n}$. As such, $|\beta_1 - \beta_0| = |k/\sqrt{n}|$ goes to 0 at the same rate as sampling noise. Intuitively, as the

sample size gets larger, the testing problem also becomes harder. The upshot is that the testing problem is non-trivial even in the limit, better modeling the finite sample problem.

A similar approach is taken in the weak instruments literature, which is concerned with the instrumental variable regressions in which the relevance condition is barely satisfied. To understand the resulting statistical pathologies, Staiger and Stock (1997) propose to model the strength of the instrument as decaying to 0 at rate $1/\sqrt{n}$, so that strength of the signal in the first stage estimation is on par with sampling uncertainty. This approach has since led to long and productive lines of inquiry (see Andrews et al. 2019 and references therein).

Our drifting parameter p_n serves a similar purpose: by letting $p_n \to 0$, we better capture the statistical properties of estimators when networks are sparse. While we do not focus on any reference level of sparsity, comparing across levels of sparsity will prove instructive.

1.3. Theoretical Results

In this section, we present our theoretical results about the property of OLS estimators under varying amounts of sparsity. In Section 1.3.1, we characterize the level of sparsity at which consistency of $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ fails. The upshot is that proxy error renders OLS estimators less robust to sparsity. In particular, eigenvector centrality is less robust to sparsity than degree under proxy error. The regimes for p_n cannot be estimated from data. Instead, we provide a rule-of-thumb for gauging the reliability of OLS estimators. Section 1.3.2 presents distributional theory for $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ under regimes of sparsity for

under which they are consistent. This leads to tools for bias correction and inference with sparse and noisily measured networks.

1.3.1. Consistency

This section presents the rates on p_n at which $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ are consistent. We also discuss the role of a_n in ensuring the consistency of $\tilde{\beta}^{(\infty)}$ and $\hat{\beta}^{(\infty)}$. Since rates on p_n cannot be estimated, we present rules-of-thumb for determining the amount of sparsity in practical applications.

We first consider the case when the true network A, is observed:

Theorem 1.1 (Consistency without Proxy Error)**.** Suppose Assumptions 1.1, 1.2 and 1.3 hold. Then,

(a) For $d \in \{1, T\}$, $\tilde{\beta}^{(d)} \xrightarrow{p} \beta^{(d)}$ if and only if $p_n \succ n^{-\frac{3}{2}}$.

(b) Suppose Assumption E1 also holds. Then, $\tilde{\beta}^{(\infty)} \xrightarrow{p} \beta^{(\infty)}$ if and only if $a_n \to \infty$.

As such, we have consistency of OLS for degree and diffusion centralities provided that the network is not *too* sparse. Under extreme sparsity, variation in $C_i^{(d)}$ becomes much smaller than variation in ε_i and it is not possible to learn about $\beta^{(d)}$. In the case of eigenvector centrality, consistency requires conditions on the normalization factor a_n but not on p_n. This is because a_n directly controls the variance of $C^{(\infty)}$, so that it is able to undo the effect of sparsity in the absence of measurement error.

Our result is similar in spirit to Conley and Taber (2011), which studies the properties of difference-in-difference (DiD) estimators when there are few treated units. In an asymptotic framework that takes the number of treated units to be fixed, the DiD estimator

is similarly inconsistent in the limit. More generally, consistency of OLS with i.i.d. data requires $\sqrt{n}\sigma_X \to \infty$, where σ_X is the variance of the regressor. Theorem 1.1 instantiates this condition for centrality regressions under sparsity.

Interestingly, the choice of a_n matters even when the network is dense. To see why, suppose $f = p_n \cdot 1$ so that $A = p_n \iota_n \iota_n'$. Then $C^{(\infty)}(A) = a_n \iota_n / \sqrt{n}$. Note that it is independent of p_n. We can then write:

$$\tilde{\beta}^{(\infty)} = \frac{\sqrt{n}}{a_n} \cdot \frac{Y'\iota_n}{\iota_n'\iota_n} = \beta^{(\infty)} + \frac{1}{a_n\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^{(d)} \ .$$

Under our assumptions, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^{(d)} \xrightarrow{d} N\left(0, \text{Var}\left(\varepsilon_i^{(d)}\right)\right)$. Therefore, $a_n \to \infty$ is necessary for the consistency of $\tilde{\beta}^{(\infty)}$.

The above example, together with Theorems 1.2 and 1.6 in the next section, makes clear that a_n has important implications for the statistical properties of $\tilde{\beta}^{(\infty)}$ and $\hat{\beta}^{(\infty)}$. We can understand this phenomenon by analogy to OLS with i.i.d. observations, in which we are able to consistently estimate β but not $\sqrt{n}\beta$.

To our knowledge, we are the first to emphasize the importance in choosing a_n appropriately. Various a_n's are used in applied work and in econometric theory. Applied researchers sometimes set $a_n = 1$ (e.g. Cruz et al. 2017; Banerjee et al. 2013). Other times, they divide $v_1(\hat{A})$ by its standard deviation, in order to interpret $\beta^{(\infty)}$ as the effect of a one standard-deviation increase in $v_1(A)$ (e.g. Chandrasekhar et al. 2018; Banerjee et al. 2019). Corollary 1.5 shows that $a_n \approx \sqrt{n}$ under some conditions on f. Cai et al. (2021), which studies eigenvector regressions under a different model for proxy error, sets a_n to $\sqrt{n}$. In the formulation of Le and Li (2020), a_n appears only implicitly and they do not prove consistency of $\hat{\beta}^{(\infty)}$. Instead, they show that $\|\hat{\beta}^{(\infty)}\hat{C}^{(\infty)} - \beta^{(\infty)}C^{(\infty)}\|_\infty \xrightarrow{p} 0$.

Of papers which study estimation of centrality statistics, Avella-Medina et al. (2020) sets $a_n = \sqrt{n}$ while Dasaratha (2020) sets $a_n = 1$. We remark that changing a_n amounts to changing the definition of $\tilde{\beta}^{(\infty)}$. The parameter of interest ultimately depends on the researcher. From the perspective of consistency, however, models with $a_n \to \infty$ are strictly preferable to those with $a_n \preccurlyeq 1$. And as we will see in Theorem 1.6, particular choices of a_n may be useful for inference.

We next consider the case with proxy error:

Theorem 1.2 (Consistency with Proxy Error)**.** Suppose Assumptions 1.1, 1.2 and 1.3 hold. Then,

(a) For $d \in \{1, T\}$, $\hat{\beta}^{(d)} \xrightarrow{p} \beta^{(d)}$ if and only if $p_n \succ n^{-1}$.

(b) Suppose also that Assumption E1 holds. Then, $\hat{\beta}^{(\infty)} \xrightarrow{p} \beta^{(\infty)}$ if $a_n \to \infty$ and

$$p_n \succ n^{-1}\sqrt{\frac{\log n}{\log\log n}}\,. \tag{1.4}$$

Suppose p_n satisfies

$$n^{-1}\left(\log\log n\right)^4 \prec p_n \prec n^{-1}\sqrt{\frac{\log n}{\log\log n}}\,. \tag{1.5}$$

Then $\hat{\beta}^{(\infty)}$ is inconsistent for $\beta^{(\infty)}$.

Theorem 1.2 gives the rates at which OLS regression on each centrality is consistent with proxy error. We summarize the rates from Theorems 1.2, together with that from 1.1, in Figure 1.1.

With proxy error, $\hat{\beta}^{(\infty)}$ is consistent under less sparsity than $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$, even when we set $a_n \to \infty$. In other words, $\hat{\beta}^{(\infty)}$ is less robust to sparsity than $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$. This

occurs because eigenvalues and eigenvectors are sensitive to noise under sparsity. Our proof uses the homogeneous Erdos-Renyi graph as a counterexample. Suppose $f = 1$ so that A is rank 1 with $v_1(A) = \iota/\sqrt{n}$. When the lower bound in Equation 1.5 is satisfied, Alt et al. (2021b) shows that $\hat{A}$ has an eigenvalue, call it ν, with a corresponding eigenvector that is approximately $\iota/\sqrt{n}$. If ν is the largest eigenvalue of $\hat{A}$, $v_1(\hat{A})$ is close to $v_1(A)$. However, when p_n satisfies Equation 1.5 , ν turns out to be much smaller than $\lambda_1(\hat{A})$. The result is that $v_1(\hat{A})$ is almost orthogonal to $v_1(A)$. Intuitively, sparsity weakens the signal in $\hat{A}$, so that its leading eigenvector is pure noise.[1] It then becomes impossible to estimate the leading eigenvector of A for OLS estimation. On the other hand, consistency of $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ only requires the mean of $\hat{A}$ to concentrate to that of A, which occurs as long as $p_n \succ n^{-1}$.

An important implication of our result is that centrality measures may have differing predictive value for outcomes in sparse regimes, not only because they differ in economic significance, but also because they differ in statistical properties. In particular, suppose diffusion centrality leads to estimates which are significantly different from 0 at some level α, while eigenvector does not. If the underlying networks are sparse, it would be erroneous to conclude that diffusion centrality is structurally meaningful while eigenvector is not, since sparsity might be driving the observed results.

Finally, let us compare the rates in Theorem 1.2 with those in Theorem 1.1. As Figure 1.1 shows, proxy error renders OLS less robust to sparsity. While $\tilde{\beta}^{(1)}$ and $\tilde{\beta}^{(T)}$ are consistent as long as $p_n \succ n^{-3/2}$, $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ now require that $p_n \succ n^{-1}$. Whereas $\tilde{\beta}^{(\infty)}$ did not require any conditions on p_n for consistency, $\hat{\beta}^{(\infty)}$ does. Moreover, this

[1] In fact, $v_1(\hat{A})$ exhibits localization. That is, its mass concentrates on the agent who happens to have the largest realized degree, which is purely a result of chance.

requirement is more stringent than that on $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$. OLS on eigenvector centrality is therefore more sensitive to proxy error than on degree or diffusion.

Remark 1.1. Avella-Medina et al. (2020) and Dasaratha (2020) provide results essentially showing that for $d \in \{1, \infty\}$, $\|\hat{C}^{(d)} - C^{(d)}\| \to 0$ with probability approaching 1 if $p_n \succ \frac{\log n}{n}$. Our focus is on the OLS estimators $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ and we find that thresholds of consistency that are strictly below $\frac{\log n}{n}$ for all $d \in \{1, T, \infty\}$.

Remark 1.2. Theorem 1.2 does not determine the behavior of $\hat{\beta}^{(\infty)}$ when $p_n \prec n^{-1} (\log \log n)^4$. Up to this threshold, we know by Alt et al. (2021b) that OLS is inconsistent only because we have descriptions of both eigenvalues and eigenvectors. To our knowledge, recent developments in random matrix theory do not provide any description of eigenvectors below this threshold. Hence, it is not clear what type of pathologies arises below $p_n \prec n^{-1} (\log \log n)^4$ and how that might affect the behavior of $\hat{\beta}^{(\infty)}$. Description of eigenvalues is more complete: below this point, we know that $\lambda_1(\hat{A})/\lambda_1(A) \to \infty$ (see Alt et al. 2021a; Benaych-Georges et al. 2019; Benaych-Georges et al. 2020). Since the estimated eigenvalues are noise, we conjecture that the estimated eigenvectors are as well. If so, we would not expect $\hat{\beta}^{(\infty)}$ to be consistent.

Remark 1.3. To improve the robustness of eigenvector centrality to sparsity, we can consider regularizing $\hat{A}$. Appendix A.3 considers such an approach and finds that consistency with regularized eigenvector obtains when $p_n \succ n^{-1}$.

Rule-of-Thumb for Determining Consistency of $\hat{\beta}^{(d)}$. Theorem 1.2 provides consistency results for $\hat{\beta}^{(d)}$ based on the rates at which $p_n \to 0$. It is therefore desirable to

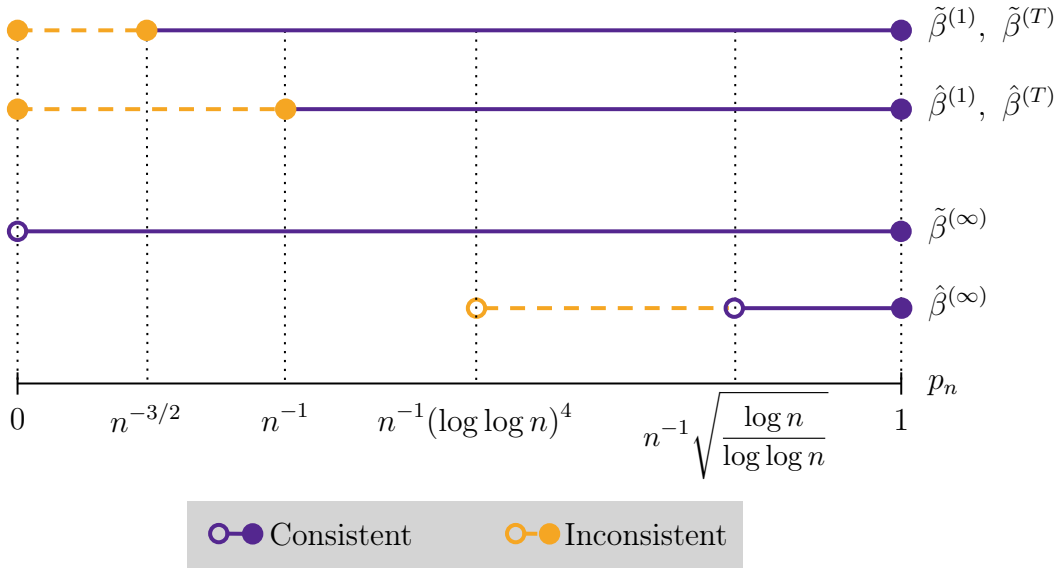

Figure 1.1. Ranges of consistency for each estimator. When the network is observed with error, regression on eigenvector centrality is less robust to sparsity than on degree or diffusion. When the network is known, much more sparsity can be accommodated.

have methods for determining if we are in the regime for which OLS is consistent with measurement error. However, p_n is a theoretical device and the rate at which it is going to 0 is not a quantity that can be estimated. Instead, we propose a rule-of-thumb for determining the regime of p_n and the consistency of $\hat{\beta}^{(d)}$.

Definition 1.8 (Connectivity)**.** Let $\hat{A}$ be a binary network. Two agents i and j are connected if $\hat{A}$ contains a path from i to j. A collection of agents $\mathcal{I}$ is connected if every pair in $\mathcal{I} \times \mathcal{I}$ is connected. We say that $\hat{A}$ is connected if $[n]$ is connected.

Definition 1.9 (Largest Component)**.** For a binary network $\hat{A}$, define

$$L_1 := \max_{\mathcal{I} \subseteq [n],\ \text{connected}} |\mathcal{I}| \ .$$

We say that $\hat{A}$ has a giant component if $L_1 \approx n$.

In words, $\hat{A}$ has a giant component if its largest connected component is a non-vanishing fraction of the total number of nodes.

It is well-known that inhomogeneous random graphs exhibit threshold behavior in connectivity and in the existence of the largest component at the following rates:

Theorem 1.3. Suppose Assumptions 1.1 and 1.2 hold. Then as $n \to \infty$,

(a) (Theorem 3.1, Bollobás et al. 2007). $p_n \approx n^{-1}$ if and only if $\hat{A}$ has a unique giant component with probability approaching 1.

(b) (Theorem 1, Devroye and Fraiman 2014). $p_n \approx n^{-1} \log n$ if and only if $\hat{A}$ is connected with probability approaching 1.

In words, if we observe that $\hat{A}$ has a giant component, we can expect that $p_n \succ n^{-1}$. If we observe that $\hat{A}$ is connected, then we can expect that $p_n \succ n^{-1} \log n \succ \sqrt{\frac{\log n}{\log n \log n}}$. Together with Theorem 1.2, this motivates the following rule-of-thumb:

Rule of Thumb 1.1.

(a) Treat $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ as consistent only if $\hat{A}$ has a giant component with $L_1 > n/2$.

(b) Treat $\hat{\beta}^{(\infty)}$ as consistent only if $\hat{A}$ is connected.

Note that if $\hat{A}$ is connected, it also has a giant component of size n. Our criteria are therefore nested. The choice of the constant $1/2$ in rule (a) ensures uniqueness of the largest component, but is technically arbitrary. Figure 1.2 provides graphical illustration.

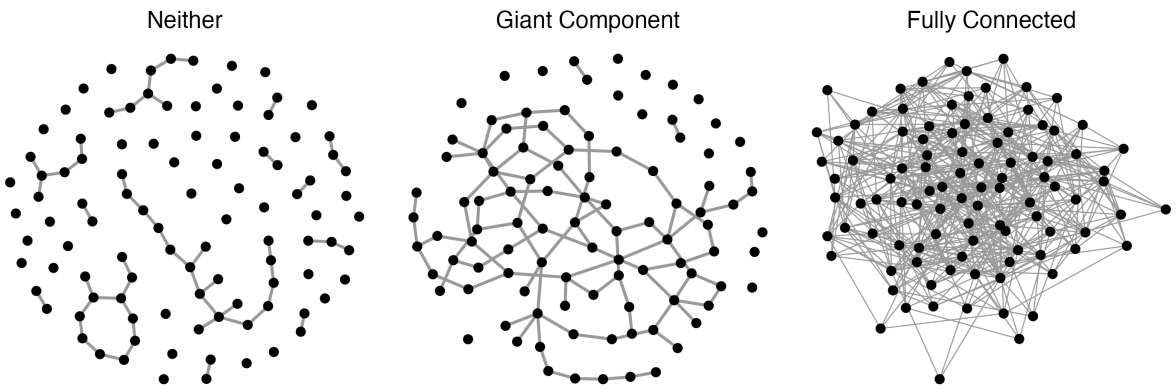

Figure 1.2. From left to right, a connected network, a disconnected network with giant component, and a disconnected network without a giant component. Networks are obtained from simulation with $f = 1$, $n = 100$ and $p_n = 1/\sqrt{n}, \sqrt{\log n}/n$ and $1/n$ respectively.

Remark 1.4. Our rule of thumb for the consistency of $\hat{\beta}^{(\infty)}$ essentially requires that $p_n \succ n^{-1} \log n$. This is not necessary. Instead, we could, for example, formulate a rule-of-thumb based on localization of the leading eigenvector. However it is less appealing to do so since such a criteria is non-nested with our criterion for a giant component, and additionally introduces a tuning parameter.

1.3.2. Distributional Theory

In this section, we study the asymptotic distributions of $\tilde{\beta}^{(d)}$ and $\hat{\beta}^{(d)}$ under sparsity and proxy error. We focus on regimes of p_n under which each estimator is consistent and find that proxy error still leads to asymptotic bias. Specifically,

$$\hat{\beta}^{(d)} \xrightarrow{p} \beta^{(d)} \quad \text{but} \quad E\left[\lim_{n\to\infty} np_n \left(\hat{\beta}^{(d)} - \beta^{(d)}\right)\right] =: B^{(d)} \neq 0 \ .$$

Furthermore, the bias may be of larger order than the standard deviation of $\hat{\beta}^{(d)}$. In this case, it would not be possible to obtain a non-degenerate limiting distribution without bias correction.

As such, we propose bias-correction and corresponding inference methods based on $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$. The distribution of $\hat{\beta}^{(\infty)}$ is more tricky to characterize. We will propose a data-dependent choice of a_n that leads to convenient properties. Readers who are only interested in the implementation of inference can refer to the summary in Table 1.1.

1.3.2.1. Centralities without Proxy Error ($\tilde{\beta}^{(1)}, \tilde{\beta}^{(T)}, \tilde{\beta}^{(\infty)}$). Our first result states that heteroskedasticity-consistent (hc) or robust t-statistics yield vaild inference in the absence of proxy error.

Theorem 1.4. Suppose Assumptions 1.1 and 1.3 hold.

(a) Suppose $p_n \succ n^{-3/2}$. Then, for $d \in \{1, T\}$,

$$\tilde{S}^{(d)} = \frac{\tilde{\beta}^{(d)} - \beta^{(d)}}{\sqrt{\tilde{V}^{(d)}}} \xrightarrow{p} \mathrm{N}(0,1) \ .$$

where $\tilde{V}^{(d)} = \left(\sum_{i=1}^n \left(C_i^{(d)}\right)^2\right)^{-2} \sum_{i=1}^n \left(C_i^{(d)}\right)^2 \left(\tilde{\varepsilon}_i^{(d)}\right)^2$.

(b) Suppose $a_n \to \infty$. Then,

$$\tilde{S}_0^{(\infty)} = \frac{\tilde{\beta}^{(\infty)} - \beta^{(\infty)}}{\sqrt{\tilde{V}^{(\infty)}}} \xrightarrow{p} \mathrm{N}(0,1) \ .$$

where $\tilde{V}^{(\infty)} = \left(\sum_{i=1}^n \left(C_i^{(\infty)}\right)^2\right)^{-2} \sum_{i=1}^n \left(C_i^{(\infty)}\right)^2 \left(\tilde{\varepsilon}_i^{(\infty)}\right)^2$.

In the above, $\tilde{\varepsilon}_i$ is as defined in Equation (1.2).

Our formulation of the t-statistic highlights that inference on $\beta^{(1)}$ and $\beta^{(T)}$ does not require the sparsity parameter p_n to be specified. This is important since p_n is in general not identified (Bickel et al. 2011) and follows from the convenient fact that the t-statistic is self-normalizing. Intuitively, the sparsity terms in the numerator and the denominator are of the same order, so that they "cancel out". Hansen and Lee (2019a) makes a similar observation in the context of cluster-dependent data: although the means of such data converge at a rate that changes based on the dependence structure within each cluster, this rate does not need to be known for estimation and inference, due to the aforementioned self-normalizing property.

We note that $\tilde{V}^{(1)} = O_p\left(n^{-3}p_n^{-2}\right)$, $\tilde{V}^{(T)} = O_p\left(n^{-2T-1}p_n^{-2T}\right)$. These are the rates of convergence for $\tilde{\beta}^{(1)}$ and $\tilde{\beta}^{(T)}$ respectively. In the absence of sparsity (i.e. if $p_n = 1$), the rate of convergence is faster than $n^{-1/2}$. This is because having a network amounts to n^2 observations. Asymptotically, the regressor $C_i^{(d)}$ has much more variation than the regression error ε_i, leading to the higher rate of convergence. Finally, we note that $\tilde{V}^{(\infty)} = O_p(a_n^{-2})$.

In the presence of proxy error, however, the above result does not obtain. The next two subsections presents distributional theory for $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$, and $\hat{\beta}^{(\infty)}$.

1.3.2.2. Degree and Diffusion Centrality under Proxy Error ($\hat{\beta}^{(1)}$, $\hat{\beta}^{(T)}$). For $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$, proxy error leads to bias and also slows down the rate of convergence. This is the content of the following theorem:

Theorem 1.5 (Inference – Degree and Diffusion)**.** Suppose Assumptions 1.1, 1.2 and 1.3 hold and that $p_n \succ n^{-1}$.

(a) Suppose $\beta^{(1)} \neq 0$. Then,

$$\hat{S}^{(1)} := \frac{\hat{\beta}^{(1)} - \beta^{(1)}\left(1 - \hat{B}^{(1)}\right)}{\beta^{(1)}\sqrt{\hat{V}^{(1)}}} \xrightarrow{d} \mathrm{N}(0,1) \ ,$$

where

$$\hat{V}^{(1)} = \frac{1}{2}\left(\sum_{i=1}^{n}\left(\hat{C}_i^{(1)}\right)^2\right)^{-2}\sum_{j\neq i}\hat{A}_{ij}\left(\hat{C}_i^{(1)} + \hat{C}_j^{(1)}\right)^2 \quad , \quad \hat{B}^{(1)} = \left(\sum_{i=1}^{n}\left(\hat{C}_i^{(1)}\right)^2\right)^{-1}\iota_n\hat{A}\iota_n \ .$$

(b) Suppose $\beta^{(T)} \neq 0$. Then,

$$\hat{S}^{(T)} = \frac{\hat{\beta}^{(T)} - \beta^{(T)}\left(1 - \hat{B}^{(T)}\right)}{\beta^{(T)}\sqrt{\hat{V}^{(T)}}} \xrightarrow{d} \mathrm{N}(0,1) \ ,$$

where

$$\hat{V}^{(T)} = \frac{1}{2}\left(\sum_{i=1}^{n}\left(\hat{C}_i^{(T)}\right)^2\right)^{-2}\cdot\delta^{4T}\cdot\iota_n'\left[\hat{A}\circ\left(\sum_{t=1}^{2T}\left(\hat{A}^{2T-t}\iota_n\right)\left(\iota_n'\hat{A}^{t-1}\right)\right)^{\circ 2}\right]\iota_n \ ,$$

$$\hat{B}^{(T)} = \left(\sum_{i=1}^{n}\left(\hat{C}_i^{(T)}\right)^2\right)^{-1}\sum_{t=1}^{2T-1} b_T(t,\delta)\cdot\iota_n\hat{A}^t\iota_n \ .$$

Here, $\circ$ denotes the entrywise product. The formula for $b_T(t,\delta)$, up to $T = 10$, can be found in Appendix A.1.

(c) Suppose for $d \in \{1, T\}$ that $\beta^{(d)} = 0$. Then,

$$\hat{S}_0^{(d)} := \frac{\hat{\beta}^{(d)}}{\sqrt{\hat{V}_0^{(d)}}} \xrightarrow{d} \mathrm{N}(0,1) \ ,$$

where

$$\hat{V}_0^{(d)} = \left(\sum_{i=1}^{n}\left(\hat{C}_i^{(T)}\right)^2\right)^{-2}\sum_{i=1}^{n}\left(\hat{C}_i^{(d)}\right)^2\left(\hat{\varepsilon}_i^{(d)}\right)^2 \ .$$

Here, $\hat{\varepsilon}_i^{(d)}$ is as defined in Equation (1.3).

Our results are stated in terms of $\hat{B}^{(d)}$ and $\hat{V}^{(d)}$ – estimators for bias and variance – though they should be understood as statements about the true bias and variance of the estimators in combination with statements about estimation feasibility. Note also that results for $\hat{\beta}^{(T)}$ specializes to that for $\hat{\beta}^{(1)}$ when setting $T = \delta = 1$.

When $\beta^{(d)} = 0$ (case (c)), our result asserts that the hc/robust variance estimator is consistent for the variance of $\hat{\beta}^{(d)}$. However, that is no longer the case when then $\beta^{(d)} \neq 0$ (cases (a) and (b)). Here, we find that $\hat{\beta}^{(d)}$ become biased. That is, $\hat{\beta}^{(d)}$ is not centered at $\beta^{(d)}$. The bias of $\hat{\beta}^{(1)}$ comprises only one term. However, bias for $\hat{\beta}^{(T)}$ comprises an exponentially growing number of terms. This provides another intuitive explanation for the poor properties of the eigenvector centrality, since as Banerjee et al. (2019) proves, can be considered the limit of diffusion centrality as $T \to \infty$. Comparing cases (a) and (b) with (c) also shows that the asymptotic distributions for $\hat{\beta}^{(d)}$ are discontinuous in $\beta^{(d)}$ at 0.

Additionally, we see that the asymptotic variance of $\hat{\beta}^{(d)}$ differs from that which is estimated by hc/robust standard error. In fact, $\hat{V}_0^{(d)}/\hat{V}^{(d)} \xrightarrow{p} 0$. Note that the difference in asymptotic variance is not the result of bias estimation. In particular, replacing $\hat{B}^{(d)}$ with its limit in probability (appropriately scaled) will not change the asymptotic variance of $\hat{S}^{(d)}$. This stands in contrast to settings such as Regression Discontinuity Design, in which estimation of the asymptotic bias leads to larger asymptotic variance in the relevant test statistic (Calonico et al. 2014).

Additionally, $\sqrt{\hat{V}^{(d)}}/\hat{B}^{(d)} = O_p(p_n)$ so that the bias is of larger order than the variance. Bias correction is therefore necessary for obtaining a non-degenerate asymptotic

distribution. To see this, write:

$$\frac{\hat{\beta}^{(d)} - \beta^{(d)}}{\sqrt{v_n}} = \underbrace{\frac{\hat{\beta}^{(d)} - \beta^{(d)} - B^{(d)}/np_n}{\sqrt{v_n}}}_{=:\Gamma_1} + \underbrace{\frac{B^{(d)}/np_n}{\sqrt{v_n}}}_{=:\Gamma_2} .$$

Suppose we chose $v_n = \mathrm{Var}(\hat{\beta}^{(d)})$. Then $\Gamma_1 \xrightarrow{d} D$, where D is some non-degenerate distribution. However, Γ_2 diverges to $+\infty$ or $-\infty$ depending on the sign of B^d. On the other hand, suppose we chose v_n so that Γ_2 is bounded. Then $\Gamma_1 \xrightarrow{d} 0$ since $\mathrm{Var}(\hat{\beta}^{(d)})/v_n \to 0$. That is, its limit is degenerate. Bias correction is thus necessary for inference.

In order for bias correction to improve mean-squared error, bias must be estimated at a sufficiently fast rate. This is not trivial for $\hat{\beta}^{(T)}$. Bias of the $\hat{\beta}^{(T)}$ comprises terms of the form $\iota_n A^t \iota_n$. However, the naive plug-in estimator $\iota_n \hat{A}^t \iota_n$ does not converge sufficiently fast for $t \geq 2$, even though it works well for $t = 1$. Using this latter fact, we recursively construct good estimators for $\iota_n \hat{A}^t \iota_n$ when $t \geq 2$, which can then be used to construct $\hat{B}^{(T)}$. The resulting estimator does not have a closed form expression in terms of T. We provide explicit formulae for $T \leq 10$ in Tables A.1 and A.2.

Hypothesis Testing. Our theory suggests the following test for $d \in \{1, T\}$.

Definition 1.10. To test the hypothesis $H_0 : \beta^{(d)} = \beta_0$ against $H_1 : \beta^{(d)} \neq \beta_0$ at the significance level of α, define

$$\phi^{(d)} = \begin{cases} \mathbf{1}\left\{ \left| \frac{\hat{\beta}^{(d)}}{\sqrt{\hat{V}_0^{(d)}}} \right| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right\} & \text{if } \beta_0 = 0 \ , \\ \mathbf{1}\left\{ \left| \frac{\hat{\beta}^{(d)} - \beta_0\left(1 - \hat{B}^{(d)}\right)}{\beta_0 \sqrt{\hat{V}^{(d)}}} \right| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \right\} & \text{otherwise.} \end{cases} \tag{1.6}$$

where Φ is the CDF of the standard normal distribution.

One-sided tests can be constructed by modifying the rejection rule in the usual way. It is immediate that the test is consistent:

Corollary 1.1 (Inference for $\beta^{(1)}$ and $\beta^{(T)}$)**.**

(a) If $\beta^{(d)} = \beta_0$, $E\left[\phi^{(d)}\right] \to \alpha$.

(b) If $\beta^{(d)} \neq \beta_0$, $E\left[\phi^{(d)}\right] \to 1$.

When $\beta^{(d)} \neq 0$, $\hat{\beta}^{(d)}$ needs to be centered by subtracting $\beta^{(d)}\left(1 - \hat{B}^{(d)}\right)$ instead of $\beta^{(d)}$. We will refer to this form of centering as *bias correction* for $\hat{\beta}^{(d)}$. As we explained at the start of Section 1.3, bias correction is necessary for $\hat{\beta}^{(d)}$ to attain a non-degenerate limiting distribution when asymptotic bias is of larger order than variance. Indeed, $\hat{B}^{(d)}/\sqrt{\hat{V}^{(d)}} = O_p\left(p_n^{-1/2}\right)$. As such, if $p_n \prec 1$,division by $\sqrt{\hat{V}^{(d)}}$ blows up $\hat{B}^{(d)}$.

In the bias for $\hat{\beta}^{(T)}$, terms with larger t's dominate those with smaller t's. When p_n is dense enough, terms with small t's may actually much smaller than $\sqrt{\hat{V}^{(T)}}$ so that they can be ignored. With only the stipulation that $p_n \succ n^{-1}$ however, a non-degenerate asymptotic distribution can only be achieved when all terms are included.

Confidence Intervals. Because $\hat{V}_0^{(d)}$ estimates variance only when $\beta^{(d)} = 0$, the usual confidence intervals based on $\hat{V}_0^{(d)}$ need not attain nominal coverage. This failure occurs for two countervailing reasons. Firstly, the quantity $\hat{V}_0$ is meant to estimate,

$$\text{Var}\left(\sum_{i=1}^{n} C_i^{(d)} \varepsilon_i\right) =: V_0^{(d)} \ .$$

However, $V_0^{(d)}$ under-estimates variance of $\hat{\beta}^{(d)}$ when $\beta^{(d)} \neq 0$. That is,

$$\text{Var}\left(\frac{\hat{\beta}^{(d)} - E[\beta^{(d)}]}{\sqrt{V_0}}\right) \to \infty \ .$$

On the other hand, the bias in $\hat{\beta}^{(d)}$ means that

$$\hat{V}_0^{(d)} \approx V_0 + \beta^{(d)} \hat{B}^{(d)} \sum_{i=1}^{n} \left(C_i^{(d)} \right)^2 .$$

The second term in the above equation can be large, such that $\hat{V}_0^{(d)}$ may exceed $\hat{V}$. This turns out to be the case in our application in Section 1.5.

To obtain confidence intervals for $\beta^{(d)}$ consider the following:

Definition 1.11. For $d \in \{1, T\}$ and a given α, let

$$\mathcal{C}_0^{(d)} := \left[\hat{\beta}^{(d)} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}_0^{(d)}} \quad , \quad \hat{\beta}^{(d)} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}_0^{(d)}} \right] . \tag{1.7}$$

Suppose $\hat{\beta}^{(d)} \geq 0$ and let

$$\mathcal{C}^{(d)} := \left[\frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}^{(d)}}} \quad , \quad \frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}^{(d)}}} \right] . \tag{1.8}$$

Finally, let $\mathcal{C}_*^{(d)} = \mathcal{C}_0^{(d)} \cup \mathcal{C}^{(d)}$.

Remark 1.5. If $\hat{\beta}^{(d)} < 0$, the upper bound in the above definition of $\mathcal{C}^{(d)}$ is smaller than the lower bound. In this case,

$$\mathcal{C}^{(d)} := \left[\frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}^{(d)}}} \quad , \quad \frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\hat{V}^{(d)}}} \right] .$$

The following is immediate:

Corollary 1.2 (Confidence Interval)**.** $P\left(\beta^{(d)} \in \mathcal{C}_*^{(d)}\right) \to 1 - \alpha$.

We can obtain one-sided confidence intervals by modifying the bounds as usual. More generally, as long as $\mathcal{C}^{(d)}$ is a $1-\alpha$ confidence interval for $\beta^{(d)} \neq 0$ and $\mathcal{C}_0^{(d)}$ is a $1-\alpha$ confidence interval when $\beta^{(d)} = 0$, their unions will produce a $1-\alpha$ confidence interval for $\beta^{(d)}$ unconditionally. In particular, it is always valid to set $\mathcal{C}_0^{(d)} = \{0\}$. This can be useful when it is not important to exclude 0 from the confidence interval. For example, suppose we want one-sided confidence intervals that upper bounds $\beta^{(d)}$. We can then consider using

$$\mathcal{C}_0^{(d)} = \{0\} \quad \text{and} \quad \mathcal{C}^{(d)} = \left(-\infty \quad , \quad \frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)} - \Phi^{-1}(1-\alpha)\sqrt{\hat{V}^{(d)}}} \right] .$$

If $\hat{\beta}^{(d)} > 0$, $\mathcal{C}_*^{(d)} = \mathcal{C}^{(d)}$. For the reasons discussed above, we can also have

$$\mathcal{C}_*^{(d)} \subsetneq \left(-\infty \quad , \quad \hat{\beta} + \Phi^{-1}(1-\alpha)\sqrt{\hat{V}_0^{(d)}} \right] .$$

As before, such a situation arises in our application (Section 1.5).

Bias Correction. Since the bias of the OLS estimators $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ can be estimated, it is reasonable to consider the following *bias-corrected* estimators:

Definition 1.12 (Bias-Corrected Estimators)**.** For $d \in \{1, T\}$, define

$$\check{\beta}^{(d)} = \frac{\hat{\beta}^{(d)}}{1 - \hat{B}^{(d)}} .$$

Bias-corrected estimators have faster rates of convergence:

Corollary 1.3. Suppose $p_n \succ n^{-1}$. For $d \in \{1, T\}$, $\check{\beta}^{(d)} - \beta^{(d)} = O_p\left(n^{-2} p_n^{-3/2}\right)$.

For reference, $\hat{\beta}^{(d)} - \beta^{(d)} = O_p\left(n^{-1} p_n^{-1}\right)$.

1.3.2.3. Eigenvector Centrality under Proxy Error ($\hat{\beta}^{(\infty)}$). We next consider inference on $\hat{\beta}^{(\infty)}$. Eigenvector centrality can be badly biased under sparsity, which makes inference challenging. However, strategic choices of a_n can overcome many of these issues. We first introduce the following simplifying assumption:

Assumption E2 (Finite Rank)**.** Suppose f has rank $R < \infty$:

$$f(u,v) = \sum_{r=1}^{R} \tilde{\lambda}_r \phi_r(u)\phi_r(v) \ , \tag{1.9}$$

where $\|\phi_r\| = 1$ for all $r \in [R]$ and if $r \neq s$,

$$\int_{[0,1]} \phi_r(u)\phi_s(u)du = 0 \ .$$

Furthermore, suppose that

$$\Delta_{\min} = \min_{1 \geq r \geq R-1} \left| \tilde{\lambda}_r - \tilde{\lambda}_{r+1} \right| > 0 \ .$$

In Equation (1.9), we express f in terms of its eigenfunctions $\{\phi_r\}_{r=1}^{R}$. Assumption E2 implies that the true network has low-dimensional structure and is satisfied by many popular network models, such as the stochastic block model (Holland et al. 1983, also see Example 1.2 below) and random dot product graphs (Young and Scheinerman 2007). This assumption is also commonly found in the networks literature (e.g. Levin and Levina 2019; Li et al. 2020), and the matrix completion literature more generally (e.g. Candès and Tao 2010; Negahban and Wainwright 2012; Chatterjee 2015; Athey et al. 2021). Importantly, existing papers on inference with eigenvectors (Le and Li 2020; Cai et al. 2021) also make this assumption. Note also that Assumption E2 implies Assumption E1.

Example 1.2 (Stochastic Block Model)**.** The Stochastic Block Model (SBM) is one of the earliest statistical models of networks. It assumes that individuals fall into groups $g \in \{1, ..., B\}$ and that the true network depends only on group membership. For example, suppose that a classroom has two groups: jocks, nerds. The SBM posits that the strength of the tie between any jock and any nerd are the same. Analogously for that between any two jocks or any two nerds, though all three ties can be of different intensity. Suppose the proportion of each group is π_g and that the link probability is $p_{g,g'} = p_{g,g'}$. Then the graphon is a step-function on $[0, 1]^2$ with B^2-steps and rank B. It is visualized in Figure 1.3.

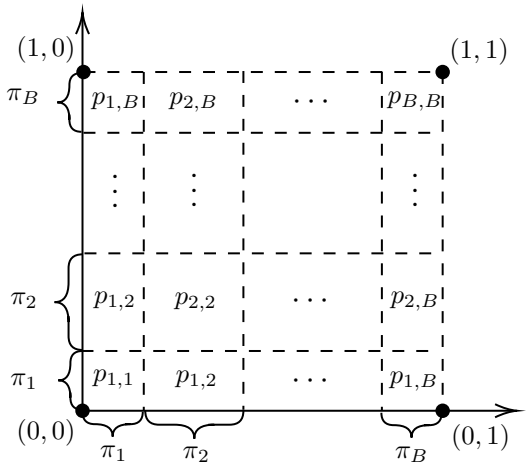

Figure 1.3. The graphon f of a stochastic block model with B blocks. f is a step-function with B^2 steps and is of rank B.

With the low-rank asumption, we can consider the asymptotic distribution of $\hat{\beta}^{(\infty)}$ in a few cases.

Theorem 1.6 (Inference – Eigenvector)**.** Suppose Assumptions 1.1, 1.2, 1.3 and E2 hold.

(a) Suppose either:

(i) $\beta^{(\infty)} = 0$, or,

(ii) $p_n \succ n^{-1} \log n$ and $a_n \prec (np_n)^{3/2}$, or,

(iii) For some $\eta > 0$,

$$p_n \succ n^{-1} \left(\frac{\log n}{\log \log n} \right)^{\frac{1}{2}+\eta} . \tag{1.10}$$

and $a_n \prec np_n$.

Then,

$$\hat{S}_0^{(\infty)} := \frac{\hat{\beta}^{(\infty)} - \beta^{(\infty)} \left(1 - \hat{B}^{(\infty)}\right)}{\sqrt{\hat{V}_0^{(\infty)}}} \xrightarrow{d} \mathrm{N}(0,1) . \tag{1.11}$$

where

$$\hat{V}_0^{(\infty)} = \left(\sum_{i=1}^{n} \left(\hat{C}_i^{(\infty)} \right)^2 \right)^{-2} \sum_{i=1}^{n} \left(\hat{C}_i^{(\infty)} \right)^2 \left(\hat{\varepsilon}_i^{(\infty)} \right)^2 \quad , \quad \hat{B}^{(\infty)} = \left(\lambda_1(\hat{A}) \right)^{-1} .$$

In the above, $\hat{\varepsilon}_i^{(\infty)}$ is as defined in Equation (1.3).

(b) Suppose $p_n \succ \frac{1}{\sqrt{n}}$, $a_n \succ n\sqrt{p_n}$ and $\beta^{(\infty)} \neq 0$. Then,

$$\hat{S}^{(\infty)} := \frac{\hat{\beta}^{(\infty)} - \beta^{(\infty)} \left(1 - \hat{B}^{(\infty)}\right)}{\sqrt{\hat{V}^{(\infty)}}} \xrightarrow{d} \mathrm{N}(0,1) ,$$

where

$$\hat{V}^{(\infty)} = 2 \left(\lambda_1(\hat{A}) \sum_{i=1}^{n} \left(\hat{C}_i^{(\infty)} \right)^2 \right)^{-2} \sum_{j \neq i} \hat{A}_{ij} \left(\left(\hat{C}_i^{(\infty)} \right)^2 + \left(\hat{C}_j^{(\infty)} \right)^2 \right) .$$

Note that the statistics above do not require R or p_n to be specified. This is useful since estimating R may be challenging in addition to p_n being unidentified (Bickel et al. 2011).

Our result describes the asymptotic distribution $\hat{\beta}^{(\infty)}$, which depends on $\beta^{(\infty)}$ and a_n. Case (a) gives conditions under which inference with hc/robust t-statistic is appropriate. As with $\beta^{(1)}$ and $\beta^{(\infty)}$, the usual test works if $\beta^{(\infty)} = 0$. However, it also works if $\beta^{(\infty)} \neq 0$ provided that a_n is small. On the other hand, if a_n is large, case (b) suggests that we get behavior that is more in line with that of $\beta^{(1)}$ and $\beta^{(\infty)}$ when target parameters are non-zero. However, to obtain the result in case (b), we require very strong conditions on p_n due to greater difficulty in controlling the behavior of estimated eigenvector, as the discussion following Theorem 1.5 explains.

When $\beta^{(\infty)} \neq 0$, the differences in case (a) and (b) arise because a_n controls the relative sizes of network proxy error and regression error. The latter dominates if a_n is sufficiently small and has the advantage of being easy to characterize. Hence, in the absence of compelling reasons for choosing a_n to be other values, researchers can consider choosing a_n for statistical convenience. In particular, if a_n is chosen so that case (a) obtains, then usual hc/robust variance estimator based t-statistic and confidence interval have the expected properties. We propose such an a_n below. However, we stress that a smaller a_n also implies a lower rate of convergence. In effect, we are changing the model from one with faster but unknown rate of convergence, to one with a rate that is slower but estimable.

Finally, our result here suggests the use of the bias-corrected estimator, as with degree and diffusion centrality:

$$\check{\beta}^{(\infty)} = \frac{\hat{\beta}^{(\infty)}}{1 - \hat{B}^{(\infty)}} \ .$$

Choice of a_n. The following data-dependent choice is convenient:

Corollary 1.4. Suppose Assumptions 1.1, 1.2, 1.3 and E2 hold. Suppose also that $a_n = \sqrt{\lambda_1(\hat{A})}$ is estimated. If p_n satisfies Equation (1.10),

$$\frac{\hat{\beta}^{(\infty)} - \beta^{(\infty)}}{\sqrt{\hat{V}^{(\infty)}}} \xrightarrow{d} \mathrm{N}(0,1) \ .$$

Since $\lambda_1(\hat{A}) \approx \tilde{\lambda}_1 n p_n$, the above choice of a_n satisfies the conditions in case (a)(iii) of Theorem 1.6, accommodating close to the maximum possible amount of sparsity for $\hat{\beta}^{(\infty)}$. It also obviates the need for bias correction since the bias is always of lower order than the variance at this rate. Furthermore, such an a_n has intuitive appeal since it implies that

$$\hat{C}^{(\infty)} \left(\hat{C}^{(\infty)}\right)' = \lambda_1(\hat{A}) v_1(\hat{A}) \left(v_1(\hat{A})\right)' = \underset{\mathrm{rank}(M)=1}{\arg\min} \left\| M - \hat{A} \right\|_2 \ .$$

In words, scaling the estimated eigenvectors to $\sqrt{\lambda_1(\hat{A})}$ means that the outer product of $\hat{C}^{(\infty)}$ is the best rank-1 approximation of $\hat{A}$. In proposing eigenvector centrality, Bonacich (1972) in fact cites this property as one of the key motivations, arguing that $\sqrt{\lambda_1(\hat{A})} v_1(\hat{A})$ can be interpreted as the "social interaction potential" of a given agent.

It is also common for applied researchers to scale eigenvector centrality by its standard deviation (e.g. Banerjee et al. 2019; Chandrasekhar et al. 2018). This is typically done so that the regression coefficient can be interpreted as the effect on outcome of a one

standard-deviation increase in eigenvector centrality. In effect, this procedure sets $a_n \approx \sqrt{n}$:

Corollary 1.5. Suppose Assumptions 1.1, 1.2, 1.3 and E2 hold. Suppose also that

$$a_n = \left(\frac{1}{n} \sum_{i=1} \left([v_1(\hat{A})]_i - \frac{1}{n} \sum_{i=1} [v_1(\hat{A})]_i \right)^2 \right)^{-1/2} .$$

Then

$$a_n = (1 + o_p(1)) \cdot \frac{\sqrt{n}}{\sqrt{1 - E\left[\phi_1(U_1)\right]^2}} .$$

As such, Case (a) (ii) of Theorem 1.6 applies when $p_n \succ n^{-2/3}$.

Note that with this choice of a_n, our theorem is able to accommodate less sparsity that if $a_n = \sqrt{\lambda_1(\hat{A})}$.

Remark 1.6. Le and Li (2020) provides methods for testing the hypothesis

$$\|\beta^{(\infty)} C^{(\infty)}\|^2 = 0$$

when $p_n \succ n^{-1/2}$. They accommodate regressions on multiple eigenvectors, but in a setting with only one eigenvector, their result asserts that the t-statistic with the homoskedastic variance estimator can be used to test the hypothesis that $\beta^{(\infty)} = 0$. Theorem 1.6 does not cover regression on multiple eigenvectors but it accommodates greater sparsity and facilitates tests of $\beta^{(\infty)} = \beta_0$ for $\beta_0 \neq 0$.

Remark 1.7. To compare our results to that of Cai et al. (2021), set $a_n = \sqrt{n}$. The condition in Case (a) (ii) of Theorem 1.6 specializes to $p_n \succ n^{-2/3}$. Cai et al. (2021) can accommodate $p_n \succ n^{-1}$ but they assume proxy error that is additive and i.i.d. Gaussian.

	With Proxy Error			No Error
	$\beta^{(1)}/\beta^{(T)}$	$\beta^{(\infty)}$ (Case 1.6(b))	$\beta^{(\infty)}$ (Case 1.6(a))	$\beta^{(1)}/\beta^{(T)}/\beta^{(\infty)}$
$H_0 : \beta^{(d)} = 0$	t-test	t-test	t-test	t-test
$H_0 : \beta^{(d)} = \beta_0 \neq 0$	Def. 1.10	Def. 1.10		
Conf. Intervals	Def. 1.11	Def. 1.11	t-stat based	t-stat based

Table 1.1. Summary of inference procedures. The hc/robust t-test is appropriate for all $\beta^{(d)}, d \in \{1, T, \infty\}$ for the null hypothesis that $\beta^{(d)} = 0$. It is also appropriate for non-zero null hypotheses (1) for all $\beta^{(d)}$ if there is no proxy error, or (2) for $\beta^{(\infty)}$ in the presence of proxy error if a_n satisfies the conditions in Theorem 1.6 case (a). Whenever the t-test is appropriate, the t-statistic based confidence intervals are also valid. In all other cases, refer to Definition 1.10 for testing and Definition 1.11 for confidence intervals.

1.4. Simulations

In this section, we present simulation evidence to support our theory. We will consider the unobserved adjacency matrix A defined as

$$A_{ij} = \begin{cases} p_n & \text{if } i \neq j \ , \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the graphon is $f = 1$. The observed adjacency matrix is $\hat{A}$, where for $i > j$, $\hat{A}_{ij} = \text{Bernoulli}(A_{ij})$. $\hat{A}_{ii} = 0$, $\hat{A}_{ji} = \hat{A}_{ij}$.

Our regression model is:

$$Y_i = \beta C_i^{(d)} + \varepsilon_i^{(d)} \ ,$$

where $C_i^{(d)}$ are centrality measures calculated on A. In this simulation, we draw $U_i \overset{\text{i.i.d.}}{\sim} U[0,1]$, $\varepsilon_i^{(d)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$, where $\varepsilon_i^{(d)} \perp\!\!\!\perp U_i$ and $\varepsilon_i^{(d)} \perp\!\!\!\perp \hat{A}_{jk}$ for all $i,j,k \in [n]$. We will set $\beta = 1$. As before, $\tilde{\beta}^{(d)}$ is used when A is observed, and $\hat{\beta}^{(d)}$ is used when only $\hat{A}$ is observed.

We revisit our three sets of results in turn: inconsistency under sparsity, bias correction and normal approximation.

1.4.1. Inconsistency Under Sparsity

The first regime of interest is $p_n = 1/n$. Theorem 1.1 asserts that $\tilde{\beta}^{(1)}$ and $\tilde{\beta}^{(T)}$ are consistent. $\tilde{\beta}^{(\infty)}$ is consistent provided that $a_n \to \infty$ for eigenvector centrality.

We start with the last claim, which is supported by Figure 1.4. For $n = 100$, we see that the choice of scaling clearly affects how well the estimator is able to concentrate around $\beta = 1$. The plots for larger values of n are qualitatively similar. This also hints at the trade-off that made in Theorem 1.6: we can choose $a_n = \sqrt{\lambda_1(\hat{A})}$ so that the distribution of $\hat{\beta}^{(\infty)}$ is easy to characterize, but this will slow down the rate of convergence. Since this subsection concerns current practice, we will set $a_n = \sqrt{n}$ for its remainder.

We return to the first claim concerning consistency of $\tilde{\beta}^{(d)}$ when $p_n = 1/n$. Figure 1.5 indeed shows the distribution of $\tilde{\beta}^{(d)}$ for each n. The estimators concentrate around β as n increases, in line with our result. However, Theorem 1.2 asserts that $\hat{\beta}^{(1)}$, $\hat{\beta}^{(T)}$ and $\hat{\beta}^{(\infty)}$ are all inconsistent when $p_n = 1/n$. Their distributions, presented in Figure 1.6, concords

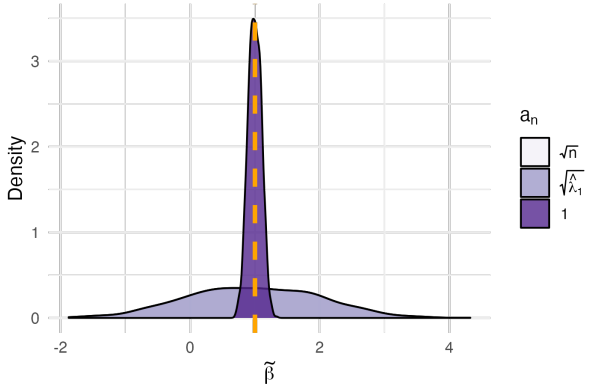

Figure 1.4. Distribution of $\tilde{\beta}^{(\infty)}$ for $n = 100$, $p_n = 1/n$ under various a_n. $\beta = 1$ (orange dashed line).

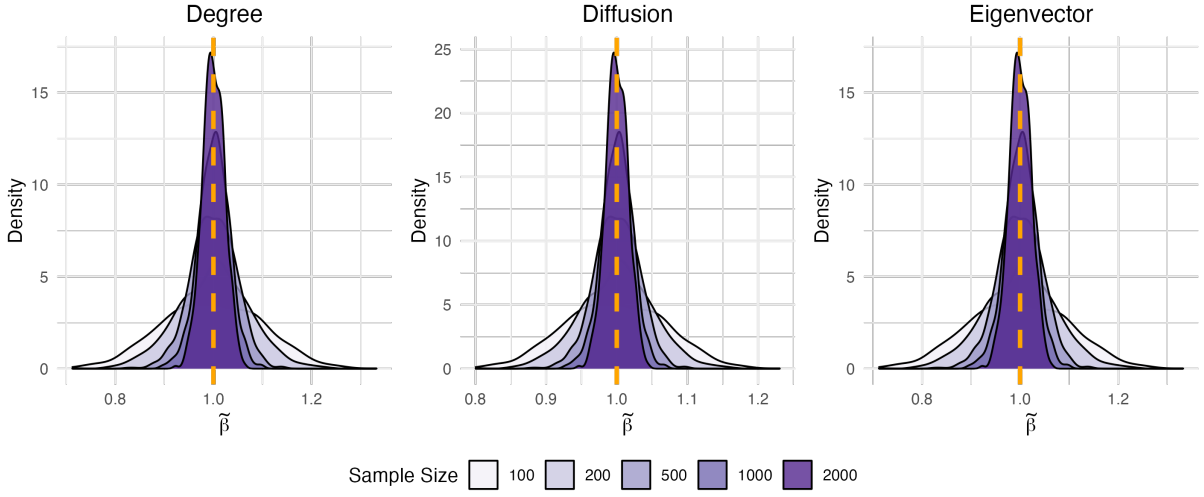

Figure 1.5. Distribution of $\tilde{\beta}^{(d)}$ for $p_n = 1/n$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

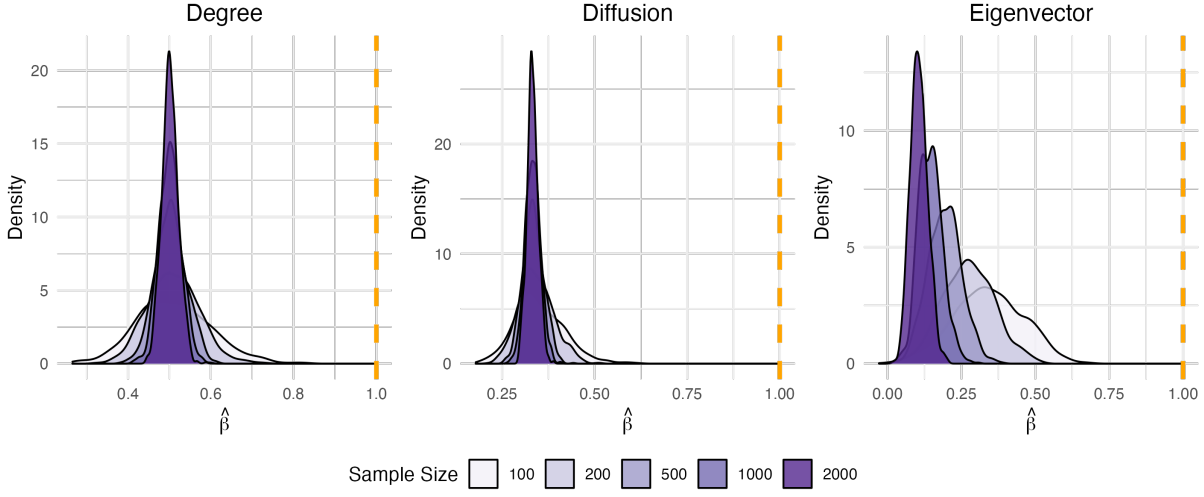

Figure 1.6. Distribution of $\hat{\beta}^{(d)}$ for $p_n = 1/n$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

with our result. Indeed, we see that $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ are attenuated by constant amount as $n \to \infty$, while $\hat{\beta}^{(\infty)}$ converges in probability to 0.

Finally, we consider the case when $p_n = n^{-1}\sqrt{\frac{\log n}{\log\log n}}$. In this regime, $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ are consistent but $\hat{\beta}^{(\infty)}$ is not. We see suggestive evidence of this in Figure 1.7, where $\hat{\beta}^{(\infty)}$ is drifting further away from β as n increases. The opposite occurs with $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$. Though the rate of convergence is slow, it is visible.

1.4.2. Bias Correction

Even in regime dense enough such that $\hat{\beta}^{(1)}$, $\hat{\beta}^{(T)}$ and $\hat{\beta}^{(\infty)}$ are consistent, they can still be subject to biases that affect their rates of convergence. This motivates the bias-corrected estimators in Definition 1.12. In this subsection, we study the effects of bias correction in the regime $p_n = 1/\sqrt{n}$.

Figure 1.8 shows the distribution of the estimators when $n = 500$. Here $a_n = \sqrt{\hat{\lambda}_1(\hat{A})}$. We see that bias correction is effective in correctly centering $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$. The same is true for $\hat{\beta}^{(\infty)}$ though to a smaller extent, in line with claims in Corollary 1.4. Results for other values of n are qualitatively similar.

1.4.3. Distributional Theory

Finally, we investigate the quality of the normal approximations proposed in Theorems 1.5 and 1.6. As before, we consider the regime $p_n = 1/\sqrt{n}$. Figure 1.9 presents the distribution of test statistics (in purple) which our theorems predict have the standard normal distribution (in gray). We see that the two distributions are indeed close. It is also common for applied researcher to compute the usual t-statistic with heteroskedaticity

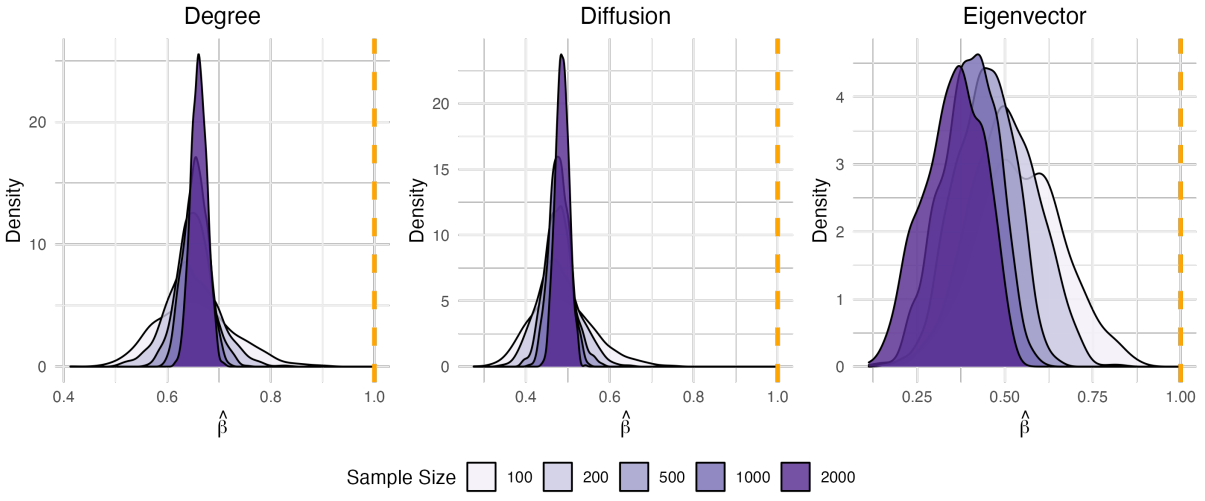

Figure 1.7. Distribution of $\hat{\beta}^{(d)}$ for $p_n = n^{-1}\sqrt{\log n / \log\log n}$. For $\tilde{\beta}^{(\infty)}$, $a_n = \sqrt{n}$. $\beta = 1$ (orange dashed line).

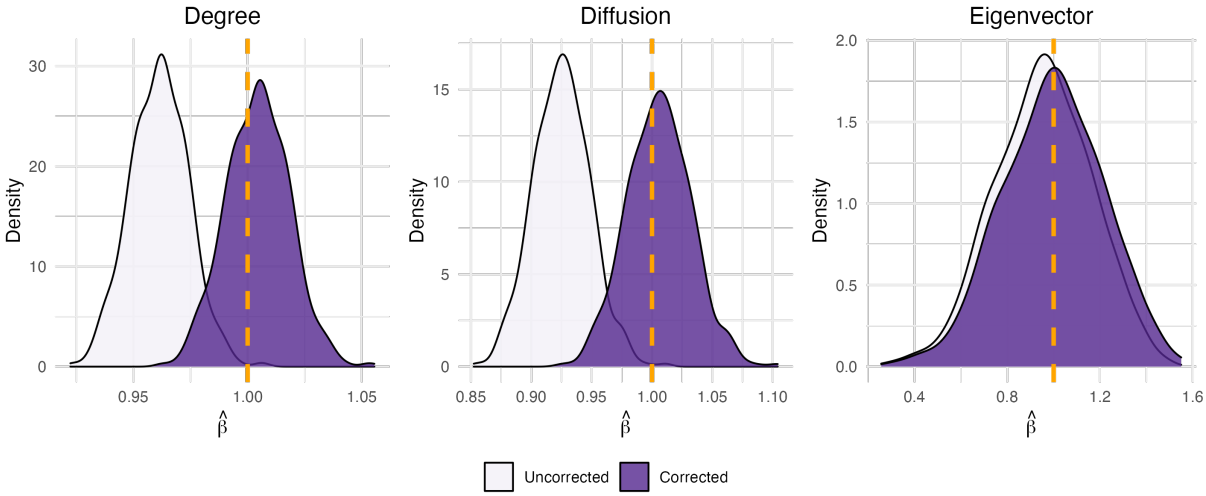

Figure 1.8. Distributions of $\hat{\beta}^{(d)}$ and their bias corrected versions $\check{\beta}^{(d)}$ for $p_n = 1/\sqrt{n}$. $\beta = 1$ (orange dashed line).

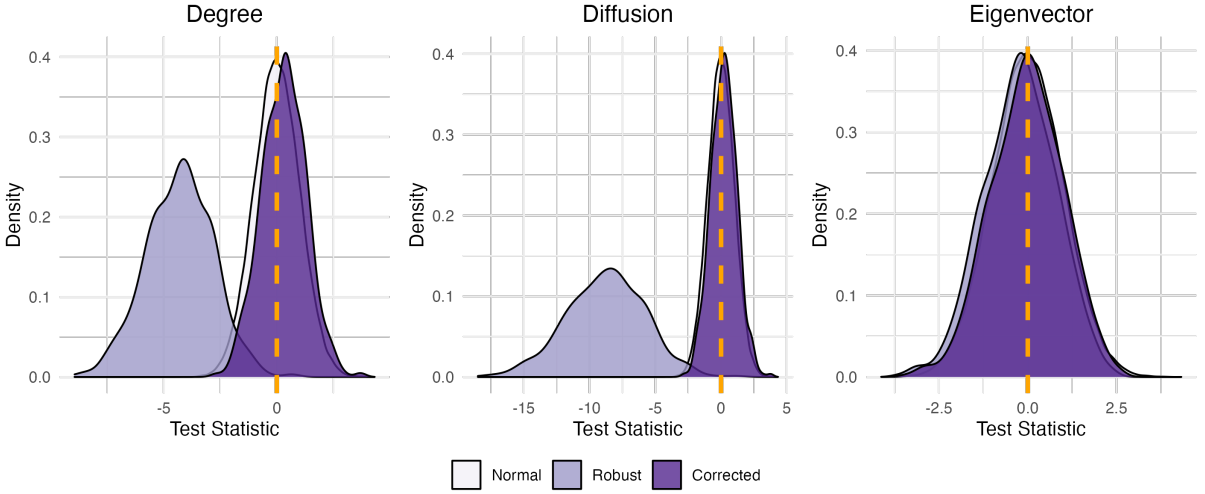

Figure 1.9. Distribution of the centered and scaled test statistics in Theorems 1.5 and 1.6. Robust refers to tests based on t-statistic with robust (heteroskedasticity consistent) standard errors.

consistent (robust) standard errors and conduct inference under the assumption that it has a standard normal distribution. For comparison, we include the distribution of the t-statistic (in lavender). Corollary 1.4 justifies the use of this statistic when $a_n = \sqrt{\hat{\lambda}_1(\hat{A})}$ but our theory for degree and diffusion centralities is based on a different statistic. Indeed, we see that the robust t-statistic can be quite far from the standard normal distribution in both location and dispersion. This suggests that our method would lead to more reliable tests.

We next examine the size and power of tests based on our distributional theory. We consider testing the hypothesis $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ at 5% level of significance. Table 1.2 presents size of the test when $\beta = 1$ is correctly specified. For degree and diffusion centralities, our theory provides test statistics which differ from the robust t-statistic. As we see from the table, the tests control size well. Tests for degree and diffusion centralities that are based on the robust t-statistic has Type I error over 50% across all sample sizes. For eigenvector centrality, our theory predicts that the robust t-statistic will perform well Indeed, it has size close to 5%. We also consider testing the hypothesis $\beta = 0$. Power for this test is presented in Table 1.3. For this null hypothesis, our theory suggests the use of the robust t-statistic. Reassuringly, the tests all have power close to 1. To understand how power changes as we vary the alternative hypothesis, we hone in on the case where $n = 500$ and $p_n = 1/\sqrt{n}$. Figure 1.10 presents the rejection probability of our test under various alternatives. We see that the our tests control size and have good power. Comparatively, tests based on the robust t-statistic have poor size control when $\beta \neq 0$. Furthermore, they can have poor power against particular alternatives owing to

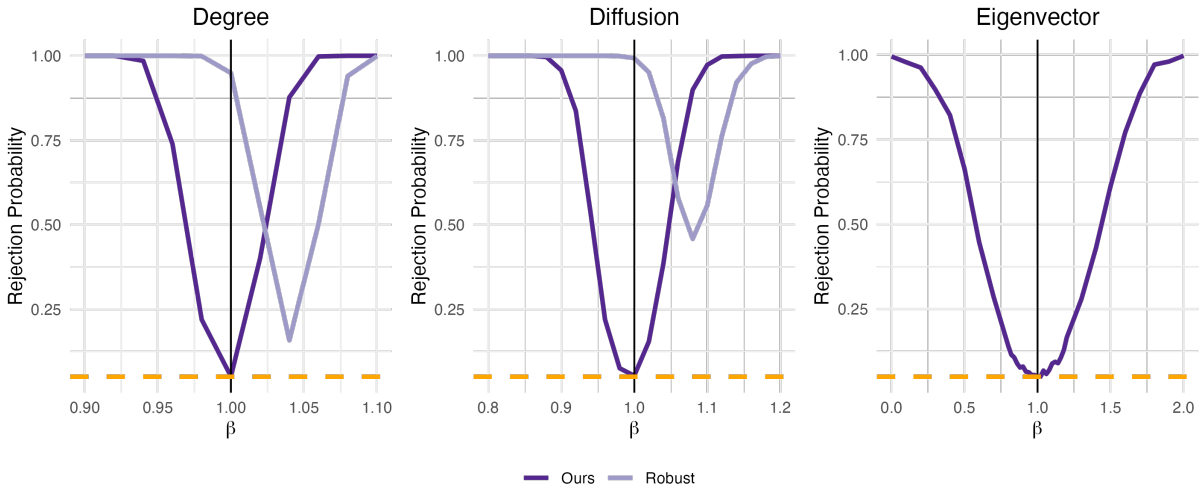

Figure 1.10. Power of the two-sided test of $H_0 : \beta = 1$ under various alternatives. Test at 5% level of significance (orange dashed line).

the bias. We conclude that our tests have desirable properties and are preferred to the test with robust t-statistic when networks are sparse and observed with noise.

1.5. Empirical Demonstration

In this section, we demonstrate the relevance of our theoretical findings via an application inspired by De Weerdt and Dercon (2006).[2] In the developing world, social insurance is an important mechanism for smoothing consumption, because of restricted access to formal credit markets (Rosenzweig 1988; Udry 1994; Fafchamps and Lund 2003; Kinnan and Townsend 2012, among many others). De Weerdt and Dercon (2006) examines the case of Nyakatoke, a village with 120 households in rural Tanzania, and find that social insurance helps households to smooth consumption following health shocks. The data they use comprises five rounds of panel data on household consumption, illness among

[2]The data is obtained from Joachim De Weerdt's website: https://www.uantwerpen.be/en/staff/joachim-deweerdt/public-data-sets/nyakatoke-network/.

p_n	Statistic		Sample Size				
			100	200	500	1000	2000
0.1	Degree	Ours	0.055	0.052	0.067	0.062	0.065
		Robust	0.656	0.673	0.690	0.668	0.674
	Diffusion	Ours	0.049	0.053	0.064	0.059	0.060
		Robust	0.889	0.894	0.887	0.871	0.898
	Eigenvector		0.045	0.043	0.037	0.056	0.044
$n^{-1/3}$	Degree	Ours	0.066	0.065	0.067	0.058	0.065
		Robust	0.330	0.450	0.573	0.705	0.783
	Diffusion	Ours	0.080	0.070	0.074	0.057	0.064
		Robust	0.645	0.734	0.813	0.888	0.934
	Eigenvector		0.045	0.042	0.051	0.042	0.058
$n^{-1/2}$	Degree	Ours	0.072	0.049	0.051	0.037	0.062
		Robust	0.659	0.801	0.949	0.993	0.999
	Diffusion	Ours	0.071	0.045	0.053	0.037	0.059
		Robust	0.881	0.948	0.993	1.000	1.000
	Eigenvector		0.077	0.045	0.050	0.050	0.047

Table 1.2. Size of 5% level two-sided tests when $\beta = 1$ is correctly specified. Robust refers to tests based on t-statistic with robust (heteroskedasticity consistent) standard errors.

other covariates, collected from February to December 2000. The authors also had access to social network data collected during the first round of the survey, in which households were asked for the identities of those who they depend on or depend on them for help. The authors then regress a household's change in consumption following illness on the mean consumption of their network neighbors, finding evidence of positive co-movements.

Another way to demonstrate the effect of social insurance on consumption smoothing could be to regress variance in consumption on network centrality measures. Specifically,

p_n	Statistic	Sample Size				
		100	200	500	1000	2000
0.1	Degree - Robust	1.000	1.000	1.000	1.000	1.000
	Diffusion - Robust	1.000	1.000	1.000	1.000	1.000
	Eigenvector	0.845	0.995	1.000	1.000	1.000
$n^{-1/3}$	Degree - Robust	1.000	1.000	1.000	1.000	1.000
	Diffusion - Robust	1.000	1.000	1.000	1.000	1.000
	Eigenvector	0.998	1.000	1.000	1.000	1.000
$n^{-1/2}$	Degree - Robust	1.000	1.000	1.000	1.000	1.000
	Diffusion - Robust	1.000	1.000	1.000	1.000	1.000
	Eigenvector	0.832	0.947	0.994	1.000	1.000

Table 1.3. Power of 5% level two-sided tests of $H_0 : \beta = 0$ when $\beta = 1$. Under this H_0, the our test statistics is the usual t-statistic with robust (heteroskedasticity-consistent) standard errors.

the regression:

$$Y_i = \beta C_i^{(d)} + \varepsilon_i^{(d)}$$

where Y_i is variance in food expenditure and $C_i^{(d)}$ is a centrality measure. The above regression could be preferred to the authors' specification if we are unsure about the covariates that reflect social assistance. For example, it might be a household's stock of savings that co-move with the decision to lend to their friends, rather than their own consumption. We might also be interested in more complex patterns of assistance, which could be summarized in an appropriate centrality measure, but which might not be tractable with covariates.

The above regression requires information on network of social insurance, in which each entry A_{ij} records the probability i lends money to j over the survey period. We can consider obtaining proxies for this network using one of the following:

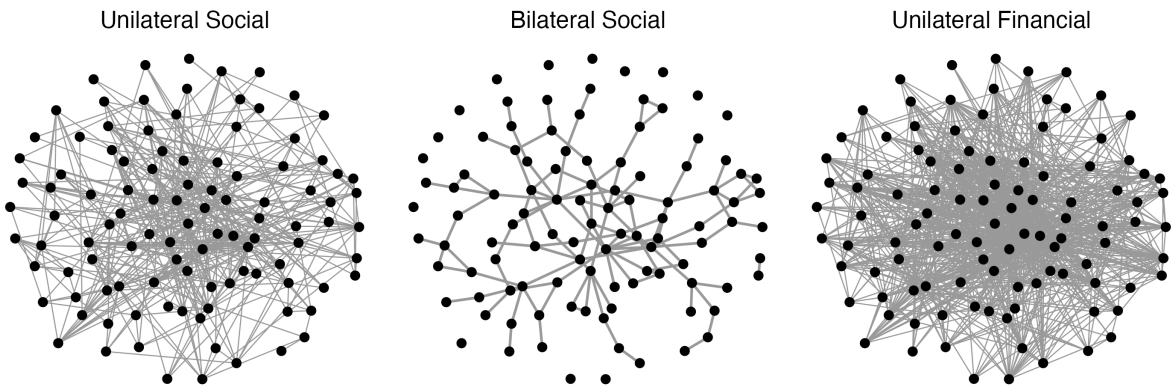

Figure 1.11. Social and Financial networks in Nyakatoke.

Unilateral Social (US). $\hat{A}_{ij} = 1$ if either i or j names the other household as a party that they could depend on or which depends on them for help.

Bilateral Social (BS). $\hat{A}_{ij} = 1$ only if both i and j names the other household as a party that they could depend on or which depends on them for help.

Unilateral Financial (UF). $\hat{A}_{ij} = 1$ if either i or j lends money to the other at least once over the survey period.

The authors study US and BS. We also consider UF since self-reported loan data is available. The networks are plotted in Figure 1.11 and the degree distributions are described in Table 1.4. By construction, BS is much sparser than US. Due to the availability of five panels, UF is denser than the other two. With $n = 120$ households, $\sqrt{n} = 11$. We might therefore be concerned that $np_n \prec \sqrt{n}$ especially for US and BS. Note that US and UF are connected. BS is disconnected but has a clear giant component. Our rule-of-thumb suggests that $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$ should across all three networks. $\hat{\beta}^{(\infty)}$ should perform well on US and UF, but likely not on BS.

$(n = 119)$	Mean	Median	Min	Max
Unilateral Social	8.02	7	1	31
Bilateral Social	2.30	2	0	10
Unilateral Financial	16.53	14	3	79

Table 1.4. Degree distributions of various networks in Nyakatoke.

Regression results are presented in Table 1.5. In this exercise, $a_n = \sqrt{\lambda_1(\hat{A})}$, $\delta = 1/\sqrt{\lambda_1(\hat{A})}$ and $T = 2$. We first note that estimated attenuation factor is the smallest (i.e. furthest from 1) in the sparsest network BS. This is in line with our result that bias is $O_p(n^{-1}p_n^{-1})$. Diffusion centrality is generally estimated to be less attenuated, because δ is small (≈ 0.2). As the last column shows, bias correction can lead to substantially different estimates. Table 1.5 also presents p-values for tests of the two-sided hypothesis that $\beta^{(d)} = 0$. Centrality statistics on BS appears to be more predictive of the variance in food consumption than that on US and UF. This highlights that researchers should not choose their network proxies on the criteria of sparsity alone. In the case of Nyakatoke, evidence suggests that US reflects "desire to link", rather than actual risk-pooling (Comola and Fafchamps 2014), such that US is a noisier proxy than BS. By the same account, the large number of discrepancies between reporting by borrowers and lenders of the same loan suggests that the loan data is subject to severe mis-reporting (Comola and Fafchamps 2017), rendering it an equally poor proxy. Among the centrality statistics on BS, eigenvector has the least predictive power by far, in line with what is suggested by our rule-of-thumb. We reiterate our warning that eigenvector centrality is less robust to sparsity than degree and diffusion, such that the p-values might reflect the poor statistical properties of the measure, rather than its lack of economic significance.

		Estimate	p-value	Atten.	Bias Corr.
Unilateral Social	Degree	-1064	0.67	0.91	-1172
	Diffusion	-4274	0.77	1.00	-4290
	Eigenvector	-12353	0.86	0.91	-13548
Bilateral Social	Degree	-11604	0.06	0.74	-15592
	Diffusion	-23672	0.16	0.95	-24883
	Eigenvector	-10543	0.93	0.78	-13434
Unilateral Financial	Degree	-412	0.70	0.96	-429
	Diffusion	-4559	0.74	1.00	-4561
	Eigenvector	-15040	0.77	0.96	-15699

Table 1.5. Regression results for various networks. Estimate is $\hat{\beta}^{(d)}$. p-value is for the two-sided test that $H_0 : \beta = 0$. Atten. is the estimated attenuation factor of $\hat{\beta}^{(d)}$ (i.e. $1 - \hat{B}^{(d)}$). Bias Corr. presents the bias corrected estimates, $\check{\beta}^{(d)}$.

Finally, we present one-sided confidence intervals for values of $\beta^{(d)}$ based on our results. These are useful for putting bounds on parameter values. In our example, a lower bound could be intuitively interpreted as the limits to informal risk-sharing, a quantity which could be useful for policymakers deciding whether or not to provide agricultural insurance. We focus on BS since it appears to be the only informative network. Results for degree and diffusion are presented in Table 1.6. Results for eigenvector are omitted since our theory is based on the usual robust t-statistic. Our confidence intervals leads to tighter lower bound than the those based on the robust t-statistic. Furthermore, as we increase the desired coverage, our confidence intervals increase much more slowly than the hc/robust confidence intervals. This is because the latter is linear in $\Phi^{-1}(1 - \alpha/2)$, whereas this term appears in the denominator of $\mathcal{C}^{(d)}$ (as in Definition 1.8).

		90%	95%	99%
Degree	Robust	(-19500, ∞)	(-21700, ∞)	(-25900, ∞)
	Ours	(-18800, ∞)	(-20000, ∞)	(-22700, ∞)
Diffusion	Robust	(-45000, ∞)	(-51000, ∞)	(-62400, ∞)
	Ours	(-25200, ∞)	(-25300, ∞)	(-25500, ∞)

Table 1.6. One-sided confidence intervals for degree and diffusion.

1.6. Conclusion

In this chapter, we studied the properties of linear regression on degree, diffusion and eigenvector centrality when networks are sparse and observed with error. We show that these issues threaten the consistency of OLS estimators and characterize the amount of sparsity at which inconsistency occurs. In doing so, we find that eigenvector centrality is less robust to sparsity than the others and that the statistical properties of the corresponding regression is sensitive to the scaling.

Additionally, we show that an asymptotic bias arises whenever the true slope parameter is not 0 and that the bias can be of larger order than the variance, so that bias correction is necessary to obtain a non-degenerate limiting distribution from the OLS estimator. Finally, we provide estimators for the bias and variance which, together with our central limit theorem, facilitates inference under sparsity and proxy error. We confirm our theoretical results via simulations, which suggest that our approximation result works better for estimation and inference when networks are sparse, particularly when compared to the use of robust standard errors and the associated t-statistics. Finally, we demonstrate the relevance of our theoretical results by studying the social insurance network in Nyakatoke, Tanzania.

In sum, our results suggest that applied researchers view their results with caution when applying OLS to sparse, proxy networks. Specifically, comparing the statistical significance of eigenvector centrality with degree or diffusion may yield misleading conclusions since they differ not only in economic significance but also statistical properties. Provided that the networks are not *too* sparse, the usual t-test is valid for null hypothesis that the slope parameter is 0. However, alternative inference procedures will be necessary for other null hypotheses. Additionally, there may be scope for improving estimation by the use of bias-corrected estimators. Estimation and inference under extreme sparsity remains an open question, though as Le et al. (2017) and Graham (2020b) show, parametric models may point to a way forward.

CHAPTER 2

A Worst-Case Randomization Test for the Level of Clustering

2.1. Introduction

Consider the following regression:

$$Y_i = X_i'\beta + U_i \quad , \quad E[X_i U_i] = 0 \ .$$

where a researcher wants to perform inference on β. If the researcher is concerned about correlation between U_i and $U_{i'}$, it is frequently helpful to group observations into independent clusters. These independent clusters can then be used to construct cluster-robust covariance estimators (CCE) as in Liang and Zeger (1986a), or for approximate randomization tests as in Canay et al. (2017a) and Cai et al. (2021).

However, these procedures require the assignment of units to clusters be known ex ante. In practice, researchers often have some freedom in choosing the level at which to cluster their standard errors. For example, those working with the American Community Survey (ACS) can cluster their data either at the individual, county or state level. Alternatively, those working with firm data from COMPUSTAT have the option to cluster at either the 4-digit, 3-digit or 2-digit Standard Industrial Classification (SIC) level, or even the firm level.

Clustering at the correct level is important for valid inference. A large body of simulation evidence shows that ignoring cluster dependence – in other words, clustering at

too fine a level – leads to type I errors that exceed the nominal error by as much as 10 times (Bertrand et al. 2004; Cameron et al. 2008a). On the other hand, clustering at excessively coarse levels can also lead to problems. For one, coarse clusters tend to be few in number. It is well-known that confidence intervals based on the cluster-robust standard errors tend to under-cover when the number of clusters is small (see Angrist and Pischke 2008 for instance), leading to poor size control. In the absence of under-coverage issues, unnecessarily coarse levels of clustering can also lead to tests with poor power since the researcher assumes less information than they actually have. Abadie et al. 2017 demonstrate via simulations, in a many-cluster setting, that CCEs based on coarse clusters can be too large. They also provide theoretical results in this vein, though they do so in the context of their "design-based" asymptotics that differ from those traditionally used to analyze clustered standard errors. Nonetheless, the problems with tests based on excessively coarse-clustering arise even with few clusters – the setting of interest for this chapter. We present a simple simulation to demonstrate these issues in Appendix B.2.

Given the above considerations, a researcher may choose to cluster at a fine level (e.g. individual or county) even when a coarse level of clustering (e.g. state), which is known to be valid, is also available. Nonetheless, they may be unsure if the fine level is appropriate. That is, whether observations across the fine clusters are approximately independent.

To help researchers assess the validity of their chosen clusters, we propose a modified randomization test that can be used as a robustness check for a given clustering specification. Our test requires the number of observations in each (fine) sub-cluster to tend to infinity, but is justified under asymptotics that take the number of (coarse) clusters and (fine) sub-clusters as fixed. Inference is difficult in this setting because scores are not

independent across sub-clusters even asymptotically, as we will explain in Section 2.2.2. Randomization tests, which typically require some type of asymptotic independence, thus cannot be directly applied. We get around this problem by searching for worst-case values of the unobserved parameters to guard against over-rejection. We describe a simple method to search for this value, so that the computational complexity of the test is of the same order as the number of sub-clusters. This is reasonable since our test is targeted towards applications with few sub-clusters. Our test has no power against negative correlation. However, ignoring negative correlation leads to variance estimators that are too large, and is thus less of an issue if the researcher is concerned about size control when performing inference on β.

To our knowledge, there are two other tests for the level of clustering. MacKinnon et al. (2020) proposes a test based on having large number of coarse clusters, relying on the wild bootstrap to improve finite sample performance. Meanwhile, Ibragimov and Müller (2016) proposes a test for the case when there are many sub-clusters. Our test, which takes the number of clusters and sub-clusters to be fixed, handles a more challenging situation, though this comes at the cost of being conservative, especially in settings with homogeneous clusters. However, as our simulations in Section 2.3 show, it has competitive power given heterogeneous clusters – a setting that could be relevant for empirical work. Indeed, our test detects correlation in the clusters chosen by Gneezy et al. (2019), demonstrating its potential usefulness in applied work (see Section 2.4). Finally, we note that the test of Ibragimov and Müller (2016) also has no power against negative correlation, although that of MacKinnon et al. (2020) does not share this limitation.

Abadie et al. (2017) takes a different approach to this issue. They argue for a "design-based" perspective on clustering, requiring researchers to determine ex ante the uncertainty that they face in either sampling or treatment assignment. For example, if the researcher believes that in their specific context, treatment assignment occurs at the sub-cluster level, then sub-clusters should be used for computing standard errors, regardless of whether or not residuals are correlated across the sub-clusters. While insightful, this approach requires researchers to answer an alternative question on which there is equally little theoretically guidance. We therefore develop our method under the "model-based" framework, in which the researcher has in mind some data-generating process that entails dependent clusters.

The remainder of this chapter is organized as follows. Section 2.2 describes our proposed test. Section 2.3 presents Monte Carlo simulations. Section 2.4 demonstrates an application to Gneezy et al. (2019). Section 2.5 concludes. Proofs are collected in Appendix B.1.

2.2. The Proposed Test

2.2.1. Model and Assumptions

In the following, we assume that the researcher has conducted inference on $\beta \in \mathbf{R}$, and seeks a robustness check for the level of clustering used for said inference. As will become clear in Section 2.2.3, using a scalar β yields computational advantages, though the test can be feasibly computed for moderate dimensions of β. For this reason and for ease of exposition we limit our discussion to the scalar case.

Consider the linear regression:

$$Y_i = X_i\beta + W_i'\gamma + U_i, \quad E[X_iU_i] = 0,\ E[W_iU_i] = 0\ , \tag{2.1}$$

where $\beta \in \mathbf{R}$ is the parameter of interest and $\gamma \in \mathbf{R}^d$ is a nuisance parameter. Suppose there are r clusters, indexed by $k \in \mathcal{K}$. Within each cluster k, there are q_k sub-clusters, indexed by $j \in \mathcal{J}_k$. Within each sub-cluster j, there are n_j individuals indexed by $i \in \mathcal{I}_j$. Let $\mathcal{J} = \bigcup_{k\in\mathcal{K}} \mathcal{J}_k$ and $\mathcal{I} = \bigcup_{j\in\mathcal{J}} \mathcal{I}_j$. Further, let $n = \sum_{j\in\mathcal{J}} n_j$ and $q = \sum_{k\in\mathcal{K}} q_k = |\mathcal{J}|$. We also write $i \in \mathcal{I}_k$ when $i \in \mathcal{I}_j$ and $j \in \mathcal{J}_k$. In the following, we suppress dependence on j and k whenever this does not cause confusion.

Assumption 2.1. Suppose that for every cluster j, there exists a vector Π_j, with a consistent estimator $\hat{\Pi}_j$, such that for all $i \in \mathcal{I}_j$:

$$X_i = W_i'\Pi_j + \varepsilon_i, \quad E[W_i\varepsilon_i] = 0\ . \tag{2.2}$$

Suppose that within a sub-cluster, W has full rank. Then $\hat{\Pi}_j$ can be chosen as the sub-cluster level OLS estimator of X on W. Otherwise, we can just drop variables until we obtain a linearly independent subset $\tilde{W}$. The entries of $\hat{\Pi}_j$ corresponding to the dropped variables can then be set to 0 while the remaining entries are chosen to be the corresponding coefficients from the sub-cluster level regression of X on $\tilde{W}$. Alternatively, if the researcher is willing to assume that Π_j is identical across clusters, $\hat{\Pi}_j$ can also be obtained from the full sample regression of X on W. Now define

$$Z_i = (X_i - W_i\Pi_j)U_i \quad \text{and} \quad \hat{Z}_i = (X_i - W_i\hat{\Pi}_j)\hat{U}_i\ , \tag{2.3}$$

where $\hat{U}_i$ is the full-sample OLS residual using equation (2.1). Suppose we know that clusters are independent, so that $E[Z_i Z_{i'}] = 0$ when $i \in \mathcal{I}_k, i' \in \mathcal{I}_{k'}$ and $k \neq k'$. That is, observations in sub-clusters from different clusters are uncorrelated. Under this assumption, we test the null hypothesis that sub-clusters within the same cluster are uncorrelated:

$$H_0 : E[Z_i Z_{i'}] = 0 \text{ for all } i \in \mathcal{I}_j, i' \in \mathcal{I}_{j'}, j \neq j' \tag{2.4}$$

against the alternative hypothesis that there exists sub-clusters within at least one cluster that exhibit correlation:

$$H_A : E[Z_i Z_{i'}] \neq 0 \text{ for some } i \in \mathcal{I}_j, i' \in \mathcal{I}_{j'} \text{ such that } j, j' \in \mathcal{J}_k, j \neq j' \ .$$

Note that changing the choice of X_i and W_i corresponds to testing different null hypotheses and could lead to differing outcomes. If a researcher wants to test the level of clustering used for inference on β, X_i should be projected onto W_i. Similarly, if inference was conducted on γ, then W_i should take the place of X_i in equation (2.3).

Remark 2.1. A researcher interested in inference on β only has to test the residualized hypothesis of equation (2.4). This is because to the first order, the asymptotic distributions of

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \quad \text{and} \quad \sqrt{n_j}\left(\hat{\beta}_j - \beta\right)$$

depend only on

$$\frac{1}{\sqrt{n}} \sum_{i \in \mathcal{I}} Z_i \quad \text{and} \quad \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} Z_i \ .$$

respectively. If the Z_i's exhibit no correlation across clusters, then conducting inference using the sub-clusters is appropriate. We flesh out this argument in Appendix B.3.

Remark 2.2. Note that tests for different coefficients require different adjustments for clustering. This is unsurprising since our null hypothesis concerns dependence between Z_i's. For intuition, consider a setting with two covariates, X_1 and X_2, which are mean 0 random variables that are independent of each other and independent of U. Consider the null hypothesis when X_1 is our covariate of interest. Now,

$$Z_{1,i} = (X_{1,i} - X_{2,i}\Pi_{1,j})\, U_i = X_{1,i}U_i \text{ since } X_1 \perp\!\!\!\perp X_2 \ ,$$

and similarly for X_2. Hence,

$$E\left[Z_{1,i}Z_{1,i'}\right] = E[X_{1,i}X_{1,i'}]E[U_iU_{i'}] \quad , \quad E\left[Z_{2,i}Z_{2,i'}\right] = E[X_{2,i}X_{2,i'}]E[U_iU_{i'}] \ .$$

If $X_{1,i}$ is independent across sub-clusters, then $E\left[Z_{1,i}Z_{1,i'}\right] = 0$. This is true even if $X_{2,i}$ is dependent within clusters, so that $E\left[Z_{2,i}Z_{2,i'}\right] \neq 0$. A similar phenomenon arises in methods employing degrees of freedom correction for inference with a small number of clusters. Here, each slope parameter in a regression may require a test with different degrees of freedom (see Bell and McCaffrey (2002), Imbens and Kolesár (2016) and Carter et al. (2017)).

Remark 2.3. As with Ibragimov and Müller (2016) and MacKinnon et al. (2020), we require the researcher to specify independent clusters which nest the potentially correlated sub-clusters. For a concrete example, these methods cannot test the the null of commuting zone level clustering against the alternative of state level clustering. However, they

can be used to test the null of county level clustering against the alternative of state level clustering. Researchers interested in testing the null of commuting zone level clustering might instead consider the alternative of clustering at the level of labor market areas.

We further assume the following:

Assumption 2.2. Suppose q and r are fixed, but $n_j \to \infty$ for all $j \in \mathcal{J}$. Let Z_i be defined as in equation (2.3). Suppose there exists $\Omega \in \mathbf{R}^{q \times q}$ such that the q-vector $S_n \xrightarrow{d} S$, where

$$S_n := \begin{pmatrix} \frac{1}{\sqrt{n_1}} \sum_{i \in \mathcal{I}_1} Z_i \\ \vdots \\ \frac{1}{\sqrt{n_q}} \sum_{i \in \mathcal{I}_q} Z_i \end{pmatrix} \quad \text{and} \quad S := N(\mathbf{0}, \Omega). \tag{2.5}$$

Further, let $\hat{\beta}$ and $\hat{\gamma}$ be the (joint) respective OLS estimators of β and γ, as defined in equation (2.1) and $\hat{\Pi}_j$ be the estimator of Π_j as defined in equation (2.2). Suppose:

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\gamma} \xrightarrow{p} \gamma, \quad \sqrt{n_j}\left(\hat{\Pi}_j - \Pi_j\right) = O_p(1) \quad \text{for all } j \in \mathcal{J}\ .$$

In other words, we assume that the errors are weakly correlated within each sub-cluster j. Imposing weak dependence within a (sub-)cluster is not an uncommon assumption (see for instance the discussion in Canay et al. (2017a) and Bester et al. (2011)). We note that under H_0, Ω is a diagonal matrix. On the other hand, under the alternative, it has a block diagonal structure due to correlation between sub-clusters.

Remark 2.4. Our assumption that $S_n \to S$ does not implicitly assume that sub-clusters have similar sizes. Intuitively, this is because our randomization test assigns

"equal weight" to each sub-cluster: each sub-cluster is normalized by its own n_j, and the sign of each $S_{n,j}$ contribute equally to the sign mismatch within its parent cluster. As such, heterogeneous sub-cluster sizes pose no issue for our test. Nonetheless, the quality of the asymptotic approximation is determined by $\min_{j\in[q]} n_j$, so the smallest cluster has to be large. We expand on this point in Appendix B.4 and explain how the restricted heterogeneity assumptions that are required for inference with clustered data are not needed in our case.

Remark 2.5. Without further assumptions on Z_i, our test requires large sub-clusters. This rules out testing the null of no clustering where there is only one observation in each sub-cluster. However, the test is valid for the null of no clustering if we are willing to assume that each Z_i is symmetrically distributed around 0. This might obtain, for example, if U is symmetric around 0 conditional on X and W. Such assumptions can be found in the econometrics literature. For example, Davidson and Flachaire (2008) use it to justify a wild-bootstrapped based F-test for the linear regression model. Nonetheless, we consider this assumption to be highly restrictive and hence justify our test via large sub-cluster asymptotics.

2.2.2. Test Statistic and Critical Value

In this subsection, we define the test statistic and explain the need to search over the worst case critical value. Before doing so, we first consider the infeasible test in which the true parameters – β, γ and Π as defined in equations (2.1) and (2.2) – are observed. Readers who are only interested in the details of implementation can skip to the end of Section 2.2.3.

2.2.2.1. Infeasible Test. Suppose we know β, γ and Π. Given Y_i and X_i, we can back out U_i and construct the vector S_n^*, whose j^{th} entry is

$$S_{n,j}^* = \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} Z_i = \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} (X_i - W_i' \Pi_j) U_i \ . \tag{2.6}$$

Given S_n^*, we can then define the infeasible test statistic:

$$T(S_n^*) = \frac{1}{r} \sum_{k \in \mathcal{K}} \left| \sum_{j \in \mathcal{J}_k} \left(\mathbf{1}(S_{n,j}^* \geq 0) - \mathbf{1}(S_{n,j}^* < 0) \right) \right| \ . \tag{2.7}$$

The inner sum is the net number of positive $S_{n,j}^*$ within each cluster k. Intuitively, if the observations across sub-clusters are independent, the net number of positive $S_{n,j}^*$ should be close to 0. Conversely, if they are positively correlated, this number will be large in absolute value, since many sub-clusters will have $S_{n,j}^*$ of the same sign. On the other hand, if they are negatively correlated, this number will be more concentrated around 0 than in the independent case. As will become clear below, our test interprets large absolute values of $T(S_n^*)$ as violation of the null. For this reason, it will not have power against negative correlation.

Remark 2.6. There are two advantages to having a test statistic that depends only on the sign of the $S_{n,j}^*$'s. Firstly, large and small realizations of $S_{n,j}^*$ contribute the same amount to $T(S_n^*)$. As such, the performance of our test is not affected even if sub-clusters have wildly differing variances, a source of heterogeneity that may be important in applied work. We demonstrate this robustness property via simulations in Section 2.3.2. Secondly, the feasible version of this test requires searching over the worst case values of the test

statistic. As will become clear in Section 2.2.3, this search is simplified by our choice of test statistic.

Next, denote by $\mathbf{G}$ the set of sign changes. $\mathbf{G}$ can be identified with the set $\{-1, 1\}^q$. That is, we can construct $\mathbf{G}$ by enumerating all vectors of length q with either 1 or -1 in each component. Then for each $g \in \mathbf{G}$,

$$gS_n^* = \begin{pmatrix} g_1 \cdot S_{n,1}^* \\ \vdots \\ g_q \cdot S_{n,q}^* \end{pmatrix} .$$

Now let $p^*(S_n^*)$ be the proportion of $T\left(gS_n^*\right)$ that are no smaller than $T\left(S_n^*\right)$:

$$p(S_n^*) = \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} \mathbf{1}\left\{T\left(gS_n^*\right) \geq T\left(S_n^*\right)\right\} . \tag{2.8}$$

The test rejects the null hypothesis when $p(S_n^*)$ is small – that is, when $T\left(S_n^*\right)$ is extreme relative to $T\left(gS_n^*\right)$:

$$\phi_n^* = \begin{cases} 1 & \text{if } p(S_n^*) \leq \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{2.9}$$

The intuition for the randomization test is as follows. Since $S_{n,j}^*$ involves only units within the same sub-cluster, under the null hypothesis, S_n^* converges to a mean-zero normal distribution with independent components. Independence, together with symmetry of normal random variables about their means, implies that for any $g \in \mathbf{G}$, gS_n^*, has the same distribution as S_n^*. Hence, the randomization distribution $\{T(gS_n^*)\}_{g \in \mathbf{G}}$ is in fact the

distribution of $T(S_n^*)$ conditional on the values of $|S_n^*|$, where $|\cdot|$ is applied component-wise. Rejecting the null hypothesis when we observe values of $T(S_n^*)$ that are extreme relative to $\{T(gS_n^*)\}_{g\in\mathbf{G}}$ therefore leads to a test with the correct size.

Note that the randomization test defined above is non-randomized. Randomization tests can also employ a randomized rejection rule for the situation when

$$\frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}}\mathbf{1}\left\{T\left(gS_n^*\right)>T\left(S_n^*\right)\right\}<\alpha \quad \text{but} \quad \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}}\mathbf{1}\left\{T\left(gS_n^*\right)\geq T\left(gS_n^*\right)\right\}>\alpha\ .$$

Using a randomized rejection rule, the randomization test will have size equal to α exactly if the necessary symmetry properties hold in finite sample. The test defined in equation (2.9) is conservative since it never rejects when the above situation occurs. However, we present the deterministic version since the test that we propose is based on it.

2.2.2.2. Naïve Test. Tests based on S_n^* are infeasible since β, γ and the Π_j's are unknown. Suppose we simply replaced Z_i with $\hat{Z}_i$ and performed the randomization test with the estimated scores. It turns out that this procedure is incorrect. To see this, let $\tilde{S}_n$ be S_n^* but with $\hat{Z}_i$ replacing Z_i. Then we can write each component of $\tilde{S}_n$ as:

$$\tilde{S}_{n,j}=\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\hat{Z}_i=\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i-W_i'\hat{\Pi}_j\right)\hat{U}_i \tag{2.10}$$

$$\approx\underbrace{\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i-W_i'\Pi_j\right)U_i}_{S_{n,j}^*}-\underbrace{\left(\hat{\beta}-\beta\right)\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i-W_i'\hat{\Pi}_j\right)^2}_{=:\,A_j}\ . \tag{2.11}$$

In the above equation, $S_{n,j}^*$ is the part that is informative about cluster structure. However, each component now has an additional nuisance term A_j that does not go away under asymptotics that take the number of sub-clusters to be fixed. Because $\hat{\beta}-\beta$ is

common across the A_j's, it induces correlation across $\tilde{S}_{n,j}$ even when the $S^*_{n,j}$'s are independent, leading potentially to over-rejection. Addressing this complication which does not arise in frameworks taking $q \to \infty$ results in the conservativeness of our test.

2.2.2.3. Feasible Test. If we knew $\hat{\beta} - \beta$, we could back out $S^*_{n,j}$ for the randomization test using equation (2.11). Since that is not possible, we propose to search over values of $\hat{\beta} - \beta$ to ensure that the test controls size when the unobserved term takes on extreme values.

For a given $\lambda \in \mathbf{R}$, let $\hat{S}_n(\lambda)$ be $q \times 1$ vector whose j^{th} entry is the following term:

$$\hat{S}_{n,j}(\lambda) := \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} \left(X_i - W_i' \hat{\Pi}_j \right) \hat{U}_i + \lambda \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} \left(X_i - W_i' \hat{\Pi}_j \right)^2 .$$

Note that if we set $\lambda^* = \hat{\beta} - \beta$, then $\hat{S}_{n,j}(\lambda^*) = S^*_{n,j} + o_p(1)$. Define:

$$T(\hat{S}_n(\lambda)) = \frac{1}{r} \sum_{k \in \mathcal{K}} \left| \sum_{j \in \mathcal{J}_k} \left(\mathbf{1}(\hat{S}_{n,j}(\lambda) \geq 0) - \mathbf{1}(\hat{S}_{n,j}(\lambda) < 0) \right) \right| .$$

For a given λ, this is just the test statistic in equation (2.7) but with $\hat{S}_{n,j}(\lambda)$ taking the place of $S^*_{n,j}$. As before, we denote by $\mathbf{G}$ the set of sign changes and write:

$$g\hat{S}_n(\lambda) = \begin{pmatrix} g_1 \cdot \hat{S}_{n,1}(\lambda) \\ \vdots \\ g_q \cdot \hat{S}_{n,q}(\lambda) \end{pmatrix} .$$

Now let $p(\hat{S}_n(\lambda))$ be the proportion of $T\left(g\hat{S}_n(\lambda)\right)$ that takes on extreme values relative to $T\left(\hat{S}_n(\lambda)\right)$:

$$p(\hat{S}_n(\lambda)) = \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} \mathbf{1}\left\{T\left(g\hat{S}_n(\lambda)\right) > T\left(\hat{S}_n(\lambda)\right)\right\} \ . \tag{2.12}$$

We can then define the randomization test as:

$$\phi_n = \begin{cases} 1 & \text{if } \sup_{\lambda\in\mathbf{R}} p(\hat{S}_n(\lambda)) \leq \alpha \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

We can then prove the following result:

Theorem 2.1. Under assumptions 2.1 and 2.2, $\limsup_{n\to\infty} \mathbb{E}[\phi_n] \leq \alpha$.

The test is a two-stage process. In the first stage, it searches for the value of λ that leads to the largest p-value. In the second stage, the test rejects if this worst-case p-value is still smaller than the desired level of significance α. Since the worst-case p-value bounds the true p-value from above, the rejection rule based on the worst-case p-value must be conservative.

As the Monte Carlo simulations in Section 2.3 shows, the test has size that could be much smaller than α under the null hypothesis. However, the same simulations also show that the test has reasonable power under the alternative hypothesis, particularly in settings where clusters are heterogeneous in their variances. The potential usefulness of our test is further seen in the empirical application (Section 2.4), where it detects dependence in the clusters chosen by Gneezy et al. (2019).

Remark 2.7. The worst-case test has no power if $r = 1$ since $\lambda = \text{median}(\{\tilde{S}_{n,j}\})$ will set exactly half the signs of $\hat{S}_{n,j}(\lambda)$ to be positive and half to be negative, so that the signs are completely balanced. However, this is no longer true with $r > 1$ since only a single value can be chosen to balance signs across multiple clusters. The implementation procedure provides further intuition for power in this test. See the next subsection.

Remark 2.8. As with standard randomization tests, $|\mathbf{G}|$ may sometimes be too large so that computation of $p(\hat{S}_n(\lambda))$ becomes onerous. In these instances, it is possible to replace $p(\hat{S}_n(\lambda))$ with a stochastic approximation. Formally, let $\hat{\mathbf{G}} = \left\{g^1, ..., g^B\right\}$, where g^1 is the identity transformation and $g^2, ..., g^B$ are i.i.d. Uniform($\mathbf{G}$). Using $\hat{\mathbf{G}}$ instead of $\mathbf{G}$ in equation (2.12) does not affect validity of theorem 2.1. For implementation, we follow Canay et al. (2017a) in evaluating the $p(\hat{S}_n(\lambda))$ completely when $q \leq 10$ and approximating it with $B = 1000$ when $q > 10$.

Remark 2.9. We advocate the use of our test as a robustness check, after a researcher has chosen a level of clustering for inference, in the same spirit that manipulation tests are routinely used in studies with regression discontinuity designs, or in tests for pre-trends in studies involving difference-in-differences. In particular, the original inference results should be presented with results of the current test, regardless of the outcome. Conceptually, this is different from using the test as a pre-test to select the level of clustering prior to inference. The distinction is important as pre-testing is known to induce uniformity issues, where inference in the second stage (on β) suffers from distortion due to mistakes in the pre-test (that happen with positive probability). These same concerns are articulated by Ibragimov and Müller (2016), who argue that their test "merely provides

empirical evidence on the plausibility of one particular clustering assumption". We take exactly the same view of our test.

2.2.3. Implementation

In this subsection we describe an efficient way of searching for $\lambda \in \mathbf{R}$. This search is simplified by the fact that $p(\hat{S}_n(\lambda))$ depends only on the sign of $\hat{S}_{n,j}(\lambda)$'s. As such, to find $\sup_{\lambda \in \mathbf{R}}$, we only need to search over sign combinations of $\hat{S}_{n,j}$. When β is scalar, the search can be completed in $O(q)$ time. This is reasonable since the test is designed for use when q is small.

Suppose for now that $\sum_{i \in \mathcal{I}_j} \left(X_i - W_i' \hat{\Pi}_j \right)^2 > 0$ for all $j \in \mathcal{J}$. Define:

$$R_j = \frac{\sum_{i \in \mathcal{I}_j} \left(X_i - W_i' \hat{\Pi}_j \right) \hat{U}_i}{\sum_{i \in \mathcal{I}_j} \left(X_i - W_i' \hat{\Pi}_j \right)^2} \ .$$

Then, $\hat{S}_{n,j}(\lambda) \geq 0 \Leftrightarrow R_j + \lambda \geq 0$. Sort the values of R_j's so that $R^{(1)} \geq R^{(2)} \geq ... \geq R^{(q)}$. We must have that $R^{(j)} + \lambda \geq 0 \Rightarrow R^{(j')} + \lambda \geq 0$ for all $j' \leq j$. Let $\hat{S}_{n,(1)}(\lambda), ..., \hat{S}_{n,(q)}(\lambda)$ denote the values of $\hat{S}_{n,j}(\lambda)$ corresponding to $R^{(1)}, ..., R^{(q)}$. Therefore, we only need to consider sequences of the form

$$\hat{S}_{n,(1)} > 0, ..., \hat{S}_{n,(j)} > 0, \quad \hat{S}_{n,(j+1)} < 0, ..., \hat{S}_{n,(q)} < 0 \ ,$$

for some cut-off j. Since the p-value, as defined in equation (2.12), depends only on the sign of $\hat{S}_n$, we can compute it using $\check{S}_n$ in the place of $\hat{S}_{n,j}(\lambda)$:

$$\check{S}_{n,(1)} = ... = \check{S}_{n,(j)} = 1, \quad \check{S}_{n,(j+1)} = ... = \check{S}_{n,(q)} = -1 \ .$$

Here, we see that even when we are searching over the worst case λ, we are only allowed to choose the cut-off point at which the signs change. We can therefore complete the search with no more than q randomization tests. Assuming that the time it takes for each test is $O(1)$, the procedure takes $O(q)$ time. The restriction that $\check{S}_{n,(j)} \geq \check{S}_{n,(j')}$ for all $j \leq j'$ also gives the test power. If all combinations of signs for the $S_{n,j}$'s were allowed, the test will always return a p-value of 1 and will have no power.

Finally, suppose there are sub-clusters such that $\sum_{i \in \mathcal{I}_j} \left(X_i - W_i' \hat{\Pi}_j \right)^2 = 0$. We can repeat the above procedure excluding these sub-clusters. In the final step, we set $\check{S}_{n,j}$ corresponding to these clusters to 0. Hence,

Remark 2.10. We can further reduce computation time by the following. Let

$$R_k^+ = \min_{j' \in \mathcal{J}_k} \{ R_{j'} \text{ greater than or equal to } > 0.5 \text{ of } \{ R_j, j \in \mathcal{J}_k \} \}$$

be the "upward-conservative" median. Also define the "downward-conservative" median:

$$R_k^- = \max_{j' \in \mathcal{J}_k} \{ R_{j'} \text{ less than or equal to } > 0.5 \text{ of } \{ R_j, j \in \mathcal{J}_k \} \} \ .$$

Now let $R^+ = \max_{k \in \mathcal{K}} R_k^+$ and $R^- = \min_{k \in \mathcal{K}} R_k^-$. We only need to consider cutoffs below R^+. Setting the sign cutoff at the argmax of R^+ results in situation in which all clusters have at least half of their entries being -1. If we now set the extreme $S_{n,j}$'s to -1, this will increase the net number of -1's in *all* clusters. Since our test is based on sign imbalance within clusters, such sequences will lead to a strictly larger test statistic and smaller p-values than if they were set to 1. For the same reason, we only need to consider cutoffs above R^-.

We summarise the implementation procedure in Algorithm 1.

Algorithm 1: Worst-Case Randomization Test

Perform full sample OLS to obtain residuals $\hat{U}_i$. Compute $\hat{\Pi}_j$ for each $j \in \mathcal{J}$.
for $j \in [q]$ **do**
  if $\sum_{i\in\mathcal{I}_j}\left(X_i - W_i'\hat{\Pi}_j\right)^2 > 0$ **then** compute: $R_j = \frac{\sum_{i\in\mathcal{I}_j}\left(X_i - W_i'\hat{\Pi}_j\right)\hat{U}_i}{\sum_{i\in\mathcal{I}_j}\left(X_i - W_i'\hat{\Pi}_j\right)^2}$
  else set $R_j = 0$.
Sort the values of R_j's such that $R^{(1)} \geq R^{(2)} \geq ... \geq R^{(q)}$.
for $j \in [q]$, $R_{(j)} \neq 0$, $R^- \leq R_{(j)} \leq R^+$ **do**
  Set $\check{S}_{n,(1)} = ... = \check{S}_{n,(j)} = 1, \quad \check{S}_{n,(j+1)} = ... = \check{S}_{n,(q)} = -1$.
  if $R_{(j)} = 0$ **then** replace $\check{S}_{n,(j)}$ with 0.
  Compute $p(\check{S}_n)$. This is as defined in equation (2.12), except with $\check{S}_n$ in place of $\hat{S}_{n,j}(\lambda)$. Save this value as $\hat{p}_j$.
if $\max_{j\in[q],R_j\neq 0}\hat{p}_j \leq \alpha$ **then** return 1. Reject the null hypothesis.
else return 0. Do not reject the null hypothesis.

2.2.4. Comparison with Existing Tests

To our knowledge, two other tests have been proposed for the level of clustering. They take either the number of sub-clusters in each cluster to infinity or the number of clusters to infinity. We assume both to be fixed. For ease of exposition, we restrict our discussion of these tests to the univariate case.

Ibragimov and Müller (2016) (IM hereafter) adopts an asymptotic framework that takes $q_k \to \infty$ for all $k \in \mathcal{K}$. Consider estimating a regression coefficient cluster-by-cluster. Let $\hat{\beta}_k$ denote coefficients estimated using only cluster k. The IM test is based on the asymptotic distribution of an estimator for the variance of $\frac{1}{r}\sum_{k=1}^{r}\hat{\beta}_k$. Let this variance be denoted by V and let $\hat{\Omega}_k^{\text{CCE}}$ be the cluster-robust variance estimator for $\hat{\beta}_k$, where the

clustering is done at the sub-cluster level using $j \in \mathcal{J}_k$. Under the null hypothesis, $\hat{\Omega}_k^{\text{CCE}}$ consistently estimates the variance of each $\hat{\beta}_k$.

Under either the null or the alternative, but maintaining the assumption that coarse clusters are independent, consider estimating V by:

$$\hat{V} = \frac{1}{r-1}\sum_{k=1}^{r}(\hat{\beta}_k - \bar{\beta})^2 \ , \qquad \bar{\beta} = \frac{1}{r}\sum_{r=1}^{k}\hat{\beta}_r \ .$$

IM show that under the null, $\hat{V} \xrightarrow{d} V^W$, where $V^W = \frac{1}{r-1}\sum_{k=1}^{r}(W_k - \bar{W})^2$ and

$$W \sim N(0, \text{diag}(\Omega_1^{\text{CCE}}, \Omega_2^{\text{CCE}}, ..., \Omega_r^{\text{CCE}})) \ .$$

The IM test constructs a reference distribution $\hat{V}^W$ by drawing W from

$$N(0, \text{diag}(\hat{\Omega}_1^{\text{CCE}}, \hat{\Omega}_2^{\text{CCE}}, ..., \hat{\Omega}_r^{\text{CCE}}))$$

and seeing if $\hat{V}$ is larger than the $(1-\alpha)^{\text{th}}$ quantile of $\hat{V}^W$.

There are two limitations to the IM test that our test does not share. Firstly, they require the regression to be estimated cluster-by-cluster. This would be infeasible in, for example, differences-in-differences set ups where treatment varies at the cluster level. Secondly, since their asymptotics take $q_k \to \infty$, we expect the test to have poor properties when q_k is small. Instead, our test is expected to have good properties even when q_k is small as long as n_j is large. These benefits come at a cost. We expect our test to perform worse if observations within sub-clusters are highly correlated, whereas the IM test allows unrestricted covariance within sub-clusters. Our test is also conservative under the null hypothesis. We note also that neither test has power against negative correlations. This

is because both tests use test statistics that take on large value relative to their reference distributions only when there is positive correlation.

MacKinnon et al. (2020) (MNW hereafter) considers an asymptotic framework that takes $r \to \infty$. In the same spirit as IM, the MNW test is a Hausman-type test based on the variance of regression coefficients. Consider the full sample regression coefficient $\hat{\beta}$. Under the null hypothesis, the (full-sample) cluster-robust covariance estimator at the sub-cluster level, denoted, $\hat{\Omega}_J^{\text{CCE}}$, is consistent for the asymptotic variance-covariance matrix.

Under either the null or the alternative, but maintaining the assumption that coarse clusters are independent, the (full-sample) cluster-robust covariance estimator at the cluster level, denoted, $\hat{\Omega}_K^{\text{CCE}}$, is consistent for the asymptotic variance-covariance matrix. Under the null hypothesis, the authors show that their test statistic converges to a standard normal distribution: $\frac{\hat{\Omega}_K^{\text{CCE}} - \hat{\Omega}_J^{\text{CCE}}}{\hat{V}^{MNW}} \xrightarrow{d} N(0,1)$ for an appropriately defined $\hat{V}^{MNW}$.

It is well known that the cluster-robust covariance estimator can be severely biased when r is small. In order to deal with such situations, the authors propose to conduct the test using the wild (sub-)cluster bootstrap. This procedure imposes the cluster structure specified in the null hypothesis when generating bootstrap samples. Under the null of no clustering, it reduces to the simple wild bootstrap. They prove the consistency of this approach in their large-r framework, showing power even against alternatives with negative correlations.

Compared to the MNW test, our test is theoretically justified when both r and q are small, provided that n_j's are large. Our test could therefore be preferable in such applications since it is presently not known if the MNW test remains valid once we take

r and q to be fixed. However, as with the IM test, the MNW test allows unrestricted covariance within sub-clusters, whereas our test is expected to have poor performance if observations within sub-clusters are highly correlated. Our test is also conservative relative to the MNW test. On the other hand, simulation evidence suggests that it has comparable performance with the MNW test when clusters have differing variances (see Section 2.3).

2.3. Monte Carlo Simulations

In this section, we examine the finite sample performance of our worst-case randomization test (WCR) together with the IM and bootstrap version of the MNW tests via Monte Carlo simulations. We also study the the naïve randomization test (NR) as described in Section 2.2.2.2. Section 2.3.1 considers the effects of varying r, q_k and n_j. Section 2.3.2 investigates the effects of cluster heterogeneity. Our data generating processes are as follows:

Model 1: Model 1 is defined by the following:

$$Y_{t,j,k} = X'_{t,j,k}\beta + \sigma_{j,k}\left(\rho V_{t,k} + \frac{1}{\sqrt{1-\phi^2}}U_{t,j,k}\right) ,$$

$$V_{t,k} \overset{\text{iid}}{\sim} N(0,1) , \quad U_{t,j,k} = \phi U_{t-1,j,k} + \varepsilon_{t,j,k}, \quad \varepsilon_{t,j,k} \overset{\text{iid}}{\sim} N(0,1) ,$$

In particular, we set $X_{t,j,k} = \beta = 1$ and $\phi = 0.25$. Errors are correlated within a sub-cluster, according to an $AR(1)$ process, with autocorrelation coefficient ϕ. ρ captures the importance of cluster level shock. Since $\frac{1}{\sqrt{1-\phi^2}}U_{t,j,k}$ has unit variance, ρ is exactly the relative variance of cluster- to sub-cluster-level shocks. $\sigma_{j,k}$ controls the variance of the unobserved term in each cluster k. Here in Section 2.3.1, we set $\sigma_{j,k} = 1$ for all

$j \in \mathcal{J}, k \in \mathcal{K}$. In Section 2.3.2, we explore the consequences of cluster heterogeneity by varying $\sigma_{j,k}$.

Model 2: This is the model used in the simulations of MacKinnon et al. (2020), with the constant omitted. Let $m_k = \sum_{j \in \mathcal{J}_k} n_j$ be the total number of observations in cluster k. Let U_k be the $m_k \times 1$ vector of $U_{t,j,k}$ for all observations in cluster k. Then

$$U_k = \rho W_\xi \xi_k + \sqrt{1-\rho^2}\,\epsilon_k \quad , \quad \epsilon_k \sim N(0, I_{m_k}) \ ,$$

where ξ_k is a 10×1 vector distributed as:

$$\xi_{k,1} \sim N(0,1) \quad , \quad \xi_{k,l} = \phi \xi_{k,l-1} + e_{k,l} \quad , \quad e_{k,l} \sim N(0, 1-\phi^2) \quad , \quad l \in \{2, ..., 10\}$$

and W_ξ is the $m_k \times 10$ loading matrix with the $(i,j)^{\text{th}}$ entry $\mathbf{1}\{j = \lfloor (i-1)10/m_k \rfloor + 1\}$. Under this model, $\frac{1}{10}$ of the observations in each cluster are correlated because they depend directly on the same $\xi_{k,l}$. In addition, there is correlation between the $\xi_{k,l}$'s since it is generated according to an AR(1) process. Observations are then ordered so that every sub-cluster contains the same number of observations that depend on each $\xi_{k,l}$. Finally, $\beta = (1,1)'$ and the two covariates are independent and generated in the same way as U. This model features more complex correlations between and within the sub-clusters. Clusters are independent and identically distributed. As in Section 5.2 of MacKinnon et al. (2020), we set $\phi = 0.5$. ρ here is directly comparable to w_ξ in their simulations.

For our simulations, we perform the test at the 5% level. 10,000 Monte Carlo simulations were drawn for each combination of the parameters. The non-standard reference distribution in IM is evaluated using 1,000 Monte Carlo draws. Wild bootstrap in MNW is evaluated using 399 draws as in their simulations.

2.3.1. Performance over values of r, q_k and n_j

To understand the size and power of each of our tests in scenarios with few clusters and few sub-clusters, we consider equal-sized clusters and sub-clusters, with $r \in \{4, 8, 12\}$, $q_k \in \{4, 8, 12\}$ and $n_j \in \{25, 50, 100\}$. We consider $\rho \in \{0, 0.5\}$.

Table 2.1 presents results under the null hypothesis ($\rho = 0$). Across the two models, we see that regardless of r, the IM test performs poorly when q_k is small. With $q_k = 4$, type I error is between 15% and 20%. By $q_k = 12$, however, the size is between 6-7%. Comparatively, our test, which is highly conservative, has type I error less than 2% across all values of q_k. The MNW and NR tests perform well across the board. Table 2.2 presents results under the alternative $\rho = 0.5$. Relative to the IM and MNW tests, our test has power that is consistently lower. In particular, our test does poorly when q_k is small. This is the weakness of the worst-case approach.

Figure 2.1 presents power of the tests for $r = 8$, $q_k = 12$, $n_j = 100$ as we vary ρ from 0 to 2 in model 1 and 0 to 1 in model 2. Across the two models, we see that the IM and MNW tests have greater power than our test. However, as ρ increases, our test quickly catches up in power.

2.3.2. Cluster Heterogeneity

As a growing body of papers document, heterogeneity across clusters can pose challenges for cluster-robust inference (see for instance Carter et al. (2017), Djogbenou et al. (2019) and Hu and Spamann (2020)). Three sources of heterogeneity are of particular concern: (i) distribution of errors, (ii) distribution of covariates, and (iii) cluster sizes. In the following, we investigate in turn how each of these issues affect tests for the level of clustering.

Size - Homogeneous Clusters

			Model 1				Model 2			
r	q_k	n_j	NR	WCR	IM	MNW	NR	WCR	IM	MNW
4	4	25	0.019	0.000	0.145	0.053	0.019	0.000	0.161	0.053
		50	0.022	0.000	0.155	0.059	0.021	0.000	0.160	0.055
		100	0.020	0.000	0.154	0.058	0.019	0.000	0.157	0.055
	8	25	0.024	0.000	0.090	0.050	0.025	0.000	0.103	0.055
		50	0.022	0.000	0.095	0.054	0.022	0.000	0.098	0.052
		100	0.022	0.000	0.094	0.050	0.026	0.000	0.098	0.053
	12	25	0.024	0.001	0.075	0.049	0.026	0.001	0.083	0.053
		50	0.022	0.000	0.075	0.050	0.026	0.000	0.086	0.057
		100	0.025	0.001	0.079	0.052	0.026	0.000	0.081	0.054
8	4	25	0.024	0.000	0.171	0.055	0.027	0.000	0.181	0.053
		50	0.024	0.000	0.162	0.051	0.025	0.000	0.170	0.051
		100	0.025	0.000	0.165	0.053	0.026	0.000	0.165	0.052
	8	25	0.025	0.000	0.095	0.051	0.026	0.000	0.101	0.049
		50	0.026	0.001	0.092	0.049	0.025	0.000	0.102	0.051
		100	0.027	0.001	0.097	0.052	0.024	0.001	0.092	0.047
	12	25	0.028	0.001	0.080	0.051	0.032	0.001	0.081	0.050
		50	0.027	0.001	0.075	0.048	0.028	0.001	0.079	0.050
		100	0.026	0.001	0.074	0.050	0.027	0.001	0.078	0.050
12	4	25	0.022	0.000	0.177	0.051	0.019	0.000	0.187	0.048
		50	0.022	0.000	0.181	0.055	0.020	0.000	0.180	0.050
		100	0.022	0.000	0.171	0.050	0.020	0.000	0.177	0.045
	8	25	0.034	0.001	0.105	0.054	0.032	0.002	0.105	0.050
		50	0.035	0.001	0.101	0.049	0.035	0.001	0.106	0.052
		100	0.032	0.001	0.100	0.050	0.030	0.001	0.094	0.044
	12	25	0.032	0.002	0.085	0.049	0.029	0.002	0.087	0.050
		50	0.030	0.002	0.076	0.047	0.030	0.003	0.086	0.054
		100	0.033	0.002	0.084	0.053	0.030	0.002	0.085	0.051

Table 2.1. Monte Carlo rejection rates under the null hypothesis $\rho = 0$ at 5% level of significance. WCR refers to our worst-case randomization test. IM is the test from Ibragimov and Müller (2016). MNW is the bootstrap version of the test in MacKinnon et al. (2020). NR is the naïve randomiztion test. r is the number of clusters, q_k is the number of sub-clusters in each cluster, and n_j is the number of individuals in each sub-cluster.

Power - Homogeneous Clusters

			Model 1				Model 2			
r	q_k	n_j	NR	WCR	IM	MNW	NR	WCR	IM	MNW
4	4	25	0.052	0.000	0.315	0.152	0.068	0.000	0.372	0.204
		50	0.061	0.000	0.313	0.157	0.130	0.000	0.529	0.336
		100	0.053	0.000	0.316	0.156	0.230	0.001	0.687	0.520
	8	25	0.122	0.006	0.383	0.291	0.175	0.010	0.458	0.380
		50	0.118	0.006	0.379	0.282	0.317	0.044	0.648	0.576
		100	0.110	0.005	0.363	0.273	0.508	0.126	0.800	0.754
	12	25	0.188	0.028	0.471	0.404	0.269	0.058	0.548	0.509
		50	0.187	0.030	0.455	0.392	0.462	0.177	0.735	0.700
		100	0.186	0.026	0.467	0.399	0.662	0.357	0.868	0.850
8	4	25	0.081	0.002	0.417	0.194	0.116	0.003	0.525	0.290
		50	0.081	0.001	0.430	0.198	0.236	0.013	0.728	0.530
		100	0.078	0.001	0.417	0.195	0.425	0.059	0.888	0.775
	8	25	0.198	0.032	0.568	0.454	0.288	0.068	0.661	0.597
		50	0.194	0.031	0.567	0.447	0.535	0.225	0.869	0.838
		100	0.189	0.028	0.551	0.442	0.782	0.494	0.965	0.957
	12	25	0.329	0.108	0.691	0.628	0.467	0.202	0.779	0.761
		50	0.321	0.106	0.692	0.627	0.731	0.477	0.934	0.925
		100	0.326	0.108	0.687	0.620	0.912	0.775	0.988	0.985
12	4	25	0.092	0.003	0.521	0.233	0.131	0.006	0.622	0.374
		50	0.092	0.002	0.516	0.235	0.288	0.032	0.839	0.674
		100	0.092	0.003	0.525	0.235	0.535	0.142	0.961	0.895
	8	25	0.279	0.065	0.701	0.565	0.409	0.140	0.786	0.737
		50	0.289	0.066	0.692	0.571	0.706	0.409	0.948	0.932
		100	0.273	0.064	0.691	0.555	0.911	0.743	0.994	0.992
	12	25	0.455	0.201	0.830	0.767	0.616	0.352	0.897	0.887
		50	0.442	0.194	0.825	0.766	0.872	0.702	0.984	0.982
		100	0.434	0.191	0.823	0.762	0.976	0.927	0.999	0.999

Table 2.2. Monte Carlo rejection rates under the alternative hypothesis $\rho = 0.5$ at 5% level of significance. WCR refers to our worst-case randomization test. IM is the test from Ibragimov and Müller (2016). MNW is the bootstrap version of the test in MacKinnon et al. (2020). NR is the naïve randomiztion test. r is the number of clusters, q_k is the number of sub-clusters in each cluster, and n_j is the number of individuals in each sub-cluster.

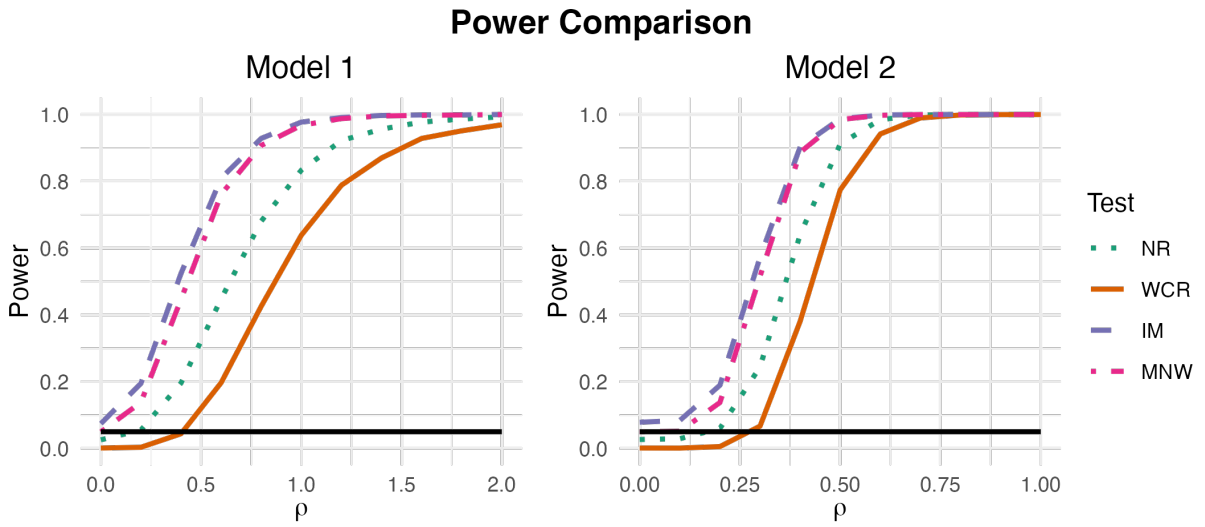

Figure 2.1. Power of various tests for level of clustering when $r = 8$, $q_k = 12$, $n_j = 100$. The black line indicates the nominal size of the tests (5%).

Whereas the previous section suggests that our test has poor performance compared to other tests, a different picture emerges once we consider heterogeneous clusters.

2.3.2.1. Distribution of Errors. We first consider what happens when clusters differ in the distribution of regression errors. Specifically, we are interested in the case in which some clusters have much larger variances in their errors than others. This might happen if some clusters are exposed to more shocks than the others, or because clusters systematically differ in certain covariates and errors are heteroskedastic. We return to model 1 but with $\sigma_{j,1} \in \{5, 10, 15\}$. That is, when all sub-clusters in cluster 1 are much noisier than the rest. Figure 2.2 plots power curves with $r = 8, q_k = 12, n_j = 100$ for $\sigma_{j,1} \in \{5, 10, 15\}$. These curves are directly comparable with Figure 2.1. Starting from the within test comparison, we see that the performance of our test is unaffected by $\sigma_{j,1}$. However, power of IM and MNW quickly degrade as $\sigma_{j,1}$ increases. Turning to the across test comparison, we see that the tests perform similarly when $\sigma_{j,1} = 5$. As $\sigma_{j,1}$ increases

to 10, our test starts to have more power than the IM and MNW tests for $\rho \geq 1$. The across test comparison also shows how the NR test fails to control size. In particular, when $\sigma_{j,1}$, an NR test with nominal size 5% could wrongly reject over 40% of the time.

We see the same patterns when sub-clusters are heterogeneous. Consider again model 1 but with $\sigma_{1,k} \in \{5, 10, 15\}$. That is, when the first sub-cluster in each cluster is much noisier than the rest. Figure 2.3 presents the results. Again, our test is not affected by changing $\sigma_{1,k}$. The power of the IM test falls by a large extent as $\sigma_{1,k}$ increases. The MNW test is also negatively affected by $\sigma_{1,k}$, though less so than the IM test.

2.3.2.2. Distribution of Covariates. We next consider the case when clusters differ in the distribution of covariates. Specifically, we are interested in the effects of having covariates that are perfectly correlated either within sub-clusters or clusters. These situations arise commonly in empirical work, when treatment assignment occurs at the sub-cluster or cluster level.

Our simulation is based on model 1 but with the following modification. When treatment is assigned at the individual level,

$$X_{t,j,k} := \begin{cases} 1 & \text{w.p. } 0.5 \\ 2 & \text{otherwise.} \end{cases}$$

where $X_{t,j,k}$ is independent across observations. When treatment is assigned at the sub-cluster level, $X_{t,j,k} = X_{j,k}$ has the same distribution as above, but is identical within sub-clusters and independent across sub-clusters (and clusters). When treatment is assigned at the cluster level, $X_{t,j,k} = X_k$ is identical within clusters and independent across clusters. The distribution of X was chosen so that β is estimable cluster-by-cluster even

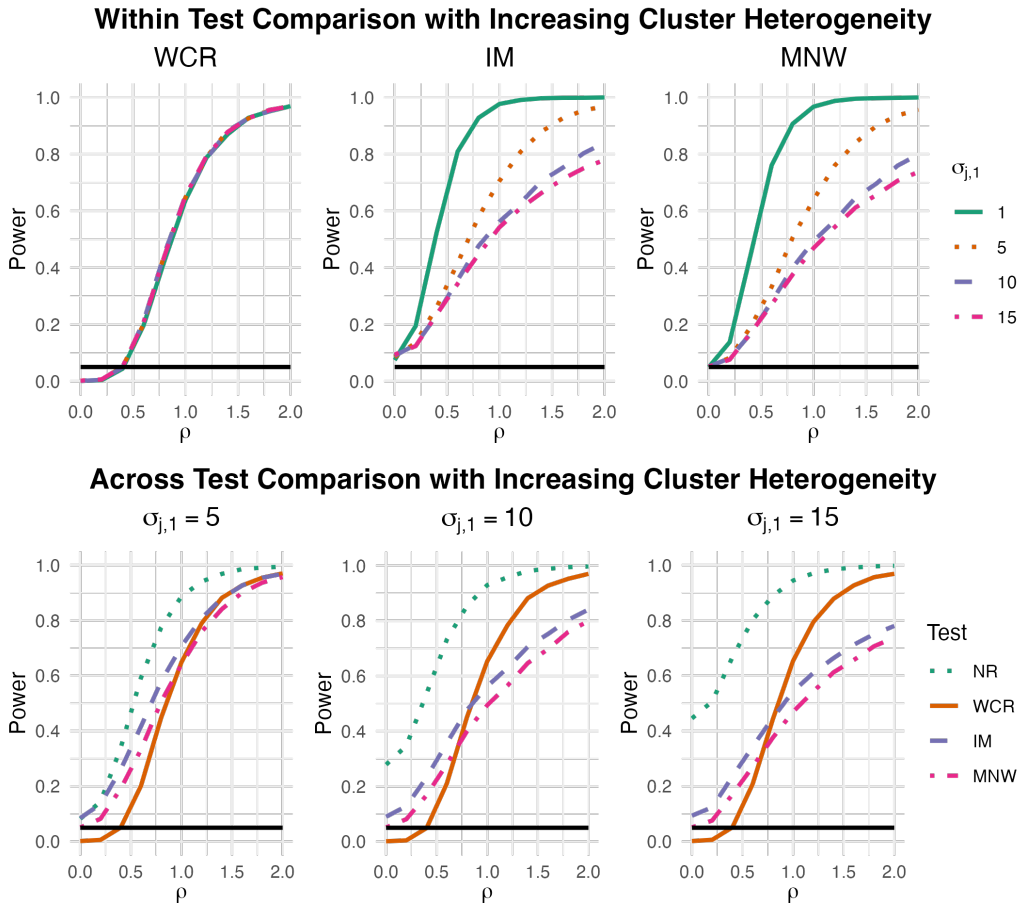

Figure 2.2. Power of various tests for level of clustering in Model 1 as $\sigma_{j,1}$ increases. Here, $r = 8$, $q_k = 12$, $n_j = 100$. The black line indicates the nominal size of the tests (5%).

when treatment is assigned at the cluster level. This condition is unlikely to be satisfied in practice when treatment is assigned at the cluster level. However, we impose it for comparison purposes, since it is necessary for the IM test.

The power curves are shown in Figure 2.4. For WCR, having treatment assignment at the sub-cluster ("Subclust") and cluster level ("Clust") turns out to slightly improve power relative to the case with treatment assignment at the individual level ("Indiv"). The size of the test is not affected. For the IM test, sub-cluster level treatment assignment slightly

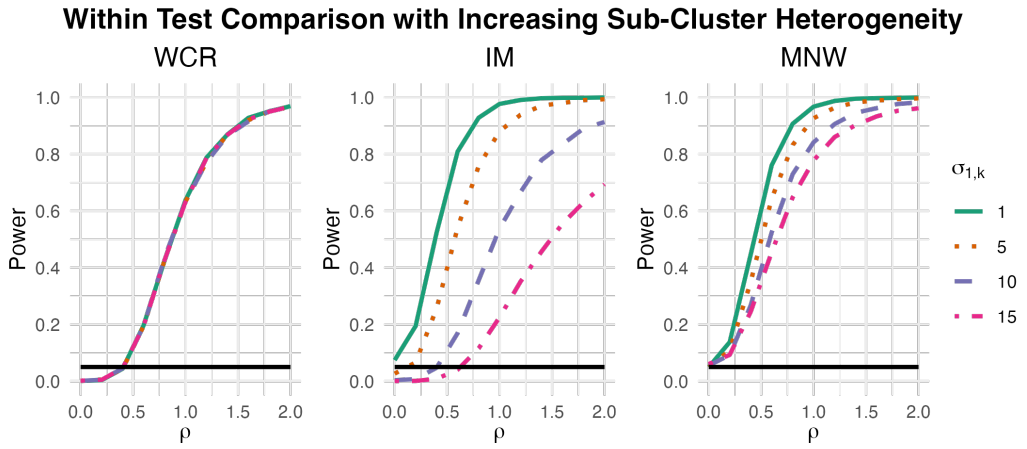

Figure 2.3. Power of various tests for level of clustering in Model 1 as $\sigma_{1,k}$ increases. Here, $r = 8$, $q_k = 12$, $n_j = 100$. The black line indicates the nominal size of the tests (5%).

worsens size control, while cluster level treatment assignment does not appear to any effect. For the MNW test, sub-cluster level treatment assignment does not appear to have any effect, while cluster-level treatment assignment seems to lower power substantially, to the point that it becomes comparable to that of the WCR.

Summing up, these simulations suggest that size control of the IM test is sensitive to heterogeneity in covariate distribution at the sub-cluster level. Meanwhile, power of the MNW test is sensitive to heterogeneity at the cluster level. On the other hand, performance of the WCR test is relatively stable across these models.

2.3.2.3. Cluster Sizes. Finally, we consider the effects of imbalanced cluster sizes. Specifically, we are interested in situations where each cluster has a sub-cluster that is much larger than the others. Our baseline model for this section follows that of Section 2.3.2.2. Inspired by Hu and Spamann (2020)'s study of clustering in the state corporate

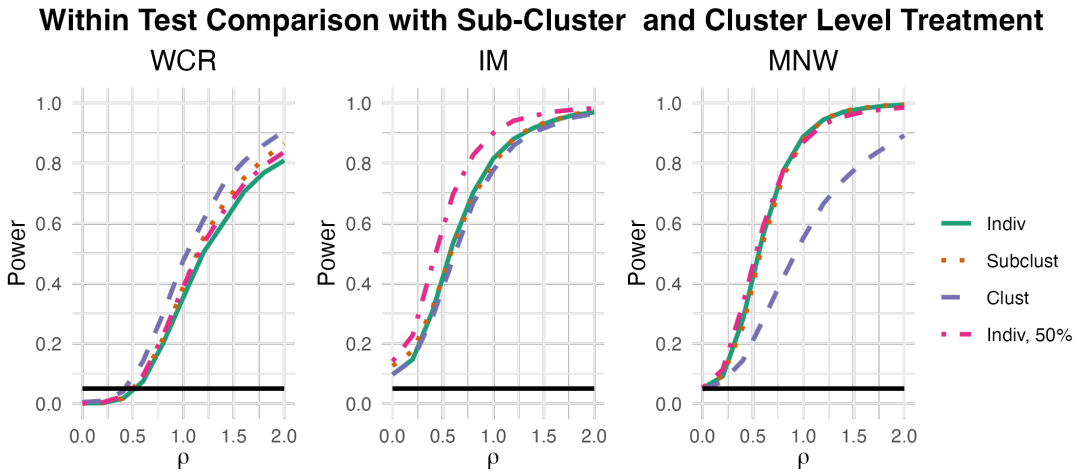

Figure 2.4. Power of various tests for level of clustering in Model 1 with different levels of treatment assignment. "Indiv", "Subclust" and "Clust" refers to treatment assignment at the individual, sub-cluster and cluster level respectively. Here, $r = 8$, $q_k = 12$, $n_j = 100$. "Indiv, 50%" refers to treatment assignment at the individual level, but each cluster also contains a large sub-cluster containing 50% of the observations in the cluster (1100), with the remaining sub-clusters having $n_j = 100$. The black line indicates the nominal size of the tests (5%).

law literature, we set each cluster to contain a sub-cluster comprising 50% of all observations in that cluster. The large sub-cluster therefore contains 1100 observations, while the remaining 11 sub-clusters continue to have size 100. The case with individual treatment is presented in Figure 2.4. Comparing the case with equally sized sub-clusters ("Indiv") and with imbalanced sub-clusters ("Indiv, 50%"), we see that imbalanced cluster sizes has negligible effects on the WCR and MNW test. It does, however, affect size control of the IM test, leading to greater over-rejection.

The MNW test becomes more slightly more sensitive to imbalanced sub-clusters once we allow treatment to be correlated within sub-cluster or clusters. Carter et al. (2017) notes a similar phenomenon: cluster imbalance and correlated covariates interact to

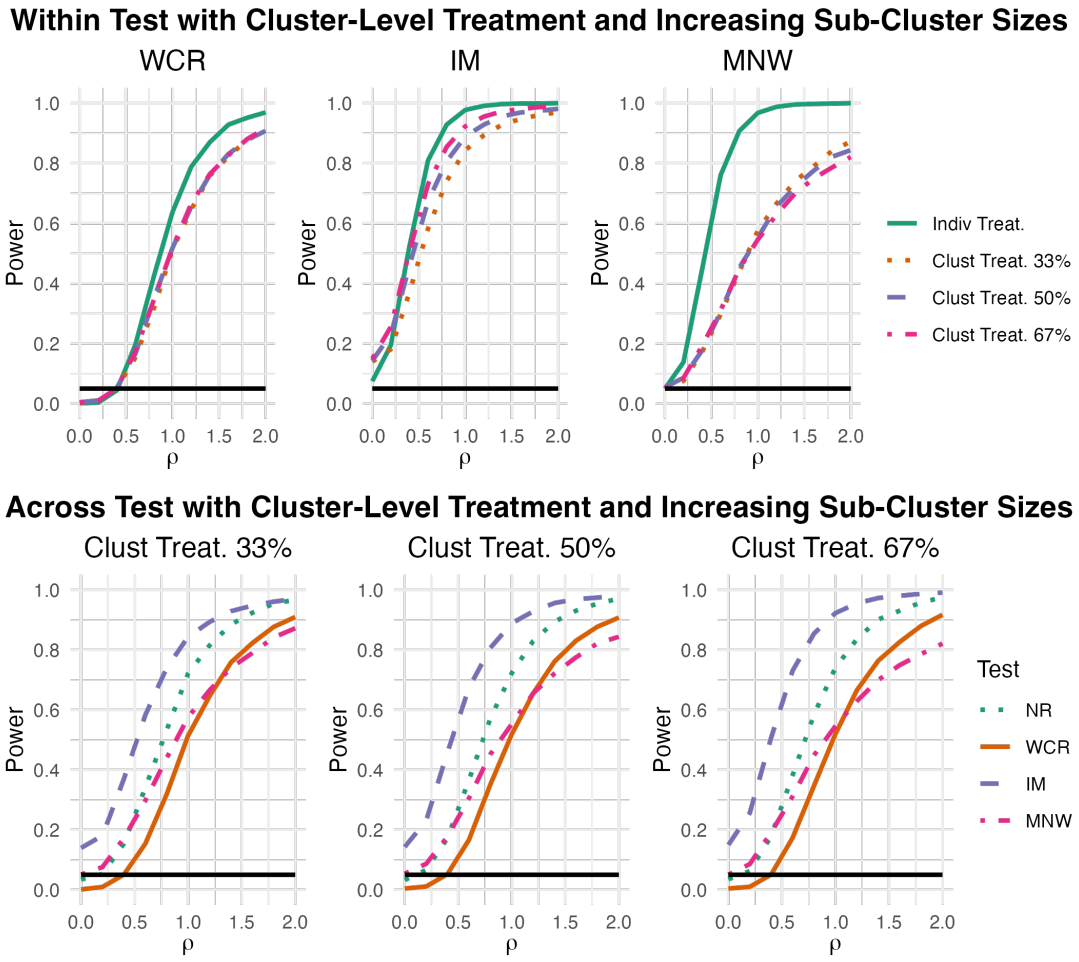

Figure 2.5. Power of various tests for level of clustering in Model 1 with treatment assignment at the cluster level. Here, $r = 8$, $q_k = 12$."Clust" refers to the case where all sub-clusters contain 100 observations. "Clust, x%" refers to the case where each cluster contains one large sub-cluster containing x% of the observations in the cluster with the remaining 11 sub-clusters having $n_j = 100$. The black line indicates the nominal size of the tests (5%).

worsen size control in cluster-robust inference using the t-test. In our setting, this effect is most pronounced with cluster-level treatment assignment, presented in Figure 2.5. Here we see that increasing the size of the outlier sub-cluster from 33% to 67% of all observations in a cluster reduces power of MNW at larger values of ρ. The performance of

the WCR test remains unchanged, so that it starts to have higher power than the MNW test at larger values of ρ. As we saw in Figure 2.4, imbalanced sub-clusters increases the IM test's type I error. This continues to be true in Figure 2.5 with cluster level treatment assignment.

All in all, the simulation evidence suggests that our test manages to maintain type I error below α when q is small, whereas the IM and NR tests may see size distortion in such a setting. The cost of size control in a fixed q setting is that the procedure is very conservative. This conservativeness limits the power of our test. However, the performance of our test is less sensitive to common sources of heterogeneity within and across clusters, such that it could become more powerful than the IM and MNW tests when clusters are heterogeneous. Indeed, as we will see in the next section, our test detects dependence in the clusters of Gneezy et al. (2019), demonstrating its potential relevance for empirical work.

2.4. Application: Gneezy et al. (2019)

In recent years, the poor performance of American students in assessment tests such as the Programme for International Student Assessment (PISA) has raised concerns among policymakers. Gneezy et al. (2019) argues that the testing gap reflects, among other things, the low effort that American students put in on tests, especially when compared to their higher scoring counterparts in other countries.

The authors test their hypothesis by a randomized controlled experiment in which students were rewarded with cash for correct answers in a 25-question test. Those assigned to the treatment group were offered roughly $1 USD per correct answer, while the control

group received no payment. Students were informed right before the test started to prevent them from changing their effort in test preparation. The experiments were conducted at 4 schools in Shanghai and 2 schools in the US. Due to logistical reasons, the authors randomized treatment at the class level for some schools and individual level for others.

Various regression analyses were conducted to study the effect of treatment on test-taking effort and test performance. Panel A in Table 3 examines whether monetary incentive increased the probability that students attempt a given question – a proxy for effort. It does so by estimating the following equation:

$$Y_{qi} = \beta Z_i + \gamma' W_i + U_{qi} \ .$$

Here, the unit of analysis is a question and Y_{qi} is an indicator for whether student i attempted question q. Z_i is the treatment indicator and W_i is a vector of control variables, which include terms such as gender, ethnicity as well as question number fixed effects. We focus on Column 1 in Panel A, which looks at US students' responses to all 25 questions in the test, and Column 4, which looks at Shanghai students' responses to the same test.

The authors present their linear regression estimate of β, together with standard errors clustered at the level of randomization. However, other levels of clustering are plausible:

- G: Group Level, that is, the level of randomization.
- S: School Level.
- SY: Experiments in Shanghai schools were conducted in 2016 and then 2018. We could plausibly interact school and year of experiment.
- ST: Schools in the US separate students into tracks (Honors, Regular, Others). We could plausibly interact school and track.

We will refer to these levels of clustering by their initials hereafter. More information on the sizes of clusters can be found in appendix B.5.

While the authors chose to cluster their standard errors by G, it seems reasonable to be concerned about correlation across individuals within the same school or among those who took the test in the same year. If these clusters were not independent, t-tests using the presented standard errors could lead to the wrong conclusions.

	Column 1			Column 4		
$\hat{\beta}$	0.037			-0.030		
	G	ST	S	G	SY	S
CCE S.E.	0.017	0.008	0.000	0.008	0.020	0.023
CCE p	0.029	0.000	0.000	0.000	0.131	0.188
Wild Bootstrap p	0.064	0.073	0.262	0.002	0.152	0.126
ART p	-	0.063	0.500	-	0.125	0.250
IM2010 p	-	0.926	0.974	-	0.748	0.816

Table 2.3. Tests for $\beta = 0$ under various levels of clustering. Based on the regressions in Table 3 Panel A of Gneezy et al. (2019).

Table 2.3 presents the OLS estimates from Gneezy et al. (2019) as well as the p-values that would be obtained from testing the null hypothesis that $\beta = 0$ using several methods. Specifically, we consider the wild cluster bootstrap (Cameron et al. (2008a)), approximate randomization tests (Canay et al. (2017a)) and the t-distribution based procedure of Ibragimov and Müller (2010), denoted IM2010. We perform these tests using the various plausible levels of clustering. For Column 1, we consider the increasingly coarse levels of clustering G, ST and S. For the US, there are no schools sampled over multiple years, so SY is the same as S. For Column 4, we consider the increasingly coarse levels of

clustering G, SY and S. In Shanghai schools, students are not separated by track, so ST is the same as S.

Remark 2.11. Gneezy et al. (2019) present clustered standard errors but do not use them for inference. Instead, they conduct randomization inference by permuting treatment status as in Young (2019). This procedure tests the null hypothesis that the distribution of the Y_{qi}'s are the same with and without treatment. This is a stronger null hypothesis than the null of 0 average treatment effect ($\beta = 0$). We believe that the latter hypothesis is typically the one of interest and test it in our Table 2.3.

Turning to the results, for column 1, we see that CCE SE's decrease as we move to increasingly coarse levels of clustering. Correspondingly, p-values from CCE-based t-tests decrease as we coarsen the clusters. Such a pattern is typically interpreted as arising from the downward bias of CCEs with few clusters (Angrist and Pischke (2008)), so that these p-values would be considered unreliable. Faced with downward bias, practitioners commonly turn to the wild cluster bootstrap. With this method, the p-values increase as we coarsen the clusters. While clustering at G and ST may lead one to conclude that there is strong evidence that $\beta \neq 0$, the p-value at S suggests the absence of strong evidence. The same phenomenon arises with approximate randomization tests: at ST there appears to be strong evidence that $\beta \neq 0$. At S, this is no longer true. With IM2010, the test does not reject in either case. We note that ART and IM2010 cannot be applied with G as the chosen level of clustering, since both methods require β to be estimated cluster-by-cluster. The results for column 4 are qualitatively similar. At G, CCE-based t-test and the wild

cluster bootstrap find strong evidence that $\beta \neq 0$. This conclusion is overturned once we cluster at either SY or S.

	Column 1			Column 4		
	G → ST	G → S	ST → S	G → SY	G → S	SY → S
WCR	0.891	1.000	1.000	0.062	0.056	1.000
IM	0.817	0.868	0.673	0.000	0.006	0.019
MNW	0.266	0.228	0.145	0.000	0.003	0.624

Table 2.4. Tests of levels of clustering applied to the regression in Table 3 Panel A of Gneezy et al. (2019).

To assess the validity of the above specifications, we apply our WCR test, the IM test and the MNW tests. Table 2.4 presents the resulting p-values. The notation $G \to S$ means that the null hypothesis involves sub-clusters G and coarse clusters S. For Column 1, clustering at G appears to be appropriate, as all 3 tests fail to reject the null hypotheses $G \to ST$ and $G \to S$. For Column 4, all 3 tests find strong evidence that sub-clusters G are inappropriate. The WCR test has higher p-values than the IM and MNW tests, likely due to its lower power. Nonetheless, they are close to 5%. The WCR and MNW tests do not reject the null hypothesis for $SY \to S$, whereas the IM test does. Given the that there are at most 2 school×year per school, the IM test is likely to over-reject. As such, we consider the conclusion of the WCR and MNW test to be more reliable in this instance. Thus, results based on clustering at SY are plausible.

All in all, we see that settings with varying numbers of clusters and sub-clusters arise in empirical work. Our test, designed for applications with few clusters and sub-clusters is relevant and appears to work well in practical settings.

2.5. Conclusion

We propose to test for the level of clustering in a regression by means of a modified randomization test. We show that the test controls size even when the number of clusters and sub-clusters are small, provided that the size of sub-clusters are relatively large. This is a challenging situation not accommodated by existing tests. To ensure size control, our procedure may be conservative when clusters are homogeneous. However, in settings with heterogeneous clusters, it has power that is comparable with other tests. As such, our test can be useful when the researcher faces an application with few sub-clusters, particularly when these clusters are likely to be heterogeneous. Finally, we note that the test is easy to implement and could serve as a helpful robustness check to researchers working with clustered data. An R package is available from the author's website.

CHAPTER 3

On the Implementation of Approximate Randomization Tests

3.1. Introduction

This chapter provides a user's guide to the general theory of approximate randomization tests (ARTs) developed in Canay et al. (2017b) when specialized to linear regressions with clustered data. Here, clustered data refers to data that may be grouped so that there may be dependence within each cluster, but distinct clusters are approximately independent in a way to be made precise below. Such data is remarkably common, including not only data that are naturally grouped into clusters, such as villages or repeated observations over time on individual units, but also data with weak temporal dependence, in which pseudo-clusters may be formed using blocks of consecutive observations. An important feature of the methodology is that it applies to commonly encountered settings in which the number of clusters is small – even as small as five. In this respect, the proposed methodology contrasts sharply and meaningfully with many commonly employed methods for inference in such settings. We briefly elaborate on this point in our discussion of related literature below.

A principal goal of this paper is to make the general theory developed in Canay et al. (2017b) more accessible by providing a step-by-step algorithmic description of how to implement the test and construct confidence intervals for the quantity of interest in these types of settings. In order to do so, we develop three novel results concerning the

methodology in Section 3.3. Our first result shows that what we view as the most natural implementation of the test, as described in Algorithm 3.1, is numerically equivalent to an alternative implementation based on weighted scores (see Algorithm 3.3). Our second result shows that when the parameter of interest is a scalar parameter, studentizing or not the t-statistic entering the test does not affect the results of the test or the associated confidence intervals. We therefore focus on the unstudentized statistic in Algorithm 3.1. Finally, our third result shows that the confidence sets for scalar parameters that are conceptually described by test inversion are indeed a closed interval of the real line. This further leads to a simple closed form expression for the lower and upper bound of the confidence intervals (see Algorithm 3.3). These results are new to this chapter and play an important role in developing simple algorithms for the implementation of ARTs.

We additionally provide a discussion of the main requirements underlying the test in Section 3.4. These requirements essentially demand that the quantity of interest is suitably estimable cluster-by-cluster. As discussed further in Section 3.4, when this is not satisfied, a researcher need not conclude that it is not possible to exploit the results in Canay et al. (2017b). Instead, several remedies are possible, including clustering more coarsely or changing the specification to ensure that this requirement is satisfied. We provide two applications that further elucidate these points: one to a linear regression with clustered data based on Meng et al. (2015) and a second to a linear regression with temporally dependent data based on Munyo and Rossi (2015). The required software to replicate these empirical exercises and to aid researchers wishing to employ the methods elsewhere is provided in both `R` and `Stata`.[1]

[1]The `Stata` and `R` packages `ARTs` can be downloaded from http://sites.northwestern.edu/iac879/software/.

The methodology described in this chapter is part of a large and active literature on inference with clustered data. Following Bertrand et al. (2004), researchers are acutely aware of the need to adjust inferences appropriately to account for this sort of dependence. Many of the most commonly employed methods for doing so, however, are inadequate for the unusually common situation in which the number of clusters is small. Conventional wisdom suggests that the number of clusters is small when it is less than forty. For example, the method described in Liang and Zeger (1986b), which has enjoyed considerable popularity due to its availability in software packages such as `Stata`, is widely acknowledged to perform poorly when this rule-of-thumb is not satisfied. Similarly, the cluster wild bootstrap described in Cameron et al. (2008b) requires either a sufficiently large number of clusters or, as shown by Canay et al. (2021a), stringent homogeneity across clusters, to perform reliably. As explained further in Section 3.4, the methods developed in Canay et al. (2017b) and described in this chapter, require neither a large number of clusters nor such homogeneity across clusters. We note that the methods by Ibragimov and Müller (2010, 2016), which are closely related to the ones described here, also do not require such restrictions, but are generally less powerful and permit testing a less rich variety of hypotheses. See Canay et al. (2017b) for further discussion of these points as well as Conley et al. (2018) for an insightful and thorough review of the related literature more broadly.

The remainder of this chapter is organized as follows. In Section 3.2, we first formalize the setting and establish some notation. We then describe the implementation of approximate randomization tests (ARTs) in an algorithmic fashion, including how to use these tests to construct confidence intervals for the quantity of interest. In Section 3.3

we present three results that play an important role in developing these algorithms. In Section 3.4, we articulate the main requirements underlying the tests and discuss remedies for cases where these requirements are not satisfied. Our two empirical applications are contained in Section 3.5. Finally, we provide some concluding remarks in Section 3.6.

3.2. Review of ARTs in regression models

We start by reviewing the inference approach proposed by Canay et al. (2017b) in the context of a linear regression model with clustered data. In order to do so, we index clusters by $j \in J \equiv \{1, \ldots, q\}$ and units in the jth cluster by $i \in I_{n,j} \equiv \{1, \ldots, n_j\}$. We also denote by $n = \sum_{j=1}^{q} n_j$ the total number of observations. The observed data consists of an outcome of interest, $Y_{i,j}$, and a vector of covariates, $Z_{i,j} \in \mathbf{R}^{d_z}$, that are related through the equation

$$Y_{i,j} = Z_{i,j}'\beta + \epsilon_{i,j} \ , \tag{3.1}$$

where $\beta \in \mathbf{R}^{d_z}$ are unknown parameters and our requirements on $\epsilon_{i,j}$ are explained below in Section 3.4. Our goal is to test

$$H_0 : c'\beta = \lambda \quad \text{vs.} \quad H_1 : c'\beta \neq \lambda \ , \tag{3.2}$$

for given values of $c \in \mathbf{R}^{d_z}$ and $\lambda \in \mathbf{R}$, at level $\alpha \in (0, 1)$. An important special case of this framework is a test of the null hypothesis that a particular component of β equals a given value, i.e.,

$$H_0 : \beta_\ell = \lambda \quad \text{vs.} \quad H_1 : \beta_\ell \neq \lambda \ ,$$

for some $\ell \in \{1, \ldots, d_z\}$, by simply setting c to be a standard unit vector with a one in the ℓth component and zeros otherwise. More generally, the approach we describe below extends immediately to the case where the hypothesis of interest involves multiple elements of β, in which case the test becomes

$$H_0 : R\beta = \Lambda \quad \text{vs.} \quad H_1 : R\beta \neq \Lambda \ , \tag{3.3}$$

for a given $p \times d_z$-dimensional matrix R and p-dimensional vector Λ, at level $\alpha \in (0, 1)$.

ARTs were developed more generally in Canay et al. (2017b) and admit a variety of different applications that go beyond the linear model considered here. For example, the method accommodates non-linear models, non-linear hypotheses, or even applications that go beyond inference with a small number of clusters (e.g., Canay and Kamat (2018) develop a variation that applies to inference in the regression discontinuity design). Here, we abstract away from the generality of the method and focus on the steps needed to use ARTs to test the null hypothesis in (3.2) in the context of the model in (3.1).

3.2.1. How to implement ARTs

The most straightforward way to test the hypotheses in (3.2) via ARTs is by following the steps described in Algorithm 3.1 below.

Algorithm 3.1 (ARTs via within-cluster estimates)**.** This implementation of ARTs involves the following steps:

Step 1: For each cluster $j \in J$, run an ordinary least squares regression of $Y_{i,j}$ on $Z_{i,j}$ using the n_j observations in cluster j. Denote the corresponding estimators

of β by

$$\{\hat{\beta}_{n,j} : j \in J\} \ .$$

Step 2: For each $j \in J$, define the random variables

$$S_{n,j} \equiv \sqrt{n_j}(c'\hat{\beta}_{n,j} - \lambda) \ , \tag{3.4}$$

and then construct the test statistic

$$T_n = \left|\frac{1}{q}\sum_{j=1}^{q} S_{n,j}\right| \ . \tag{3.5}$$

Step 3: Let $\mathbf{G} = \{1, -1\}^q$, so $g = (g_1, \ldots, g_q) \in \mathbf{G}$ is simply a q-dimensional vector with elements g_j being either 1 or -1. For any element $g \in \mathbf{G}$, define

$$T_n(g) = \left|\frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}\right| \ . \tag{3.6}$$

Step 4: Compute the $1-\alpha$ quantile of $\{T_n(g) : g \in \mathbf{G}\}$ as

$$\hat{c}_n(1-\alpha) \equiv \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_n(g) \leq u\} \geq 1-\alpha\right\} \ . \tag{3.7}$$

Step 5: Compute the test as

$$\phi_n \equiv I\{T_n > \hat{c}_n(1-\alpha)\} \ , \tag{3.8}$$

where T_n is as in (3.5) and $\hat{c}_n(1-\alpha)$ is as in (3.7). The associated p-value is

$$\hat{p}_n \equiv \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{T_n(g) \geq T_n\} \ , \tag{3.9}$$

where $T_n(g)$ is as in (3.6).

Algorithm 3.1 involves five steps that are easy to implement from a computational standpoint, but some of the steps deserve some clarification. Step 1 involves q within-cluster regressions that lead to q estimates of β. This essentially demands that the parameter β is identified cluster-by-cluster, and may fail to hold if some of the variables in the vector $Z_{i,j}$ are constant within cluster. We discuss possible remedies for this problem in Section 3.4 and illustrate their use in one of the applications in Section 3.5. An important feature of the method is that from Step 2 onwards, the original data is no longer needed as all calculations only involve the q estimators of the parameter β obtained in Step 1.

Step 2 defines a type of unstudentized t-statistic that is appropriate for the null hypothesis in (3.2). We discuss the connection to its studentized version in Section 3.3.2 below. If the null hypothesis of interest is the one in (3.3), then a Wald-type test statistic could be used instead, i.e.,

$$T_n^{\text{wald}} \equiv q\Big(\frac{1}{q}\sum_{j=1}^{q} S_{n,j}\Big)' \Sigma_S^{-1} \Big(\frac{1}{q}\sum_{j=1}^{q} S_{n,j}\Big) \ , \tag{3.10}$$

where

$$S_{n,j} \equiv \sqrt{n}(R\hat{\beta}_{n,j} - \Lambda) \quad \text{and} \quad \Sigma_S \equiv \frac{1}{q}\sum_{j=1}^{q} S_{n,j} S_{n,j}' \ .$$

Step 3 does not require one to recompute the estimates of β. It rather uses the q estimates from Step 1 and applies sign changes to the q-dimensional vector $\{S_{n,j} : j \in J\}$. Since the cardinality of $\mathbf{G}$ is $|\mathbf{G}| = 2^q$, it exceeds 2000 when $q > 10$ and in such cases it may be convenient to use a stochastic approximation. This may be done while still controlling the rejection probability under the null hypothesis (see Canay et al., 2017b,

Remark 2.2). Formally, in this case we let

$$\hat{\mathbf{G}} \equiv \{g^1, \ldots, g^B\} \,, \tag{3.11}$$

where $g^1 = \iota \equiv (1, \ldots, 1)$ is the identity vector and $g^b = (g_1^b, \ldots, g_q^b)$, for $b = 2, \ldots, B$, are i.i.d. Rademacher random variables; i.e., each g_j^b equals ± 1 with equal probability. To retain validity of the test regardless of the value of B, we require that $g^1 = \iota$. We note, however, that the power of the test may still depend on B. For this reason, we implement Algorithm 3.1 with $\hat{\mathbf{G}}$ replacing $\mathbf{G}$ everywhere and set $B = 1000$ (or any other reasonably large number chosen by the analyst).

Step 4 requires computing the $1-\alpha$ quantile of $\{T_n(g) : g \in \mathbf{G}\}$, which can be typically obtained by sorting the values of $\{T_n(g) : g \in \mathbf{G}\}$ and then taking the $\lceil |\mathbf{G}|(1-\alpha)\rceil^{\text{th}}$ highest element in the ordered list. Thus, if we denote the ordered values of $\{T_n(g) : g \in \mathbf{G}\}$ by

$$T_n^{(1)} \leq T_n^{(2)} \leq \cdots \leq T_n^{(B)} \,,$$

then we may define $\hat{c}_n(1-\alpha)$ in (3.7) as $\hat{c}_n(1-\alpha) = T^{(\lceil |\mathbf{G}|(1-\alpha)\rceil)}$. This representation suggests that the test may have trivial power for very low values of q. For example, when $\alpha = 10\%$, this problem arises if $q \leq 4$. For $q = 5$ the test already has non-trivial power and is only slightly conservative under the null. Similarly, when $\alpha = 5\%$ the test has non-trivial power for any $q \geq 6$.

Step 5 is straightforward and it provides both the test ϕ_n and the p-value $\hat{p}_n$. Each of these correspond to the non-randomized version of ARTs as opposed to their randomized counterparts (see Remark 2.4 in Canay et al., 2017b) since practitioners often prefer tests

that do not involve exogenous randomness. In any case, the differences between the randomized and non-randomized versions of the test have been found to be minimal in simulations (see, e.g., Canay et al., 2017b).

3.2.2. How to compute confidence intervals

We now discuss how to compute confidence intervals for the parameter $c'\beta$ by developing a novel algorithm that exploits the properties derived in Section 3.3.3. As before, a particularly important case is when c selects the ℓth component of β and then the confidence set is simply a confidence interval for β_ℓ. Conceptually we can simply form the confidence set by collecting all values of $c'\beta$ that cannot be rejected by our test at level α. That is, for the test ϕ_n in (3.8) we define

$$C_n = \{\lambda \in \mathbf{R} : \phi_n = 0 \text{ when testing } H_0 : c'\beta = \lambda\} \ . \tag{3.12}$$

In an asymptotic framework where $n \to \infty$ while q remains fixed, Canay et al. (2017b) show that ϕ_n is asymptotically level α under H_0. It follows from that result that, by construction, C_n covers $c'\beta$ with probability at least equal to $1-\alpha$ asymptotically. In Section 3.3.3 we show that C_n is indeed a closed interval in $\mathbf{R}$ and so it takes the form

$$C_n = [\lambda_l, \lambda_u] \ , \tag{3.13}$$

where λ_l is the *smallest* value of λ that cannot be rejected by ϕ_n and λ_u is the *largest* value of λ that cannot be rejected by ϕ_n. The analysis in Section 3.3.3 also reveals that λ_l and λ_u admit simple closed-form representations that we exploit to develop Algorithm 3.2 below.

Algorithm 3.2 (ART-based confidence intervals for $c'\beta$)**.** For $\{\hat{\beta}_{n,j} : j \in J\}$ as defined in Step 1 of Algorithm 3.1, the construction of the confidence interval involves the following steps:

Step 1: For every $g \in \mathbf{G}$, compute the following objects,

$$a(g) \equiv \frac{1}{q}\sum_{j=1}^{q}\sqrt{n_j}g_j \ , \quad b(g) \equiv \frac{1}{q}\sum_{j=1}^{q}\sqrt{n_j}g_j c'\hat{\beta}_{n,j} \ , \quad \text{and} \quad \lambda_0 \equiv \frac{b(\iota)}{a(\iota)} \ , \tag{3.14}$$

where $\iota = (1, \ldots, 1) \in \mathbf{G}$ is the vector with all ones.

Step 2: For every $g \in \mathbf{G}$ define

$$\lambda_l(g) \equiv \begin{cases} \frac{b(\iota)}{a(\iota)}\frac{|a(\iota)|}{|a(\iota)|+|a(g)|} + \frac{b(g)}{a(g)}\frac{|a(g)|}{|a(\iota)|+|a(g)|} & \text{if } \frac{b(g)}{a(g)} \leq \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\iota)}{a(\iota)}\frac{|a(\iota)|}{|a(\iota)|-|a(g)|} - \frac{b(g)}{a(g)}\frac{|a(g)|}{|a(\iota)|-|a(g)|} & \text{if } \frac{b(g)}{a(g)} > \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\iota)}{a(\iota)} - \frac{|b(g)|}{a(\iota)} & \text{if } |a(g)| = 0 \\ -\infty & \text{if } g = \pm\iota \end{cases} . \tag{3.15}$$

and

$$\lambda_u(g) \equiv \begin{cases} \frac{b(\iota)}{a(\iota)}\frac{|a(\iota)|}{|a(\iota)|+|a(g)|} + \frac{b(g)}{a(g)}\frac{|a(g)|}{|a(\iota)|+|a(g)|} & \text{if } \frac{b(g)}{a(g)} \geq \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\iota)}{a(\iota)}\frac{|a(\iota)|}{|a(\iota)|-|a(g)|} - \frac{b(g)}{a(g)}\frac{|a(g)|}{|a(\iota)|-|a(g)|} & \text{if } \frac{b(g)}{a(g)} < \lambda_0 \text{ and } |a(g)| \neq 0 \\ \frac{b(\iota)}{a(\iota)} + \frac{|b(g)|}{a(\iota)} & \text{if } |a(g)| = 0 \\ +\infty & \text{if } g = \pm\iota \end{cases} . \tag{3.16}$$

Step 3: Compute the lower bound λ_l in the confidence interval (3.13) as the α quantile of $\{\lambda_l(g) : g \in \mathbf{G}\}$, i.e.,

$$\lambda_l \equiv \inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{\lambda_l(g) \leq u\} \geq \alpha\right\} . \tag{3.17}$$

Compute the upper bound λ_u in the confidence interval (3.13) as the negative of the α quantile of $\{-\lambda_u(g) : g \in \mathbf{G}\}$, i.e.,

$$\lambda_u \equiv -\inf\left\{u \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\{-\lambda_u(g) \leq u\} \geq \alpha\right\} . \tag{3.18}$$

Report the confidence interval C_n as in (3.13).

Algorithm 3.2 requires three steps that are straightforward to compute and that exploit the results in Section 3.3.3. We refer the reader to that section for the details on why λ_l and λ_u admit the expressions in (3.17) and (3.18), respectively.

3.3. Three results on implementation of ARTs

Before we review the main requirement underlying ARTs, we present three properties related to the implementation of ARTs that we believe practitioners should be aware of and that are novel to this chapter. The first property establishes a connection between the implementation of ARTs as described in Algorithm 3.1 and an alternative implementation based on weighted scores. The second property establishes the numerical equivalence of ARTs for the null in (3.2) when the test statistics in (3.5) is replaced by its studentized version. The third and final result shows that ARTs confidence set for $c'\beta$ is indeed a

closed interval in $\mathbf{R}$ and provides a representation for the upper and lower bounds of the interval that lead to Algorithm 3.2.

3.3.1. Equivalence with weighted scores

It turns out that ARTs can be implemented by an algorithm that does not involve estimating the parameter β within each cluster. This alternative algorithm involves replacing Steps 1 and 2 in Algorithm 3.1 by the two alternative steps described in Algorithm 3.3 below, while keeping Steps 3 to 5 unaffected.

Algorithm 3.3 (ARTs via within-cluster weighted scores)**.** This implementation of ARTs involves the following steps:

Step 1′: Run a full-sample least squares regression of $Y_{i,j}$ on $Z_{i,j}$ subject to the restriction imposed by the null hypothesis, i.e., $c'\beta = \lambda$. Denote by $\hat{\epsilon}^{\mathrm{r}}_{i,j}$ the restricted residuals from this regression and by $\hat{\beta}^{\mathrm{r}}_{n}$ the restricted LS estimator of β.

Step 2′: For each cluster $j \in J$, define

$$S_{n,j} \equiv c'\hat{\Omega}^{-1}_{n,j}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}\hat{\epsilon}^{\mathrm{r}}_{i,j} \ , \tag{3.19}$$

where

$$\hat{\Omega}_{n,j} \equiv \frac{1}{n_j}\sum_{i\in I_{n,j}} Z_{i,j}Z'_{i,j} \tag{3.20}$$

is a $d_z \times d_z$ matrix that is assumed to be full rank with inverse $\hat{\Omega}^{-1}_{n,j}$.

Steps 3-5: Same as in Algorithm 3.1.

Note that Steps 3-5 remain unchanged given the alternative definition of $S_{n,j}$ in Step 2′. When it comes to Steps 1 and 2, there are two differences worth discussing. The first difference is that Step 1′ requires a single full-sample restricted least squares estimator of β as opposed to the q cluster-by-cluster estimators in Step 1 of Algorithm 3.1. The second difference is that Step 2′ is based on within-cluster weighted scores as opposed to the centered within-cluster estimates of β in Step 2 of Algorithm 3.1. Interestingly, these two implementations are numerically equivalent and so implementing ARTs via Algorithm 3.1 or Algorithm 3.3 leads to identical results. To see this formally, it is enough to show that $S_{n,j}$ as defined in (3.4) and (3.19) are the same using the following argument. For each $j \in J$,

$$\begin{aligned}
S_{n,j} &\equiv c'\hat{\Omega}_{n,j}^{-1}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}\hat{\epsilon}_{i,j}^{\mathrm{r}} \\
&= c'\hat{\Omega}_{n,j}^{-1}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}(Y_{i,j} - Z_{i,j}'\hat{\beta}_n^{\mathrm{r}}) \\
&= c'\hat{\Omega}_{n,j}^{-1}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}Y_{i,j} - c'\hat{\Omega}_{n,j}^{-1}\frac{1}{\sqrt{n_j}}\sum_{i\in I_{n,j}} Z_{i,j}Z_{i,j}'\hat{\beta}_n^{\mathrm{r}} \\
&= \sqrt{n_j}(c'\hat{\beta}_{n,j} - c'\beta) - \sqrt{n_j}(c'\hat{\beta}_n^{\mathrm{r}} - c'\beta) \\
&= \sqrt{n_j}(c'\hat{\beta}_{n,j} - \lambda)\ ,
\end{aligned}$$

where the fourth equality follows by adding and subtracting $\sqrt{n_j}c'\beta$ and the last equality holds because $c'\hat{\beta}_n^{\mathrm{r}} = c'\beta = \lambda$ under the null hypothesis in (3.2). It thus follows that $S_{n,j}$ in (3.4) and in (3.19) are identical and so ARTs can be alternatively implemented via Algorithm 3.1 or 3.3. The following lemma summarizes our discussion above:

Lemma 3.1. Let $\hat{\Omega}_{n,j}$ in (3.20) be full rank for each $j \in J$. Denote by C_n a confidence interval for $c'\beta$ computed using Algorithm 3.1 and by C_n' a confidence interval for $c'\beta$ computed using Algorithm 3.3. Then $C_n = C_n'$.

3.3.2. Equivalence with studentized version of the *t*-statistic

The ART defined in (3.8) of Algorithm 3.1 is based on the unstudentized test statistic T_n defined in (3.5). It may perhaps appear more desirable to instead consider the studentized version of this test statistic as studentization commonly improves performance in a variety of other settings. Here, we prove that this is not the case for ARTs when the null hypothesis is the one in (3.2) and that both versions of the test statistic lead to numerically identical results.

To see this, start by defining the studentized version of the test statistic in (3.5) as $T_n^{\mathrm{s}} \equiv T_n^{\mathrm{s}}(\iota)$, where for each $g \in \mathbf{G}$,

$$T_n^{\mathrm{s}}(g) \equiv \sqrt{q}\frac{\left|\frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}\right|}{\hat{\sigma}_{\mathrm{s}}(g)} \quad \text{and} \quad \hat{\sigma}_{\mathrm{s}}(g) \equiv \sqrt{\frac{1}{q}\sum_{j=1}^{q}\left(g_j S_{n,j} - \frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}\right)^2} \,. \tag{3.21}$$

Then note that

$$\hat{\sigma}_{\mathrm{s}}^2(g) = \frac{1}{q}\sum_{j=1}^{q} g_j^2 S_{n,j}^2 - \left(\frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}\right)^2 = V_n - T_n^2(g) \,,$$

where $V_n \equiv \frac{1}{q}\sum_{j=1}^{q} g_j^2 S_{n,j}^2$ does not depend on g as $g_j^2 = 1$ for all $j \in J$. It follows that we can write the studentized test statistic as

$$T_n^{\mathrm{s}}(g) = \sqrt{q}\frac{T_n(g)}{\sqrt{V_n - T_n^2(g)}} \,.$$

Since the function $x \mapsto \frac{x}{\sqrt{1-x^2}}$ is strictly increasing for $x \in [0,1)$, it follows that $T_n^{\mathrm{s}}(g)$ is a strictly monotonic transformation of $T_n(g)$ for each $g \in \mathbf{G}$. We conclude that $I\{T_n(g) \geq T_n(\iota)\} = I\{T_n^{\mathrm{s}}(g) \geq T_n^{\mathrm{s}}(\iota)\}$ for all $g \in \mathbf{G}$ and so the ART based on $T_n(g)$ and $T_n^{\mathrm{s}}(g)$ are identical. This discussion is summarized in the following lemma:

Lemma 3.2. Let $\hat{\Omega}_{n,j}$ in (3.20) be full rank for each $j \in J$. Denote by C_n a confidence interval for $c'\beta$ computed using Algorithm 3.1 and by C_n' a confidence interval for $c'\beta$ computed using Algorithm 3.1 with T_n^{s} in place of T_n and $T_n^{\mathrm{s}}(g)$ in place of $T_n(g)$. Here, $T_n^{\mathrm{s}}(g)$ is given by (3.21) and T_n^{s} is understood to be $T_n^{\mathrm{s}}(\iota)$, where ι is the identity transformation. Then, $C_n = C_n'$.

3.3.3. Convexity of the confidence intervals

The ART-based confidence intervals for $c'\beta$ defined in (3.12) can be computed by test inversion. From a computational standpoint, however, computing confidence sets by test inversion may be cumbersome and the resulting set may not even be an interval. That is, it may not be closed and convex. In this section we prove that this is not a concern for ART-based confidence intervals for $c'\beta$ and so such confidence intervals could be easily computed by a standard bisection algorithm. In fact, our results go even further. We derive closed form expressions for the lower and upper bounds of the confidence interval that imply that computing ART-based confidence intervals for $c'\beta$ is straightforward from a computational standpoint. In order to derive these results, we slightly change our notation to make explicit the dependence on λ of each of the elements entering the test

in (3.8). To this end, let

$$T_n(g,\lambda) \equiv \left|\frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}(\lambda)\right| \quad \text{where} \quad S_{n,j}(\lambda) = \sqrt{n_j}(c'\hat{\beta}_{n,j} - \lambda) \ ,$$

and note that $T_n = T_n(\iota, \lambda)$. Using this notation, we can re-write the confidence interval in (3.12) as

$$C_n = \left\{ \lambda \in \mathbf{R} : \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\left\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\right\} \geq \alpha \right\} \ ,$$

which is simply the values of λ for which the p-value of the test, as defined in (3.9), is not below α. In order to show that this confidence set is a closed interval, we claim that the p-value

$$\hat{p}_n(\lambda) = \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\left\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\right\} \tag{3.22}$$

is equal to 1 for $\lambda_0 \equiv b(\iota)/a(\iota)$, monotonically increasing for any $\lambda < \lambda_0$, and monotonically decreasing for any $\lambda > \lambda_0$. The next lemma formalizes this result.

Lemma 3.3. Let $\hat{\Omega}_{n,j}$ in (3.20) be full rank for each $j \in J$. Let $a(g)$, $b(g)$, and λ_0 be defined as in (3.14). The p-value in (3.22) equals

$$\hat{p}_n(\lambda) = \begin{cases} \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\left\{\lambda \geq \lambda_l(g)\right\} & \text{for } \lambda < \lambda_0 \\ 1 & \text{for } \lambda = \lambda_0 \\ \frac{1}{|\mathbf{G}|}\sum_{g\in\mathbf{G}} I\left\{\lambda \leq \lambda_u(g)\right\} & \text{for } \lambda > \lambda_0 \end{cases} \ , \tag{3.23}$$

where $\{\lambda_u(g) : g \in \mathbf{G}\}$ and $\{\lambda_l(g) : g \in \mathbf{G}\}$ are defined in Algorithm 3.2.

Proof. It is useful to re-write $T_n(g,\lambda)$ in terms of $a(g)$ and $b(g)$. To this end, note that

$$T_n(g,\lambda) \equiv \left|\frac{1}{q}\sum_{j=1}^{q} g_j S_{n,j}(\lambda)\right| = \left|\frac{1}{q}\sum_{j=1}^{q} g_j\sqrt{n_j}c'\hat{\beta}_{n,j} - \lambda\frac{1}{q}\sum_{j=1}^{q} g_j\sqrt{n_j}\right|$$

$$= |b(g) - \lambda a(g)| \; . \tag{3.24}$$

Given $g \in \mathbf{G}$ and $a(g) \neq 0$, $T_n(g,\lambda)$ is a "V-shaped" function of λ taking the value 0 at $\frac{b(g)}{a(g)}$ and with slope $-|a(g)|$ for all $\lambda < \frac{b(g)}{a(g)}$ and slope $|a(g)| \leq a(\iota)$ for all $\lambda > \frac{b(g)}{a(g)}$. Figure 3.1 illustrates this for three values of g.

First, note that $I\{T_n(g,\lambda_0) \geq T_n(\iota,\lambda_0)\} = I\{T_n(g,\lambda_0) \geq 0\} = 1$ for all $g \in \mathbf{G}$ and so it follows immediately that $\hat{p}_n(\lambda_0) = 1$.

Second, restrict attention to the set $\Lambda^+ \equiv \{\lambda \in \mathbf{R} : \lambda > \lambda_0\}$ where $T_n(\iota,\lambda)$ is linearly increasing. In order to prove that $\hat{p}_n(\lambda)$ takes the form in (3.23) we prove that $I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = I\{\lambda \leq \lambda_u(g)\}$ for each $g \in \mathbf{G}$ by dividing the argument into three cases.

Case 1: Consider $g \in \mathbf{G}$ such that $a(g) \neq 0$ and $|a(g)| \neq a(\iota)$. Since $|a(g)| < a(\iota)$, it follows that $T_n(g,\lambda)$ and $T_n(\iota,\lambda)$ intersect only once on Λ^+ and this holds regardless of whether $\frac{b(g)}{a(g)} < \lambda_0$ or $\frac{b(g)}{a(g)} \geq \lambda_0$ (see Figure 3.1 for a graphical illustration of each of these cases). Denote the intersection point by $\lambda_u(g)$ and note that $T_n(g,\lambda) \geq T_n(\iota,\lambda)$ for all $\lambda_0 < \lambda \leq \lambda_u(g)$ and $T_n(g,\lambda) < T_n(\iota,\lambda)$ for all $\lambda > \lambda_u(g)$. Conclude that on Λ^+,

$$I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = I\{\lambda \leq \lambda_u(g)\} \; . \tag{3.25}$$

Simple algebra shows that the intersection point $\lambda_u(g)$ takes the form in (3.16).

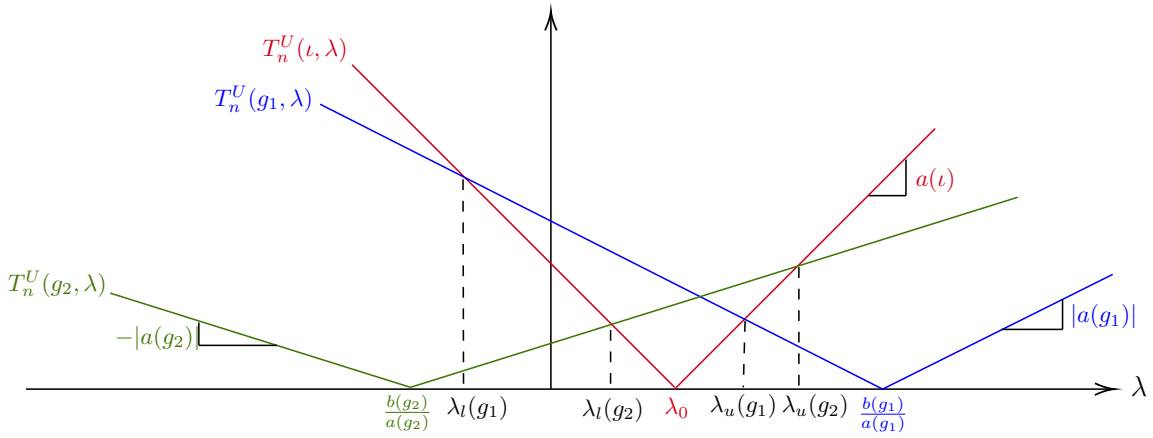

Figure 3.1. $T_n(g,\lambda)$ as functions of λ for $g \in \{\iota, g_1, g_2\}$.

Case 2: Consider $g \in \mathbf{G}$ such that $a(g) = 0$. Note that $b(\iota) - \lambda a(\iota) < 0$ for $\lambda \in \Lambda^+$. It thus follows that for $\lambda \in \Lambda^+$,

$$I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = I\{|b(g)| \geq |b(\iota) - \lambda a(\iota)|\} = I\left\{\lambda \leq \frac{b(\iota)}{a(\iota)} + \frac{|b(g)|}{a(\iota)}\right\},$$

and so (3.25) holds in this case with $\lambda_u(g) = \frac{b(\iota)}{a(\iota)} + \frac{|b(g)|}{a(\iota)}$, as defined in (3.16).

Case 3: Consider $g \in \mathbf{G}$ such that $|a(g)| = a(\iota)$ and so $g = \pm\iota$. If $g = \iota$, $I\{T_n(g,\lambda) \geq T_n(\iota,\lambda)\} = 1$ for all $\lambda \in \mathbf{R}$. We conclude that (3.25) holds with $\lambda_u(g) = \infty$. If $g = -\iota$, then we have that $a(-\iota) = -a(\iota)$ and $b(-\iota) = -b(\iota)$ so that $\frac{b(-\iota)}{a(-\iota)} = \lambda_0$ and again $I\{T_n(-\iota,\lambda) \geq T_n(\iota,\lambda)\} = 1$ for all $\lambda \in \mathbf{R}$. We conclude that (3.25) holds with $\lambda_u(g) = \infty$, as defined in (3.16). This completes the proof of (3.23) for the case $\lambda \in \Lambda^+$.

Finally, the construction for $\lambda \in \Lambda^- \equiv \{\lambda \in \mathbf{R} : \lambda < \lambda_0\}$ parallels the one for $\lambda \in \Lambda^+$ so we omit the arguments here. Putting all the cases together, (3.23) follows and this completes the proof. □

Figure 3.2 illustrates the p-value in (3.23) as a function of λ for the groups in Figure 3.1. Since $\hat{p}_n(\lambda)$ is right continuous and increasing for $\lambda < \lambda_0$, we can define λ_l as the

smallest value of λ for which $\hat{p}_n(\lambda) \geq \alpha$. Such value exists and is unique. Similar, since $\hat{p}_n(\lambda)$ is left continuous and decreasing for $\lambda > \lambda_0$, we can define λ_u as the largest value of λ for which $\hat{p}_n(\lambda) \geq \alpha$. Such value exists and is again unique. This argument leads to the representation of C_n in (3.13), showing that ART-based confidence intervals for $c'\beta$ are indeed intervals in $\mathbf{R}$. Furthermore, note that (3.23) implies that the smallest value of λ for which $\hat{p}_n(\lambda) \geq \alpha$ can be defined as

$$\inf \left\{ \lambda \in \mathbf{R} : \frac{1}{|\mathbf{G}|} \sum_{g \in \mathbf{G}} I \left\{ \lambda \geq \lambda_l(g) \right\} \geq \alpha \right\} ,$$

which is just the definition of the α quantile of $\lambda_l(g)$, as defined in Algorithm 3.2. A similar result holds for λ_u and so C_n can be computed in closed form by Algorithm 3.2.

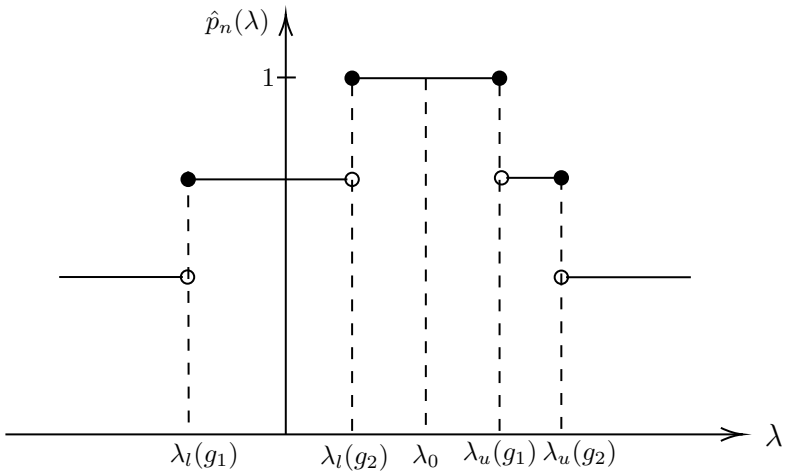

Figure 3.2. $\hat{p}_n(\lambda)$ as a function of λ.

3.4. What we need for ARTs to work

The main requirement underlying ARTs is Assumption 3.1 in Canay et al. (2017b). This assumption guarantees that the test delivers rejection probabilities under the null

hypothesis that are close to the nominal level α in an asymptotic framework where $n \to \infty$ and q remains fixed. In the context of the linear model in (3.1), this translates into the following two conditions summarized in Assumption 3.1 below.

Assumption 3.1. Let $\{\hat{\beta}_{n,j} : j \in J\}$ be the cluster-by-cluster estimators of β defined in Algorithm 3.1. Assume that:

(a) $\{\hat{\beta}_{n,j} : j \in J\}$ jointly converge in distribution at some (possibly unknown) rate; i.e.,

$$\begin{pmatrix} a_{n,1}(\hat{\beta}_{n,1} - \beta) \\ \vdots \\ a_{n,q}(\hat{\beta}_{n,q} - \beta) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} S_1 \\ \vdots \\ S_q \end{pmatrix} \tag{3.26}$$

for a sequences $a_{n,j} \to \infty$ and random variables $(S_1, \ldots, S_q)'$.

(b) The limiting random variables $(S_1, \ldots, S_q)'$ are invariant to sign changes, i.e.,

$$(g_1 S_1, \ldots, g_q S_q) \overset{d}{=} (S_1, \ldots, S_q) \ , \tag{3.27}$$

for any g in $\mathbf{G}$, where $\mathbf{G}$ is defined in Step 4 of Algorithm 3.1.

Condition (3.26) holds, for example, when $Z_{i,j}$ and $\epsilon_{i,j}$ are uncorrelated and the analyst assumes some form of weak dependence within clusters that permits the application of an appropriate central limit theorem. In such a case, (3.26) typically holds with $a_{n,j} = \sqrt{n_j}$ and each S_j being a mean-zero normal random variable. In fact, under the commonly used assumption of independent clusters, it also follows that $S_j \perp\!\!\!\perp S_{j'}$ for any $j \neq j'$. In this case the normally distributed random variables may not be identically distributed but are indeed independent. Condition (3.27), in turn, requires each S_j to be symmetrically

distributed around zero and independent of each other. This is immediately satisfied when each S_j is a mean-zero normal random variable and clusters are independent. Importantly, these assumptions allow for the normally distributed random variables to have different variances across clusters; a type of heterogeneity not allowed by the cluster wild bootstrap approach popularized by Cameron et al. (2008b) and later studied formally by Canay et al. (2021a).

Remark 3.1. The asymptotic normality in (3.26) arises frequently in applications, but is not necessary for the validity of ARTs. All that is required is that the estimators $\{a_{n,j}(\hat{\beta}_{n,j} - \beta) : j \in J\}$ have a limiting distribution that is the product of q distributions that are symmetric about zero. This may even hold in cases where the estimators have infinite variances or are inconsistent. See Canay et al. (2017b, Remark 4.5) for additional discussion on this point. □

Remark 3.2. It is worthwhile to contrast the requirements of Assumption 3.1 with those of "classical" methods, such as those described in Liang and Zeger (1986b). These latter methods permit arbitrary dependence within each cluster, but require the size of the clusters to be small and the number of clusters to be large. As described above, Assumption 3.1(a), on the other hand, permits the number of clusters to be small, but requires the size of the clusters to be large and weak dependence within each cluster. We emphasize, however, that these restrictions are commonly employed in establishing the validity of other methods in settings with a small number of clusters, including, for example, the t-test approach by Ibragimov and Müller (2010) and the wild bootstrap Canay et al. (2021a). □

Remark 3.3. We focus our exposition on the case where $Z_{i,j}$ is exogenous but we emphasize that the conditions in (3.26) and (3.27) typically hold in instrumental variable (IV) models. Accommodating IV to ARTs then only requires modifying Step 1 in Algorithm 3.1 so that the least squares regression is replaced with the appropriate IV regression. Steps 2-6 remain unaffected. □

An implicit requirement behind ARTs that deserves further comments lies in Step 1 of Algorithm 3.1, which requires that the analyst runs cluster-by-cluster regressions. This step implicitly assumes that the parameter β is identified within each cluster. In practice, this means that the matrix $\hat{\Omega}_{n,j}$ in (3.20) must be invertible for each $j \in J$ and hence the same requirement applies to Algorithm 3.3. This restriction may be substantially important in some applications and so here we discuss common ways in which the problem may manifest and two alternative remedies.

One case in which running least squares cluster-by-cluster is not feasible is when the coefficient of interest is associated with a variable that only varies across clusters. For example, consider the model in (3.1) and partition $Z_{i,j}$ into a constant term, a scalar variable that only varies across clusters, $Z_j^{(1)}$, and another variable that varies across and within clusters, $Z_{i,j}^{(2)}$. That is,

$$Y_{i,j} = \beta_0 + Z_j^{(1)}\beta_1 + Z_{i,j}^{(2)}\beta_2 + \epsilon_{i,j} \ , \tag{3.28}$$

where the analysts' interest lies in the coefficient β_1, i.e., $c'\beta = \beta_1$. Clearly, the regression in Step 1 of Algorithm 3.1 would not separately identify β_0 and β_1 as $Z_j^{(1)}$ is perfectly colinear with the constant term. The matrix $\hat{\Omega}_{n,j}$ in (3.20) is simply singular. This

situation arises, for example, in the empirical application considered by Canay et al. (2017d) where $j \in J$ indexes schools and the variable of interest is a treatment indicator at the school level. A natural remedy in a situation like this is clustering more coarsely (e.g., by combining clusters) to obtain variation within the re-defined clusters. This is possible for ARTs since the validity of the method does not rely on having a large number of clusters and thus it can afford to work with coarser clustering. In fact, in certain settings combining clusters may be quite natural. For example, Canay et al. (2017d) re-defined clusters as "pairs" of schools (as opposed to just schools) given that the treatment assignment mechanism of the experiment was a matched pairs design and so the pairs used at the randomization stage represented natural groupings. In other settings where it is less clear how to group clusters, any grouping that satisfies the requisite identification condition leads to a valid test, but it may be further desirable to combine such tests to limit concerns about "data snooping" across groupings. To this end, results in DiCiccio et al. (2020) on combining tests may be relevant.

Remark 3.4. A quick inspection of (3.28) may lead the analyst to believe there is a workaround that does not involve combining clusters if one instead uses some estimator of β_0 from a full sample regression. For example, the full sample least squares estimator $\hat{\beta}_{n,0}$ from the regression in (3.1). Then, assuming for simplicity that $Z_j^{(1)} \neq 0$ for all $j \in J$, one may consider modifying Step 1 in Algorithm 3.1 by running a regression of $Y_{i,j}$ on an intercept and $Z_{i,j}^{(2)}$ (not including $Z_j^{(1)}$) and then redefining $\{\hat{\beta}_{n,j} : j \in J\}$ as the difference between the within cluster intercept estimates, $\hat{\beta}_{j,0}$ and the full sample estimate $\hat{\beta}_{n,0}$, i.e., $\hat{\beta}_{n,j} = \hat{\beta}_{j,0} - \hat{\beta}_{n,0}$. Such strategies unfortunately introduce dependence between the q

estimators of β (as they all depend on $\hat{\beta}_{n,0}$) and thus end up violating one of the two main conditions needed for ARTs to be asymptotically valid; mainly condition (3.27). □

Another case where the lack of identification within cluster may manifest is when the variable of interest actually varies within clusters but the model specification involves other variables that are collinear with some other variable (including the variable of interest or the constant term) within clusters. For example, consider the model in (3.1) where instead of individuals indexed by $i \in I_{n,j}$, units within cluster are indexed over time $t \in T$. Partition $Z_{j,t}$ into the variable of interest, $Z_{j,t}^{(1)}$, and time fixed effects δ_t. That is,

$$Y_{j,t} = Z_{j,t}^{(1)}\beta_1 + \sum_{\tilde{t} \in T} I\{\tilde{t} = t\}\delta_{\tilde{t}} + \epsilon_{j,t} \ . \tag{3.29}$$

It then follows that, within each cluster $j \in J$, the time fixed effect δ_t absorbs all the variation in $Z_{j,t}^{(1)}$ and so β_1 is not identified. In cases like this the analyst could again combine clusters to obtain variation within the re-defined clusters. An alternative remedy is to change the specification by, for example, replacing the time fixed effect with a cluster-specific time trend. Such specification is more restrictive than the time fixed effect in the sense that it imposes a linear trend but, at the same time, is more general as it allows for heterogeneity across clusters in the linear trend. We illustrate this approach in the application we consider in Section 3.5.1.

The need to identify β within each cluster is in our view the main limitation of ARTs, but a limitation that needs to be dealt with in certain settings. One may then wonder why not simply use some other inference method that is valid when the number of clusters is small and that does not rely on estimating β cluster-by-cluster. Perhaps the most

popular approach in that category is the cluster wild bootstrap popularized by Cameron et al. (2008b) and recently studied formally by Canay et al. (2021a). While not having to estimate β within each cluster represents an advantage over ARTs, this additional flexibility comes at a cost in terms of the degree of heterogeneity that the model can deal with. In particular, the results in Canay et al. (2021a) show that the cluster wild bootstrap is expected to work well in settings with a small number of clusters *as long as* the clusters are "homogeneous," in a sense made precise in Canay et al. (2021a). Intuitively, it is required that the variance covariance matrix $\hat{\Omega}_{n,j}$ defined in (3.20) is the same across clusters (up to scalar multiplication). Such stringent homogeneity condition is not required for ARTs to work well, as the method allows clusters to be arbitrarily heterogeneous as long as $\hat{\Omega}_{n,j}$ is invertible for $j \in J$.

Remark 3.5. For ease of exposition, we have written the requirement in (3.26) in terms of the differences $\hat{\beta}_{n,j} - \beta$, but it is possible to replace it with the differences $c'\hat{\beta}_{n,j} - c'\beta$ (or $R\hat{\beta}_{n,j} - R\beta$, depending on the null hypotheses of interest). In most cases, re-writing the condition in this way is not useful, but it is in cases where $c'\beta$ is identified within each cluster while β is not. For example, consider the model in (3.28) when the coefficient of interest is β_2 as opposed to β_1, i.e, $c'\beta = \beta_2$. In that case the entire term $\beta_0 + Z_j^{(1)}\beta_1$ may be absorbed into a cluster-specific intercept without affecting the identification and estimation of $c'\beta = \beta_2$ within each cluster. □

3.5. Empirical applications

In this section we apply ARTs as described in Algorithm 3.1 and ART-based confidence intervals as described in Algorithm 3.2 in the context of two distinct empirical applications.

The `R` and `Stata` packages and codes required to replicate the results in this section are available as part of the online supplemental material.

3.5.1. Meng, Qian and Yared (2015)

Meng et al. (2015, MQY) argue that China's Great Famine, from 1959 to 1961, was the result of an inflexible food procurement policy by the central government. To make this point, they show that food production and mortality become positively correlated during the time of famine, when this coefficient is otherwise negative or not significantly different from 0 in normal times.

MQY consider the following regression,

$$Y_{j,t+1} = Z_{j,t}^{(1)}\beta_1 + Z_{j,t}^{(2)}\beta_2 + \delta_t + \epsilon_{j,t}$$

where j indexes provinces (ranging from 1 to 19) and t indexes years (ranging from 1953 to 1982). Here,

$$\begin{aligned}
Y_{j,t+1} &= \log(\text{number of deaths in province } j \text{ during year } t+1) \\
Z_{j,t}^{(1)} &= \log(\text{predicted grain production in province } j \text{ during year } t) \\
&\qquad \times I\{t \text{ is a famine year}\} \\
Z_{j,t}^{(2)} &= \log(\text{predicted grain production in province } j \text{ during year } t) \\
\delta_t &= \text{time fixed effects .}
\end{aligned}$$

In this application the level of clustering is a province, and so in order to apply ARTs as described in Section 3.2.1, one needs to estimate $\beta = (\beta_1, \beta_2)'$ and δ_t province-by-province. This illustrates one of the situations where including time fixed effects province-by-province is infeasible for the implementation of ARTs, given that the only source of remaining variation within a province is indeed time. The second identification problem described in Section 3.4 then arises. As we discussed in that section, one way to deal with this issue consists of replacing the time fixed effects with a cluster-specific time trend, i.e., in Step 1 of Algorithm 3.1 estimate

$$Y_{j,t+1} = Z_{j,t}^{(1)}\beta_1 + Z_{j,t}^{(2)}\beta_2 + \gamma_j t + \epsilon_{j,t} \ . \tag{3.30}$$

We will refer to this as Analysis #1. In addition, we also consider the following alternative specifications studied by MQY:

- Analysis #2: Repeating Analysis #1 using only data between 1953 and 1965.
- Analysis #3: Repeating Analysis #1 using four additional autonomous provinces.
- Analysis #4: Repeating Analysis #2 using four additional autonomous provinces.
- Analysis #5: Repeating Analysis #1 using actual rather than constructed grain production.
- Analysis #6: Repeating Analysis #2 using actual rather than constructed grain production.

As with Analysis #1, the above analyses differ from their MQY counterparts only in that a linear time trend $\gamma_j t$ replaces time fixed effects δ_t. Table 3.1 summarizes the number of clusters and the number of observations for each of these analyses. We caution, however, that in this application, in addition to the number of clusters being small, the number of

observations within each cluster may also be small. See Remark 3.2 for further discussion in relation to Assumption 3.1(a).

Analysis	# of Clusters	Min. Size	Med. Size	Max. Size	Mean
#1, #5	19	29	30	30	29.95
#2, #6	19	12	13	13	12.95
#3	23	29	30	30	29.96
#4	23	12	13	13	12.96

Table 3.1. Cluster Information. 'Min. Size', 'Med. Size', 'Max. Size' denote the minimum, the median, and the maximum size of clusters.

Meng et al. (2015) consider the following two null hypotheses of interest,

$$H_0^{(1)} : \beta_1 = 0 \quad \text{and} \quad H_0^{(2)} : \beta_1 + \beta_2 = 0 \ . \tag{3.31}$$

In Table 3.2 we replicate the main table in Meng et al. (2015) using cluster robust standard errors (CCE) and also include the results associated with ARTs for both $H_0^{(1)}$ and $H_0^{(2)}$ in (3.31). For $H_0^{(1)}$ we report p-values and 95% confidence intervals, while for $H_0^{(2)}$ we just report p-values following MQY. The authors note in footnote 33 that using the cluster wild bootstrap led to similar results as those presented in their main table so we do not include cluster wild bootstrap results here either.

We comment on the following main features of Table 3.2:

(1) For the null hypothesis $H_0^{(1)}$ associated with the parameter β_1, the ART p-values are of comparable magnitude to traditional CCE p-values. Similarly, ART-based confidence intervals are of roughly the same length as those obtained based on CCE although the ART-based confidence intervals do not contain the LS estimates. This is because ART-based confidence intervals are centered around the

	#1	#2	#3	#4	#5	#6
LS Estimate: β_1	0.063	0.057	0.071	0.067	0.064	0.058
CCE: Province						
se	0.007	0.007	0.007	0.008	0.007	0.007
p-value	0.000	0.000	0.000	0.000	0.000	0.000
95% CI	[0.050, 0.077]	[0.043, 0.071]	[0.057, 0.086]	[0.051, 0.083]	[0.051, 0.078]	[0.044, 0.071]
ART						
p-value	0.000	0.002	0.000	0.000	0.000	0.000
95% CI	[0.032, 0.055]	[0.018, 0.047]	[0.038, 0.066]	[0.028, 0.067]	[0.032, 0.058]	[0.029, 0.050]
$\beta_1 + \beta_2 = 0$						
CCE p-value	0.050	0.009	0.059	0.005	0.266	0.363
ART p-value	0.098	0.571	0.096	0.487	0.080	0.001
Observations	569	246	689	298	569	246
Short Sample	No	Yes	No	Yes	No	Yes
Auto. Region	No	No	Yes	Yes	No	No
Pred. Grain Prod.	Yes	Yes	Yes	Yes	No	No

Table 3.2. Results for Analyses #1-6, comparable to those in Table 2 of Meng, Qian and Yared (2015). 'LS Estimate' denotes the full sample OLS estimate for β_1. CCE refers to cluster-robust standard errors. ART p-values are obtained using Algorithm 3.1. ART-based 95% confidence intervals are obtained using Algorithm 3.2.

mean of the province-by-province estimates, which may not necessarily be equal to the full sample LS estimate of β_1.

(2) For the null hypothesis $H_0^{(2)}$ associated with the parameter $\beta_1 + \beta_2$, the ART p-value is sometimes higher and sometimes lower than the CCE p-value depending on the specification. Given the relatively small number of clusters in this application, the ART p-values are likely to be more reliable than those associated with CCE as CCE is known to perform poorly when the number of clusters is not sufficiently large.

3.5.2. Munyo and Rossi (2015)

Munyo and Rossi (2015) study criminal recidivism of former prisoners by looking at the relationship between the number of inmates released from incarceration on a given day and the number of offenses committed on the same day. They claim that the liquidity constraints that inmates face on the day of release increase the likelihood of recidivism on the same day. Using data of 2631 days between January 1st 2004 and March 15 2011 collected from the criminal incidents reports in Montevideo in Uruguay, they estimate the following linear model by least squares

$$Y_t = Z_t'\beta + \epsilon_t$$

where t indexes days and

Y_t = the total number of offenses on day t

Z_t = the total number of inmates released, temperature, rainfall, hours of sunshine on day t, a dummy for holidays, a dummy for December 31st and a yearly trend.

We refer to this as Analysis #1. Munyo and Rossi (2015) additionally consider the following four analyses:

- Analysis #2: Z_t includes a daily trend in place of a yearly trend.
- Analysis #3: Z_t includes a monthly trend in place of a yearly trend.
- Analysis #4: Z_t includes an intra-month daily trend, month- and year- level fixed effects and their interactions in place of a yearly trend.

- Analysis #5: Z_t includes month- and year- level fixed effects and their interactions in place of a yearly trend.

Analysis #5 is their preferred specification. Munyo and Rossi (2015) report the results of these analyses in Table 2 in their paper. They report least squares estimates of β with Newey-West heteroskedasticity-autocorrelation-consistent (HAC) standard errors. In addition, they report ART p-values as described in Algorithm 3.1 for the null hypothesis that $H_0 : c'\beta = 0$ as in (3.2), where c selects the coefficient on the total number of inmates released on day t.

In this application the level of clustering is not naturally determined by the data, but pseudo-clusters may be formed using blocks of consecutive observations under the assumption of weak temporal dependence. In order to apply ARTs as described in Algorithm 3.1 we then form q pseudo-clusters by dividing the data into q consecutive blocks of size $b_n = \lfloor n/q \rfloor$ where $n = 2631$ is the number of total observations. More concretely, we define the jth pseudo-cluster as

$$X_j^{(n)} = \{(Y_t, Z_t')' : t = (j-1)b_n + 1, \cdots, jb_n\} \quad \text{where} \quad j = 1, \cdots, q-1 \ ,$$

and let the last qth pseudo-cluster contain all the remaining $n - b_n(q-1)$ observations. Note that in this application the number of pseudo-clusters q is a tuning parameter that the analyst must specify. Munyo and Rossi (2015) set $q = 10$. We repeat their analyses with alternative values of q and investigate how sensitive the results are to this choice. The relevant cluster information is given in Table 3.3.

Table 3.4 shows LS estimates of β, p-values for the hypothesis in (3.2), and 95% confidence intervals for each analysis. Following Munyo and Rossi (2015), we report

# of Clusters (q)	Cluster Size
8	328
10	263
16	164

Table 3.3. Pseudo-cluster size for different values of q in Munyo and Rossi (2015).

Specification	#1	#2	#3	#4	#5
LS Estimate	0.225	0.260	0.259	0.225	0.234
HAC					
se	0.124	0.123	0.123	0.096	0.096
p-value	0.068	0.034	0.034	0.019	0.015
95% CI	[-0.017, 0.468]	[0.02, 0.5]	[0.019, 0.5]	[0.038, 0.413]	[0.046, 0.421]
ART: q=8					
p-value	0.008	0.023	0.023	0.102	0.102
95% CI	[0.124, 0.429]	[0.035, 0.391]	[0.035, 0.391]	[-0.07, 0.397]	[-0.067, 0.418]
ART: q=10					
p-value	0.002	0.014	0.014	0.063	0.053
95% CI	[0.141, 0.603]	[0.068, 0.446]	[0.068, 0.458]	[-0.023, 0.431]	[-0.003, 0.452]
ART: q=16					
p-value	0.002	0.006	0.006	0.027	0.010
95% CI	[0.131, 0.444]	[0.097, 0.369]	[0.087, 0.371]	[0.02, 0.324]	[0.056, 0.367]
Observations	2631	2631	2631	2631	2631
Time Trend	Year	Day	Month	Intra-month Day	None
Time Fixed Effect	No	No	No	Yes	Yes
Controls	No	No	No	No	No

Table 3.4. Results for Analyses #1-5, comparable to those in Table 2 of Munyo and Rossi (2015). 'LS Estimate' denotes the full sample LS estimate of β. HAC refers to the heteroskedasticity and autocorrelation consistent standard error. ART p-values are obtained using Algorithm 3.1. ART-based 95% confidence intervals are obtained using Algorithm 3.2.

results based on HAC standard errors. The table also shows ART p-values as described in Algorithm 3.1 and ART-based 95% confidence intervals as described in Algorithm 3.2 for $q = 8$, $q = 10$, and $q = 16$.

We summarize the main findings of the results in Table 3.4 as follows:

(1) The choice of q is important for the results of ARTs but currently there is no theory developed to choose this tuning parameter according to some data dependent criteria. The smaller q is, the more observations are available within each cluster. Having more observations per cluster is important for one of the requirements behind ARTs, mainly (3.26). A small value of q, however, tends to affect the power of ARTs despite not really affecting the control of the rejection probability under the null hypothesis. This feature can be seen in Table 3.4, where ARTs p-values are decreasing in q across different specifications. In this application, where there are still over a hundred observations when $q = 16$, a larger value of q like $q = 10$ or $q = 16$ may be preferable to smaller values, like $q = 8$, based on power considerations. Note, however, that except in Analyses #4–5, where the choice of q determines whether the null hypothesis is rejected at a given significance level, the results for Analyses #1–3 are in all agreement at a 5% level.

(2) Overall, the test results based on standard t-test with HAC standard errors are consistent to those of ARTs when $q = 16$. Both methods reject the null hypothesis $H_0 : c'\beta = 0$ at a 10% nominal level across different specifications. The results support the authors' argument that the release of inmates from incarceration increase the chance of re-offenses on the day of release.

3.5.3. Computational gains of the new algorithm

Tables 3.5 and 3.6 report four alternative ways to compute ART-based confidence intervals in the two empirical applications we consider in this chapter; Meng et al. (2015) and

Munyo and Rossi (2015). The first alternative is to compute the confidence intervals by a simple grid search algorithm. The second alternative involves a bi-section algorithm. We implement both of these methods using a studentized and an unstudentized test statistic to illustrate the result in Section 3.3.2. The last alternative is to simply use Algorithm 3.2, as reported in Sections 3.5.1 and 3.5.2. In each case, we also report computational times to illustrate the computational advantages of the algorithm we propose in this chapter. The `R` and `Stata` codes required to replicate the results in this section are available as part of the online supplemental material.

Starting from Table 3.5, we see that grid search take a significant amount of time to compute. Our convexity result (Lemma 3.3) facilitates the use of the bisection method, cutting implementation time by a factor of over 50. Moving from the bisection method to Algorithm 3.2 further leads to a speed up of at least 2 times. A similar pattern emerges in Table 3.6. Furthermore, comparing specification with $q = 8$ that with $q = 16$, the speed advantage of our method becomes far starker. For $q = 16$, grid search takes almost 100 times as long as the bisection method. The bisection method, meanwhile, takes close to 10 times as long as Algorithm 3.2.

3.6. Concluding remarks

The goal of this chapter is to make the general theory developed in Canay et al. (2017b) more accessible by providing a step-by-step algorithmic description of how to implement the test and construct confidence intervals in linear regression models with clustered data, as well as clarifying the main requirements and limitations of the approach. The main two takeaways are the following. First, ARTs-based confidence intervals for scalar parameters

	Grid Search		Bisection		ART
	Stud.	Unstud.	Stud.	Unstud.	
#1	[0.032, 0.055] 19.65	[0.032, 0.055] 6.66	[0.032, 0.055] 0.31	[0.032, 0.055] 0.11	[0.032, 0.055] 0.06
#2	[0.018, 0.047] 46.63	[0.018, 0.047] 16.43	[0.018, 0.047] 0.35	[0.018, 0.047] 0.12	[0.018, 0.047] 0.02
#3	[0.038, 0.066] 24.50	[0.038, 0.066] 8.67	[0.038, 0.066] 0.30	[0.038, 0.066] 0.09	[0.038, 0.066] 0.03
#4	[0.028, 0.067] 62.52	[0.028, 0.067] 21.56	[0.028, 0.067] 0.34	[0.028, 0.067] 0.11	[0.028, 0.067] 0.02
#5	[0.032, 0.058] 19.47	[0.032, 0.058] 6.76	[0.032, 0.058] 0.27	[0.032, 0.058] 0.11	[0.032, 0.058] 0.01
#6	[0.029, 0.050] 23.32	[0.029, 0.050] 8.26	[0.029, 0.050] 0.30	[0.029, 0.050] 0.11	[0.029, 0.050] 0.01

Table 3.5. Computational gains of Algorithm 3.2 relative to grid search and bisection algorithms in the applications of Section 3.5.1. The top row for each specification is the confidence interval. The bottom row is time in seconds. For the bisection search, our tolerance is set to the absolute value of the LS estimate, divided by 1000. For comparability, we set the step-size of the grid search to the same value.

in linear regression models can be characterized in closed form and thus are straightforward to implement in practice. Algorithms 3.1 and 3.2 provide a clear explanation of how to apply ARTs in linear models, and the companion `Stata` and `R` packages available as part of the supplemental material are intended to facilitate doing so. Second, our discussion on the main requirements behind ARTs hopefully show that understanding the trade-offs between ARTs and other popular alternatives for inference with a small number of clusters, like the cluster wild bootstrap, is fundamental for practitioners to choose a method that aligns well with the features of their application. In particular, while ARTs essentially demand that the parameter of interest is suitably estimable cluster-by-cluster without imposing restrictions on the degree of heterogeneity across clusters, the cluster

		Grid Search		Bisection		ART
		Stud.	Unstud.	Stud.	Unstud.	
$q = 8$	#1	[0.124, 0.429] 2.62	[0.124, 0.429] 0.92	[0.124, 0.429] 0.11	[0.124, 0.429] 0.02	[0.124, 0.429] 0.06
	#2	[0.035, 0.391] 3.85	[0.035, 0.391] 1.40	[0.036, 0.391] 0.08	[0.036, 0.391] 0.03	[0.035, 0.391] 0.00
	#3	[0.035, 0.391] 4.09	[0.035, 0.391] 1.50	[0.035, 0.390] 0.08	[0.035, 0.390] 0.03	[0.035, 0.390] 0.00
	#4	[-0.070, 0.397] 9.29	[-0.070, 0.397] 3.37	[-0.070, 0.397] 0.11	[-0.070, 0.397] 0.03	[-0.070, 0.397] 0.00
	#5	[-0.067, 0.418] 9.20	[-0.067, 0.418] 3.18	[-0.067, 0.418] 0.07	[-0.067, 0.418] 0.04	[-0.067, 0.418] 0.01
$q = 10$	#1	[0.141, 0.603] 30.19	[0.141, 0.603] 10.41	[0.141, 0.603] 0.33	[0.141, 0.603] 0.11	[0.141, 0.603] 0.01
	#2	[0.068, 0.446] 26.61	[0.068, 0.446] 9.34	[0.068, 0.446] 0.33	[0.068, 0.446] 0.13	[0.069, 0.445] 0.00
	#3	[0.067, 0.458] 28.37	[0.067, 0.458] 9.78	[0.068, 0.458] 0.33	[0.068, 0.458] 0.11	[0.068, 0.458] 0.02
	#4	[-0.024, 0.431] 32.75	[-0.024, 0.431] 11.47	[-0.024, 0.430] 0.32	[-0.024, 0.430] 0.11	[-0.023, 0.430] 0.02
	#5	[-0.003, 0.452] 31.67	[-0.003, 0.452] 11.02	[-0.003, 0.452] 0.34	[-0.003, 0.452] 0.11	[-0.003, 0.451] 0.02
$q = 16$	#1	[0.124, 0.447] 373.86	[0.124, 0.447] 127.07	[0.124, 0.447] 3.19	[0.124, 0.447] 1.11	[0.124, 0.447] 0.13
	#2	[0.098, 0.364] 451.67	[0.098, 0.364] 153.20	[0.098, 0.364] 3.16	[0.098, 0.364] 1.11	[0.097, 0.364] 0.14
	#3	[0.088, 0.368] 415.67	[0.088, 0.368] 142.55	[0.088, 0.368] 3.25	[0.088, 0.368] 1.10	[0.087, 0.368] 0.16
	#4	[0.014, 0.325] 703.13	[0.014, 0.325] 248.68	[0.015, 0.325] 3.43	[0.015, 0.325] 1.14	[0.014, 0.325] 0.11
	#5	[0.048, 0.365] 572.54	[0.048, 0.365] 193.69	[0.048, 0.365] 3.09	[0.048, 0.365] 1.10	[0.047, 0.365] 0.14

Table 3.6. Computational gains of Algorithm 3.2 relative to grid search and bisection algorithms in the applications of Section 3.5.2. The top row for each specification is the confidence interval. The bottom row is time in seconds. For the bisection search, our tolerance is set to the absolute value of the LS estimate, divided by 1000. For comparability, we set the step-size of the grid search to the same value.

wild bootstrap requires the clusters to be sufficiently homogeneous (see Canay et al., 2021a) without demanding identification of the parameter of interest cluster-by-cluster.

CHAPTER 4

On the Performance of the Neyman Allocation with Small Pilots

4.1. Introduction

A growing literature on experiment design provides researchers with tools for reducing the asymptotic variance of their average treatment effect (ATE) estimates. Many do so in the context of two-wave experiments, where the researcher has access to a pilot study that can be used to improve the main study. Pilots are typically assumed to be large, allowing population parameters to be well-estimated. However, large pilots may not be realistic in practice. In this chapter, we study the implications of small pilots for experiment design through the lens of the Neyman Allocation.

The Neyman Allocation (Neyman 1934) is a simple method for minimizing the variance of the difference-in-means estimator of ATE. In a setting without covariates, suppose that the standard deviations of the treated and control outcomes are known. The Neyman Allocation assigns more units to either treatment or control in proportion to the ratio of their standard deviations. Intuitively, the optimal experiment entails more measurements of the noisier quantity. Since the variances are not known in practice, the feasible Neyman Allocation (FNA) estimates the variances using the pilot study and then plugs the estimates into the assignment rule.

The FNA is an important part of many experiment design procedures. In the econometrics literature alone there are several notable works. Hahn et al. (2011) propose to

estimate the variance of outcome and control groups conditional on covariate value, implementing the FNA conditional on covariates. In a similar vein, Tabord-Meehan (2021) employs tree-based techniques to stratify units based on their covariates. Units are then assigned to treatment and control based on the FNA conditional on strata. Meanwhile, Cytrynbaum (2021) proposes local randomization to select representative units for participation and treatment in experiments. In what Cytrynbaum (2021) terms the "fully efficient" case, treatment proportion conditional on the randomization group is chosen by the FNA. Despite their differences, the above papers study their proposals in asymptotic frameworks that take the pilot size to infinity. Their analyses, appropriate for large pilots, essentially assume that population parameters are arbitrarily well-estimated from the pilots alone. In practice, pilots are often conducted for logistical reasons and may be small. In such settings, accurately estimating the relevant variances may be difficult.

To understand the implications of small pilots, we study the properties of the Neyman Allocation in a novel asymptotic framework for two-wave experiments. Our framework takes the main wave sample size to infinity while the pilot size remains fixed. We show that when uncertainty in parameter estimation is non-negligible in the limit, the FNA converges in distribution to a mixture of normal distributions. Furthermore, we find that the FNA can do worse than the naive, balanced allocation that assigns half the units to treatment and half to control. This occurs when outcomes have similar variances across treatment and control, i.e. when outcomes are relatively homoskedastic with respect to treatment status. To assess how much homoskedasticity exists in practice, we examine the first 10 completed experiments in the AER RCT Registry. We ask the hypothetical question: if researchers conducted these studies as a two-wave experiment with a random sample from

the same population, would they do better with the FNA or with balanced randomization? We find that the treatment and control groups are often highly homoskedastic across a range of outcomes and across experiments. This suggests that if faced with a small pilot, the authors of those studies would likely not have benefited from implementing the FNA. Finally, we show that as pilot sizes increase, the amount of heteroskedasticity needed for the FNA to be preferable to the balanced allocation decreases, but at a rate that depends on the kurtosis of the outcome variables. Hence, even when researchers believe they are in a setting with high heteroskedasticity, they may want to avoid the FNA if they also believe that the outcomes are fat-tailed.

Our findings suggest that researchers should be cautious when designing experiments using the FNA with small pilots. However, even when pilots are large, methods which condition on many covariates may end up estimating the FNA using a small conditional sample. Furthermore, if researchers believe that units exhibit cluster-dependence – a common assumption in empirical work – the number of "effective" observations may be smaller still, impeding the estimation of the FNA. We note that this chapter specifically addresses the use of the FNA in reducing the asymptotic variance of the difference-in-means estimator in two-wave experiments. It does not speak to papers which take the treatment assignment probability as given, such as Bai (2022).

This chapter is most similar to papers pointing out a similar issue in the design of sequential experiments. Melfi and Page (1998) argue by simulation that treatment assignment rules based on estimated outcomes can do worse than non-adaptive rules due to estimation noise. Theoretical analysis is provided in Hu and Rosenberger (2003), in an

asymptotic framework that does not nest ours. This chapter is also related to those studying small sample problems in experiments. de Chaisemartin and Ramirez-Cuellar (2019) is concerned with the problems small strata pose for inference. Bruhn and McKenzie (2009) considers the effectiveness of various randomization strategies in achieving balance when a single-wave experiment is small. Finally, we note that there is a large literature discussing the large-pilot properties of the Neyman Allocation for alternative criteria such as power (e.g. Brittain and Schlesselman 1982, Azriel et al. 2012), minimax optimality (e.g. Bai 2021) or ethical considerations (e.g. Chapter 8 of Hu and Rosenberger 2006). These criteria fall outside the scope of this present chapter.

The remainder of this chapter is organized as follows. Section 4.2 presents the theoretical framework. Section 4.3 contains analytical and simulation results using a stylized toy example. Our main theoretical results can be found in section 4.4. We assess the level of homoskedasticity in selected empirical applications in section 4.5. Section 4.6 concludes the chapter. All proofs as well as additional empirical examples are contained in the online appendix.

4.2. Framework

We use a standard binary treatment potential outcomes framework assuming an infinite superpopulation. The potential outcomes are $(Y(0), Y(1))$, where $Y(0)$ denotes the potential outcome under control or status quo and $Y(1)$ denotes the potential outcome under treatment or the innovation.

Assumption 4.1. Potential outcomes $(Y(0), Y(1))$ have finite second moments. The vector of means is μ and the covariance matrix is Σ where

$$\mu = \mathbb{E}\left[\begin{pmatrix} Y(0) \\ Y(1) \end{pmatrix}\right] = \begin{pmatrix} \mu(0) \\ \mu(1) \end{pmatrix}$$
$$\Sigma = \text{Var}\left[\begin{pmatrix} Y(0) \\ Y(1) \end{pmatrix}\right] = \begin{pmatrix} \sigma^2(0) & \rho \cdot \sigma(0) \cdot \sigma(1) \\ \rho \cdot \sigma(0) \cdot \sigma(1) & \sigma^2(1) \end{pmatrix}.$$

Additionally, assume potential outcome variances are positive so that $\sigma^2(a) > 0$ for each $a \in \{0, 1\}$.

The estimand of interest is the Average Treatment Effect (ATE), $\theta = \mathbb{E}[Y(1) - Y(0)]$. To estimate the ATE, the experimenter conducts a two-wave experiment. The smaller first wave, also known as the pilot, and used to inform the experimenter about aspects of the design of the larger main wave (i.e. the second wave). The following assumptions about the two experimental waves will be maintained throughout the chapter. For notational clarity, the tilde symbol (e.g. $\widetilde{X}$) refers to quantities associated with the pilot.

Assumption 4.2. Potential outcomes in the pilot, denoted $\left\{\widetilde{Y}_i(0), \widetilde{Y}_i(1)\right\}_{i=1}^m$, consist of m i.i.d. draws from the distribution of the random vector $(Y(0), Y(1))'$. Treatment is randomly assigned in the pilot so that denoting assignments by $\left\{\widetilde{A}_i\right\}_{i=1}^m$,

$$\left\{\widetilde{Y}_i(0), \widetilde{Y}_i(1)\right\}_{i=1}^m \perp\!\!\!\perp \left\{\widetilde{A}_i\right\}_{i=1}^m.$$

Assumption 4.3. Potential outcomes in the main wave, denoted by $\{Y_i(0), Y_i(1)\}_{i=1}^n$, are n i.i.d. draws from the distribution of the random vector $(Y(0), Y(1))'$ and are independent to potential outcomes and treatment assignments in the pilot. That is,

$$\{Y_i(0), Y_i(1)\}_{i=1}^n \perp\!\!\!\perp \left\{\widetilde{Y}_i(0), \widetilde{Y}_i(1), \widetilde{A}_i\right\}_{i=1}^m.$$

The experimenter has access to an exogenous randomization device for assigning treatments in the main wave. That is, there is an i.i.d. sample of Uniform$[0, 1]$ random variables, $\{U_i\}_{i=1}^n$, available to the researcher and satisfying

$$\{U_i\}_{i=1}^n \perp\!\!\!\perp \left[\{Y_i(0), Y_i(1)\}_{i=1}^n, \left\{\widetilde{Y}_i(0), \widetilde{Y}_i(1), \widetilde{A}_i\right\}_{i=1}^m\right].$$

We describe a number of allocation schemes that the experimenter could implement in the main wave using the randomization devices $\{U_i\}_{i=1}^n$.

- **Simple random assignment:** For a given $p \in (0, 1)$, let

$$A_{p,i} = \mathbb{I}\{U_i \leq p\}.$$

The associated observed outcomes in this case are denoted $Y_{p,i} = Y_i(0)(1 - A_{p,i}) + Y_i(1)A_{p,i}$ and the estimator for the average treatment effect is the difference-in-means estimator:

$$\widehat{\theta}_p = \frac{\frac{1}{n}\sum_{i=1}^n Y_{p,i}A_{p,i}}{\frac{1}{n}\sum_{i=1}^n A_{p,i}} - \frac{\frac{1}{n}\sum_{i=1}^n Y_{p,i}(1 - A_{p,i})}{\frac{1}{n}\sum_{i=1}^n (1 - A_{p,i})}.$$

Balanced randomization corresponds to choosing $p = \frac{1}{2}$. We will refer to the associated treatment assignment rule as the balanced allocation.

- **The Infeasible Neyman Allocation:** For any given $p \in (0,1)$, elementary arguments show that

$$\sqrt{n}\left(\widehat{\theta}_p - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(1)}{p} + \frac{\sigma^2(0)}{1-p}\right).$$

The optimal choice of p to minimize the variance of the limiting distribution is the Neyman Allocation:

$$p_* = \frac{\sigma(1)}{\sigma(1) + \sigma(0)} \ .$$

The optimal treatment scheme, associated observed outcomes and difference-in-means estimator are denoted by

$$\begin{aligned} A_{p_*,i} &= \mathbb{I}\left\{U_i \leq p_*\right\} \\ Y_{p_*,i} &= Y_i(0)\left(1 - A_{p_*,i}\right) + Y_i(1)A_{p_*,i} \\ \widehat{\theta}_{p_*} &= \frac{\frac{1}{n}\sum_{i=1}^n Y_{p_*,i}A_{p_*,i}}{\frac{1}{n}\sum_{i=1}^n A_{p_*,i}} - \frac{\frac{1}{n}\sum_{i=1}^n Y_{p_*,i}\left(1 - A_{p_*,i}\right)}{\frac{1}{n}\sum_{i=1}^n \left(1 - A_{p_*,i}\right)} \ . \end{aligned} \tag{4.1}$$

Implementing the Neyman Allocation requires knowledge of the quantities $\sigma(0), \sigma(1)$ and as such is infeasible.

- **The Feasible Neyman Allocation (FNA):** One feasible implementation is to use the pilot data to form a plug-in estimator for p_*. We start by using pilot data to estimate potential outcome variances:

$$\widetilde{\sigma}^2_m(a) = \ \frac{1}{m_a - 1}\sum_{i=1}^m \left(\widetilde{Y}_i \mathbb{I}\left\{\widetilde{A}_i = a\right\} - \frac{1}{m_a}\sum_{i=1}^m \widetilde{Y}_i \mathbb{I}\left\{\widetilde{A}_i = a\right\}\right)^2 \ ,$$

where $m_a = \sum_{i=1}^{m} \mathbb{I}\left\{\widetilde{A}_i = a\right\}$. The feasible Neyman Allocation is

$$\widetilde{p} = \begin{cases} \frac{\widetilde{\sigma}_m(1)}{\widetilde{\sigma}_m(1)+\widetilde{\sigma}_m(0)} & \text{if } \widetilde{\sigma}_m(1), \widetilde{\sigma}_m(0) > 0 \ , \\ \frac{1}{2} & \text{otherwise} \ . \end{cases} \tag{4.2}$$

The latter case in (4.2) (where at least one $\widetilde{\sigma}_m(a) = 0$) avoids division by zero and additionally avoids the case where the entire main wave sample gets assigned to a single treatment arm. If the potential outcomes are continuously distributed, this latter case happens with zero probability. The distribution of $\widetilde{p}$ depends on m, but we omit the sample size subscript for notational convenience. The associated treatment allocation scheme, observed outcomes and difference-in-means estimator are denoted by

$$\begin{aligned} A_{\widetilde{p},i} &= \mathbb{I}\{U_i \leq \widetilde{p}\} \\ Y_{\widetilde{p},i} &= Y_i(0)\left(1 - A_{\widetilde{p},i}\right) + Y_i(1)A_{\widetilde{p},i} \\ \widehat{\theta}_{\widetilde{p}} &= \frac{\frac{1}{n}\sum_{i=1}^{n} Y_{\widetilde{p},i}A_{\widetilde{p},i}}{\frac{1}{n}\sum_{i=1}^{n} A_{\widetilde{p},i}} - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_{\widetilde{p},i}\left(1 - A_{\widetilde{p},i}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(1 - A_{\widetilde{p},i}\right)} \ . \end{aligned} \tag{4.3}$$

4.3. Toy Example

In this section, we illustrate the main problems with the FNA in small samples with a toy model. In the context of this simple example, we ask the question: when does the FNA do worse than the balanced allocation? It turns out that this happens for a range of plausible values of population parameters. Section 4.4 describes the extension of our findings into more general settings.

We assume in this section that the potential outcomes are bivariate normal:

$$\begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) \ .$$

Suppose we have a pilot of size m where m is even. Suppose treatment is assigned deterministically as follows:

$$\widetilde{A}_i = \begin{cases} 1 & \text{if } 1 \leq i \leq \frac{m}{2} \\ 0 & \text{otherwise.} \end{cases}$$

As such, treatment assignment is strongly balanced in the pilot. Using standard arguments, it follows that the variance estimates are distributed as independent χ^2 random variables:

$$\widetilde{\sigma}^2_m(1) \perp\!\!\!\perp \widetilde{\sigma}^2_m(0) \quad , \quad \left(\frac{m}{2} - 1\right)\left(\frac{\widetilde{\sigma}^2_m(a)}{\sigma^2(a)}\right) \sim \chi^2_{\frac{m}{2}-1} \ .$$

Conditional on the pilot sample, the feasible Neyman Allocation assigns

$$\widetilde{p} = \frac{\widetilde{\sigma}_m(1)}{\widetilde{\sigma}_m(0) + \widetilde{\sigma}_m(1)}$$

proportion of the units in the main to treatment. Suppose for simplicity again that we assign the first $n\widetilde{p}$ units to treatment and the remainder to control (rounding $n\widetilde{p}$ if necessary). Then,

$$\widehat{\theta}_{\widetilde{p}} = \left(\frac{1}{n\widetilde{p}} \sum_{i=1}^{n\widetilde{p}} Y_i(1)\right) - \left(\frac{1}{n - n\widetilde{p}} \sum_{i=n\widetilde{p}+1}^{n} Y_i(0)\right) \ .$$

Terms in the above expression have conditional distributions:

$$\left[\frac{1}{\sqrt{n\widetilde{p}}}\sum_{i=1}^{n\widetilde{p}} Y_i(1) \;\middle|\; \widetilde{p}\right] \sim \mathcal{N}\left(0, \sigma^2(1)\right) \quad , \quad \left[\frac{1}{\sqrt{n-n\widetilde{p}}}\sum_{i=n\widetilde{p}+1}^{n} Y_i(0) \;\middle|\; \widetilde{p}\right] \sim \mathcal{N}\left(0, \sigma^2(0)\right) \ .$$

Furthermore, they are independent. Hence,

$$\left[\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) \;\middle|\; \widetilde{p}\right] \sim \mathcal{N}\left(0\,,\ \frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right) \ .$$

Taking expectation over $\widetilde{p}$, we have that:

$$\text{Var}\left[\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right)\right] = \mathbb{E}\left[\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right] \ . \tag{4.4}$$

Letting $\widetilde{p}$ be constant at p, we obtain the variance of the difference-in-means estimator under simple random assignment as a special case:

$$\text{Var}\left[\sqrt{n}\left(\widehat{\theta}_{p} - \theta\right)\right] = \frac{\sigma^2(1)}{p} + \frac{\sigma^2(0)}{1-p} \ .$$

Our goal is then to compare the two expressions above when we set $p = \frac{1}{2}$. First, note that we can rewrite (4.4) as

$$\mathbb{E}\left[\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right] = \mathbb{E}\left[\left(1 + \frac{1}{Z_m}\frac{\sigma(0)}{\sigma(1)}\right)\sigma^2(1) + \left(1 + Z_m\frac{\sigma(1)}{\sigma(0)}\right)\sigma^2(0)\right] \ ,$$

where $Z_m \sim \sqrt{F\left(\frac{m}{2}-1, \frac{m}{2}-1\right)}$. The Neyman Allocation does worse than balanced randomization whenever the following obtains:

$$\begin{aligned}
&\mathbb{E}\left[\left(1+\frac{1}{Z_m}\frac{\sigma(0)}{\sigma(1)}\right)\sigma^2(1)+\left(1+Z_m\frac{\sigma(1)}{\sigma(0)}\right)\sigma^2(0)\right] \geq 2\sigma^2(1)+2\sigma^2(0) \\
\iff &\mathbb{E}\left[\frac{1}{Z_m}+Z_m\right]\sigma(1)\sigma(0) \geq \sigma^2(1)+\sigma^2(0) \\
\iff &\frac{\sigma^2(1)}{\sigma^2(0)}-\mathbb{E}\left[\frac{1}{Z_m}+Z_m\right]\frac{\sigma(1)}{\sigma(0)}+1 \leq 0 \\
\iff &\frac{\sigma^2(1)}{\sigma^2(0)}-2\mathbb{E}\left[Z_m\right]\frac{\sigma(1)}{\sigma(0)}+1 \leq 0 \ .
\end{aligned}$$

The final implication uses the fact that Z_m is reciprocally symmetric under bivariate normality and balanced randomization. By the quadratic formula, the above inequality satisfied if and only if

$$\frac{\sigma(1)}{\sigma(0)} \in C_m := \left[\ \mathbb{E}\left[Z_m\right]-\sqrt{\mathbb{E}\left[Z_m\right]^2-1}\ ,\ \mathbb{E}\left[Z_m\right]+\sqrt{\mathbb{E}\left[Z_m\right]^2-1}\ \right] \ . \tag{4.5}$$

Note that reciprocal symmetry together with Jensen's inequality guarantees that the discriminant is strictly positive:

$$\mathbb{E}[Z_m]^2 = \mathbb{E}[Z_m]\mathbb{E}\left[\frac{1}{Z_m}\right] > 1 \ .$$

Furthermore, we have that

$$\left(\mathbb{E}\left[Z_m\right]+\sqrt{\mathbb{E}\left[Z_m\right]^2-1}\right)\left(\mathbb{E}\left[Z_m\right]-\sqrt{\mathbb{E}\left[Z_m\right]^2-1}\right) = 1 \ .$$

In other words, the interval has the form

$$C_m = \left[\frac{1}{c_m}, c_m\right] \ ,$$

where $c_m > 1$ is the upper bound in (4.5). Hence, there is a range of parameter values under which the FNA does strictly worse than balanced randomization. We first note that $1 \in C_m$ for all m. This is intuitive since $p = \frac{1}{2}$ is the infeasible Neyman Allocation when $\sigma(1)/\sigma(0) = 1$. Secondly, $x \in C_m \Leftrightarrow 1/x \in C_m$. That is, the relative performance of the FNA to the balanced allocation does not change when we relabel treatment and control.

Simulation Evidence

Given an underlying distribution, it is simple to compute C_m by Monte Carlo integration. In this subsection, we present the values of C_m for some simple models and argue that for plausible values of $\sigma(1)/\sigma(0)$, the FNA performs worse than the balanced allocation.

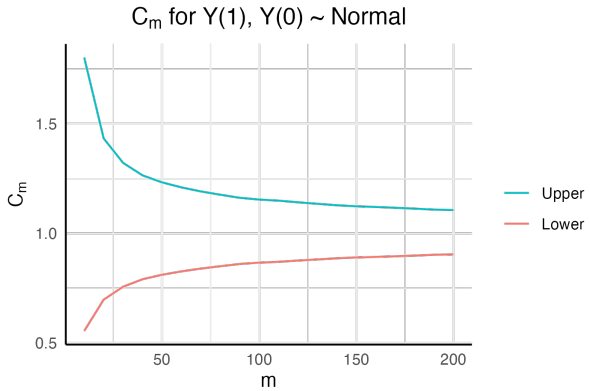

Figure 4.1. C_m when $Y(1) \sim \mathcal{N}(\mu(1), \sigma^2(1))$ and $Y(0) \sim \mathcal{N}(\mu(0), \sigma^2(0))$. Monte Carlo integration using 10,000 draws.

We start with the toy model, where $Y(1) \sim \mathcal{N}(\mu(1), \sigma^2(1))$ and $Y(0) \sim \mathcal{N}(\mu(0), \sigma^2(0))$. Results are shown in Figure 4.1. The set of parameter values over which the FNA does worse is larger when m is smaller. In fact, we show in Section 4.4 that if the data generating process (DGP) is sub-Gaussian, the length of C_m is $O\left(m^{-1/2}\right)$. In the toy model, $C_{20} = [0.70, 1.43]$, while C_{50} is $[0.81, 1.23]$. Suppose instead that $Y(1) \sim \mu(1) + \sigma(1)\chi_1^2$, $Y(0) \sim \mu(0) + \sigma(0)\chi_1^2$. Figure 4.2 shows that C_m is wider across the range of m. In particular, $C_{20} = [0.49, 2.22]$, while $C_{50} = [0.64, 1.61]$. While the intervals may appear rather narrow at first glance, we provide numerous examples in Section 4.5 in which the amount of heteroskedasticity falls within this range.

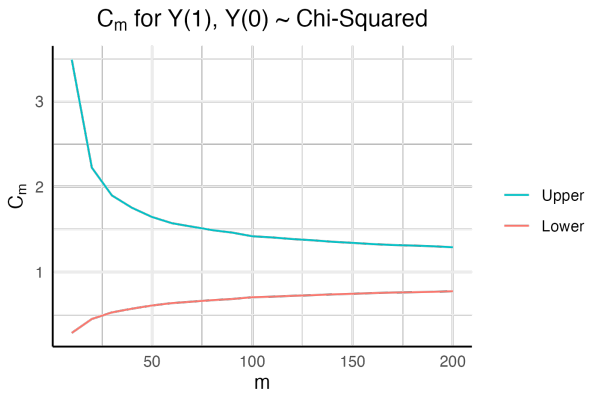

Figure 4.2. C_m when $Y(1) \sim \mu(1) + \sigma(1)\chi_1^2$ and $Y(0) \sim \mu(0) + \sigma(0)\chi_1^2$. Monte Carlo integration using 10,000 draws.

As our toy model shows, it is the estimation of $\widetilde{\sigma}(1)$ and $\widetilde{\sigma}(0)$ that causes problems when m is small. It is therefore intuitive that C_m will be wide when the distributions are fat tailed, that is, when kurtosis is high. As such, we consider the following parametrization: $Y(1) \sim \mu(1) + \sigma(1) \cdot \text{Pareto}(l, s)$ and $Y(0) \sim \mu(0) + \sigma(0) \cdot \text{Pareto}(l, s)$. Here, l and s are the location and scale parameters respectively. Figure 4.3 plots C_m for $l = 1$ and $s \in \{2, 3, 4\}$. Indeed, we see that the bands are much wider than in the previous models.

For $s = 2$, even when $m = 100$, $C_m = [0.40, 2.51]$, which, given the examples in Section 4.5 appear to be fairly extreme amounts of heteroskedasticity. Finally, we note that, C_m decreases in width as we move from $s = 2$ to $s = 4$.

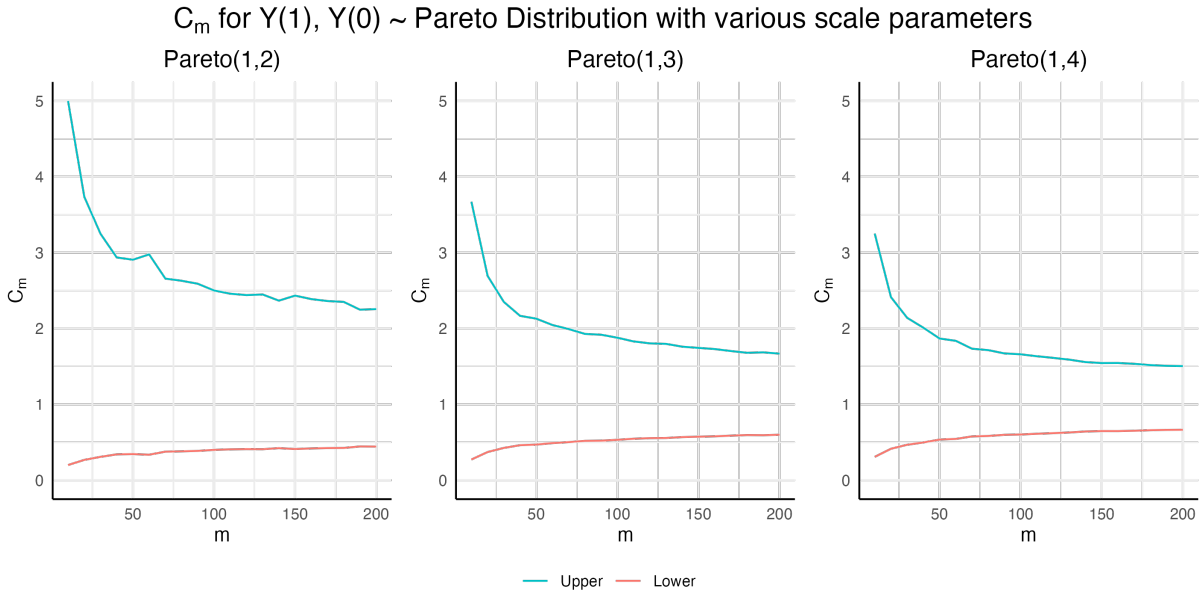

Figure 4.3. C_m when $Y(1) \sim \mu(1) + \sigma(1) \cdot \text{Pareto}(1, s)$, $Y(0) \sim \mu(0) + \sigma(0) \cdot \text{Pareto}(1, s)$ and $s \in \{2, 3, 4\}$. Monte Carlo integration using 10,000 draws.

In sum, we see that the FNA does worse than the balanced allocation when the treatment and control groups are relatively homoskedastic. Furthermore, C_m can be quite large when m is small. The small pilot problem is exacerbated when observations exhibit cluster dependence, so that the "effective observations" are fewer in number. Small pilot issues can also arise when researchers perform stratified randomization with many strata, so that each stratum ends up with few observations. In Section 4.5, we argue that many empirical applications in fact have fairly homoskedastic outcomes. As such, unless a researcher has reason to believe that their outcomes are highly heteroskedastic, they should exercise caution in using the FNA with small pilots.

4.4. Theoretical Results

In this section, we study the theoretical properties of the FNA. We first review the case where $n, m \to \infty$. Next, we study the FNA in our asymptotic framework where $n \to \infty$ but m is fixed, highlighting implications of our results for empirical researchers.

Large-m asymptotics. Consider an asymptotic regime where both $m, n \to \infty$, corresponding to situations where both the pilot and the main wave are large. It is straightforward to show that $\widetilde{p} \xrightarrow{p} p_*$. Furthermore, we have that:

Proposition 4.1. Under Assumptions 4.1, 4.2 and 4.3, as $m, n \to \infty$

$$\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_*\right),$$

where

$$\Sigma_* = \frac{\sigma^2(1)}{p_*} + \frac{\sigma^2(0)}{1 - p_*} = (\sigma(1) + \sigma(0))^2.$$

In other words, $\widehat{\theta}_{\widetilde{p}}$ and $\widehat{\theta}_{p_*}$ have the same limiting distribution after suitable centering and scaling. This occurs because in the limit, noise coming from estimating p_* with $\widetilde{p}$ is negligible in comparison to the sampling error of the difference-in-means estimator. Researchers employing the large m framework essentially assume that $\widetilde{p}$ is an arbitrarily good estimator for p_* so that its sampling error can be ignored. In practice, error in $\widetilde{p}$ can be large, particularly when m is small. Asymptotic approximations that do not take this into account will likely perform poorly in finite sample.

Fixed-m Asymptotics. To better understand the behavior of the FNA under small pilots, we study its properties in a novel asymptotic framework that takes m to be fixed, even as $n \to \infty$.

A growing literature in econometrics uses fixed-sample asymptotics to study settings in which the "effective sample size" is small. For example, when data exhibit cluster-dependence, settings with few clusters pose unique challenges for estimation and inference. To tailor their analyses to these problems, papers such as Ibragimov and Müller (2010), Canay et al. (2017c) and Canay et al. (2021b) employ asymptotic frameworks in which the number of clusters is fixed in the limit. Similarly, to model difference-in-differences studies involving few treated units, Conley and Taber (2011) keep the number of treated units fixed even as the number of untreated units tend to infinity. In the same vein, inference for regression discontinuity designs typically involves few observations around the discontinuity. To capture this, Canay and Kamat (2017) analyze a permutation test under an asymptotic regime with a fixed number of observations on either side of the discontinuity.

Our approach is similar in spirit to these papers. In keeping m fixed, we have that $\widetilde{p}$ is a noisy estimate of p_* even in the limit. Preserving this important feature of the statistical problem makes our framework more appropriate for analyzing experiments with small pilots. In this setting, $\widehat{\theta}_{\widetilde{p}}$ converges in distribution to a mixture of Gaussians instead of $\mathcal{N}(0, \Sigma_*)$. Furthermore, the form of the limiting mixture distribution depends on the distribution of $\widetilde{p}$. This is the content of the following proposition:

Proposition 4.2. Under Assumptions 4.1, 4.2 and 4.3, if m remains fixed as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) \xrightarrow{d} \mathcal{L}_m \ ,$$

where $\mathcal{L}_m$ is a random variable whose distribution takes the form

$$\mathbb{P}\left(\mathcal{L}_m \leq t\right) = \int_0^1 \Phi\left(\frac{t}{s\left(p\right)}\right) \, G_m(\mathrm{d}p),$$

Φ is the CDF of $\mathcal{N}(0,1)$, G_m is the distribution of $\widetilde{p}$ and $s(\cdot)$ is defined by

$$s(p) = \sqrt{\frac{\sigma^2(1)}{p} + \frac{\sigma^2(0)}{1-p}} \, .$$

Remark 4.1. The weak limit $\mathcal{L}_m$ in Proposition 4.2 has mean zero. To see this, note that conditional on a value of the assignment probability p, i.e. conditional on the event $\widetilde{p} = p$, $\mathcal{L}_m$ has a $\mathcal{N}(0, s(p))$ distribution (where $s(p)$ is as defined in Proposition 4.2). The conclusion then holds by the Law of Iterated Expectations.

To see the intuition for our result, recall that for each $p \in (0,1)$,

$$\sqrt{n}\left(\widehat{\theta}_p - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(1)}{p} + \frac{\sigma^2(0)}{1-p}\right) = \mathcal{N}\left(0, s^2(p)\right).$$

When m is held fixed, in the limit as $n \to \infty$, $\widetilde{p}$ remains a non-degenerate random variable. In particular, the limiting distribution of $\widetilde{p}$ is its finite-sample distribution. Thus, the distribution of $\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right)$ becomes a mixture of the marginal distributions of the process

$$\left\{\sqrt{n}\left(\widehat{\theta}_p - \theta\right) : p \in (0,1)\right\} \, ,$$

where the mixing distribution is the distribution of $\widetilde{p}$, denoted here by G_m.

Implication for Experiments. Our results have implications for the use of the FNA when pilot sizes are small. In this sub-section we compare the asymptotic variances of $\widehat{\theta}_{\widetilde{p}}$, $\widehat{\theta}_{p_*}$ as well as $\widehat{\theta}_p$ with $p = \frac{1}{2}$. Since these estimators all have limit distributions that have

mean zero (after centering around θ and scaling by $\sqrt{n}$), comparing asymptotic mean squared errors is equivalent to comparing the variances of the limit distributions. We first show that $\widehat{\theta}_{\widetilde{p}}$ has larger asymptotic variance in the fixed-m regime than in the large-m regime. We next show that under reasonable ranges of parameter values, the asymptotic variance of $\widehat{\theta}_{\widetilde{p}}$ can exceed that of $\widehat{\theta}_p$ with $p = \frac{1}{2}$.

Remark 4.2. By Proposition 4.2 and Remark 4.1, $\widehat{\theta}_{\widetilde{p}}$ has a limit distribution which has mean zero. Given that $\widehat{\theta}_{p_*}$ and $\widehat{\theta}_p$ also have mean zero normal limit distributions, comparing asymptotic mean squared errors is equivalent to comparing the variances of the limit distributions.

We begin with the corollary:

Corollary 4.1. Under Assumptions 4.1, 4.2 and 4.3, suppose m remains fixed as $n \to \infty$. $\mathcal{L}_m$ has variance:

$$\mathbb{E}\left[\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right] > \Sigma_* \ .$$

In words, the asymptotic variance of $\widehat{\theta}_{\widetilde{p}}$ is larger under the fixed-m regime than under the large-m regime. When pilots are small, uncertainty in $\widetilde{p}$ may be large and could affect the asymptotic variance of $\widehat{\theta}_{\widetilde{p}}$. In particular, $\widehat{\theta}_{\widetilde{p}}$ will not be able to attain the optimal asymptotic variance of the infeasible allocation p_*. Conventional large-m asymptotics may be too optimistic about the effectiveness of the Neyman Allocation with small pilots. We expect our analysis to better capture the behavior of $\widehat{\theta}_{\widetilde{p}}$ when pilots are small.

In addition to not attaining Σ_*, the $\widehat{\theta}_{\widetilde{p}}$ can do worse than $\widehat{\theta}_p$ for certain values of $\sigma^2(1)$ and $\sigma^2(0)$, as our next two results asserts. For convenience, define the following:

Definition 4.1. Let C_m be the set such that for a pilot of size m, $\widehat{\theta}_{\widetilde{p}}$ has higher asymptotic variance than $\widehat{\theta}_p$ if and only if $\frac{\sigma(1)}{\sigma(0)} \in C_m$. Let $|C_m|$ denote the length of C_m.

Definition 4.2. Let

$$Z_m = \frac{\widetilde{\sigma}_m(1)}{\sigma(1)} \Big/ \frac{\widetilde{\sigma}_m(0)}{\sigma(0)} \ .$$

and

$$B_m = \frac{1}{2}\,\mathbb{E}\left[\frac{1}{Z_m} + Z_m\right] \ .$$

For our first result, we characterize the region C_m in terms of B_m.

Proposition 4.3. Under Assumptions 4.1, 4.2 and 4.3, suppose m remains fixed as $n \to \infty$. Then

$$C_m = \left[\frac{1}{c_m}\,,\, c_m\right] \ . \tag{4.6}$$

Furthermore,

$$c_m = B_m + \sqrt{B_m^2 - 1} > 1 \quad \text{and} \quad |C_m| = 2\sqrt{B_m^2 - 1} \ > \ 0 \ .$$

The properties of C_m are intuitive. Firstly, $x \in C_m$ implies that $1/x \in C_m$. That is, the relative performance of the FNA to the balanced allocation does not change when we relabel treatment and control. Secondly, $1 \in C_m$. This is because when $\sigma(1)/\sigma(0) = 1$, the balanced allocation is optimal. Finally, note that $|C_m|$ depends on the bias of $\widetilde{\sigma}_m(1)/\widetilde{\sigma}_m(0)$ and $\widetilde{\sigma}_m(0)/\widetilde{\sigma}_m(1)$. In particular, if both terms are unbiased, $B_m = 1$ and $|C_m| = 0$. However, $|C_m|$ is strictly positive as long as $\widetilde{p}$ is not degenerate.

The exact properties of C_m depends on the underlying distributions of potential outcomes. To understand its behavior in a more general setting, we study its first-order approximation. This yields the following:

Proposition 4.4. Under Assumptions 4.1, 4.2 and 4.3, suppose $n \to \infty$. Suppose additionally that $Y_i(1)$ and $Y_i(0)$ are sub-Gaussian. Then,

$$C_m = \left[1 - \sqrt{\frac{V}{m}} + \delta_m^-\,,\, 1 + \sqrt{\frac{V}{m}} + \delta_m^+\right] \,,$$

where $\delta_m^+, \delta_m^- = o\left(\frac{1}{\sqrt{m}}\right)$ and

$$V = \frac{1}{4}\left(\frac{\mathbb{E}\left[(Y(1) - \mu(1))^4\right]}{\sigma^4(1)} + \frac{\mathbb{E}\left[(Y(0) - \mu(0))^4\right]}{\sigma^4(0)} - 2\right) \,.$$

Provided that the potential outcomes are sub-Gaussian, the relative efficiency of $\widehat{\theta}_{\widetilde{p}}$ and $\widehat{\theta}_p$ under $p = 1/2$ is, to a first order, determined by the kurtosis of $Y_i(1)$ and $Y_i(0)$. Intuitively, if the potential outcomes have fatter tails, $\widetilde{p}$ will be poorly estimated, leading to larger variance in $\widehat{\theta}_{\widetilde{p}}$.

Furthermore, $|C_m|$ is shrinking to 0 at the rate $1/\sqrt{m}$. Letting $m \to \infty$, $\widehat{\theta}_{\widetilde{p}}$ has weakly lower asymptotic variance across all parameter values. Hence, we recover the classic result concerning the optimality of the Neyman Allocation. When m is small, however, C_m can be wide, as Proposition 4.3 suggests. As we will argue in Section 4.5, many empirical applications have $\sigma(1)/\sigma(0)$ close to 1, so that for small m, they fall within the range in which balanced randomization is preferred.

While the sub-Gaussian assumption limits the applicability of Proposition 4.4, it covers binary and bounded outcomes, which are relevant in empirical work. Furthermore, we

consider it to be a negative result: even when potential outcomes are well-behaved, the FNA is sensitive to the kurtosis of the potential outcomes. It can still perform poorly relative to the balanced allocation as a result.

Remark 4.3. Using an argument based on Taylor expansions, we can weaken the sub-Gaussian assumption to finiteness of the first 14^{th} moments. The proof is available on request.

4.5. Empirical Evidence of Approximate Homoskedasticity

To assess the amount of heteroskedasticity that empirical researchers face, we revisit the first 10 completed experiments in the AER RCT Registry. In each experiment, we ask the following question: Suppose the authors had access to a small pilot prior to the main study, would they have done better using the FNA instead of balanced randomization? In each experiment presented, we use the full experimental sample to estimate the standard deviations of each treatment arm and compute the corresponding ratios to see if these are close to one. In practice, researchers cannot do this given a small pilot since they do not have access to consistent variance estimators. Our findings suggest that these authors would likely not have not better with the FNA. We present two examples in this section: Avvisati et al. (2014) is an experiment in which the outcomes are relatively homoskedastic. Ashraf et al. (2006) contains outcome variables which are heteroskedastic. In this example, we also provide estimates of the interval C_m and show that it will be wide even when m is large. The remaining eight experiments are qualitatively similar to Avvisati et al. (2014) and can be found in Section C.2 of the online appendix.

4.5.1. Avvisati et al. (2014)

There is a significant body of research examining the impact of school-level factors such as class size or teacher quality on educational performance of students. These are typically seen as the primary instruments for educational policy intervention. A large body of work also examines the impact of parental inputs on educational outcomes. Avvisati et al. (2014) study whether or not parental inputs can be effectively manipulated through simple participation programs at schools. They do so via a large-scale randomized control trial in middle schools in the Créteil educational district of Paris. The experiment targeted families of 6th graders and the program consisted of a sequence of three meetings with parents every 2–3 weeks. The sessions focused on how parents can help their children by participating at school and at home in their education and additionally, included advice on how to adapt to results in end-of-term report cards. Participation in the program was randomized at the class level – half of the classes at a given school were assigned to the participation program. Classes are groups of 20–30 students. The overall sample comprised of 183 classes and a total of 4,308 students. The study tracked three types of outcomes: (1) parental involvement attitudes and behaviour; (2) children's behaviour, namely truancy, disciplinary record and work effort; and (3) children's academic results. Since randomization was done at the class-level, we examine heteroskedasticity with respect to treatment status at both the individual level and at the class level.

Table 4.1 reports student-level standard deviations in treatment and control groups, as well as their ratios, for the main outcomes of interest. These are the outcomes considered in Tables 2, 3 and 5 in Avvisati et al. (2014). The ratios are all close to one, so that by and large the treatment and control groups are relatively homoskedastic at the student

level. This indicates that if the experimenters were to run a randomized control trial in the same population with treatment assigned at the individual-level, the FNA would likely yield no improvement relative to the balanced allocation.

Table 4.2 reports standard deviations and their ratios in class-level means of outcomes. This corresponds to a scenario in which classes are the units of interest, with class-level means as the relevant outcomes. We first calculate class-level means and then compute their standard deviations across classes for the treatment and control groups respectively. The standard deviation ratios for class-level means are by and large also close to one. Hence, if the hypothetical experiment was to be conducted at the class-level, the treatment and control groups would still be relatively homoskedastic. In this case, the FNA would again not improve upon the balanced allocation.

4.5.2. Ashraf et al. (2006)

A large body of economic models posit that individuals have time inconsistent preferences, exhibiting more impatience for near-term trade-offs than for future trade-offs. The implication of these models is that those who engage in commitment devices ex ante may improve their welfare. To test this hypothesis, Ashraf et al. (2006) conducted an RCT in the Philippines, in which individuals were offered randomly offered the chance to open a SEED (Save, Earn, Enjoy Deposits) account. Money deposited into the account cannot be withdrawn until the owner reached a goal, such as reaching a savings amount or until a pre-specified month in which they expected large expenditures.

Partnering with a rural bank in Mindanao, they authors surveyed 1,777 of their existing or former clients, of which 842 were placed into the treatment group, while 469 were

Table 4.1. Student-Level Heteroskedasticity in Avvisati et al. (2014).

	Outcome Variables	Treatment	Control	Treat./Cont.
Parental Involvement	Global parenting score	0.34	0.34	1.01
	School-based involvement score	0.66	0.63	1.05
	Home-based involvement score	0.59	0.57	1.04
	Understanding and perceptions score	0.53	0.55	0.97
	Parent-school interaction	0.40	0.40	1.01
	Parental monitoring of school work	0.43	0.41	1.05
Behavior	Absenteeism	6.29	8.63	0.73
	Pedagogical team: Behavioral score	0.73	0.74	0.98
	Pedagogical team: Discipl. sanctions	1.20	1.18	1.02
	Pedagogical team: Good conduct	0.49	0.47	1.04
	Pedagogical team: Honors	0.28	0.32	0.89
	Teacher assessment: Behavior in class	0.48	0.49	0.98
	Teacher assessment: School work	0.49	0.50	1.00
Test Scores	French (Class grade)	3.73	3.70	1.01
	Mathematics (Class grade)	4.26	4.25	1.00
	Average across subjects (Class grade)	2.87	2.88	1.00
	Progress over the school year	0.49	0.49	0.99
	French (Uniform test)	0.99	1.01	0.98
	Mathematics (Uniform test)	0.99	1.02	0.98

placed in the control group. As treatment involved receiving a briefing on the importance of savings, the remaining 466 individuals were placed in the marketing group, receiving the briefing but not access to SEED. We focus on the Table VI of Ashraf et al. (2006), containing results on saving behavior. The parameter of interest is the Intent-to-Treat effect, with approximately 25% of the treated taking up treatment. Here, the authors find that relative to the control group, the treatment group had a higher change in savings 6 months (6m) and 12 months (12m) after treatment. Comparing treatment to marketing group led to weaker but still positive results.

We present standard deviations of the outcomes as well as their ratios in Table 4.3. The outcomes of interest are

Table 4.2. Class-Level Heteroskedasticity in Avvisati et al. (2014).

	Outcome Variables	Treatment	Control	Treat./Cont.
Parental Involvement	Global parenting score	0.12	0.12	0.97
	School-based involvement score	0.24	0.35	0.69
	Home-based involvement score	0.20	0.15	1.33
	Understanding and perceptions score	0.20	0.22	0.94
	Parent-school interaction	0.13	0.12	1.09
	Parental monitoring of school work	0.13	0.14	0.96
Behavior	Absenteeism	2.21	3.45	0.64
	Pedagogical team: Behavioral score	0.24	0.27	0.88
	Pedagogical team: Discipl. sanctions	0.36	0.36	1.01
	Pedagogical team: Good conduct	0.22	0.23	0.95
	Pedagogical team: Honors	0.10	0.11	0.87
	Teacher assessment: Behavior in class	0.16	0.16	1.01
	Teacher assessment: School work	0.15	0.14	1.02
Test Scores	French (Class grade)	1.32	1.31	1.01
	Mathematics (Class grade)	1.76	1.84	0.96
	Average across subjects (Class grade)	0.83	0.93	0.89
	Progress over the school year	0.16	0.14	1.11
	French (Uniform test)	0.42	0.42	1.01
	Mathematics (Uniform test)	0.40	0.43	0.93

(1) Change in Total Balance (6m) (Δ *Tot. Bal. (6m)*),

(2) Change in Total Balance (12m) (Δ *Tot. Bal. (12m)*),

(3) Change in Total Balance exceeds 0% (12m) (Δ *Tot. Bal.* $> 0\%$ *(12m)*),

(4) Change in Total Balance exceeds 20% (12m) (Δ *Tot. Bal.* $> 20\%$ *(12m)*).

We first note that Δ *Tot. Bal.* $> 0\%$ *(12m)* and Δ *Tot. Bal.* $> 20\%$ *(12m)* are binary outcomes which are relatively homoskedastic. Δ *Tot. Bal. (6m)* and Δ *Tot. Bal. (12m)*, measured in Philippine pesos, exhibit more heteroskedasticity. In particular, comparing the treatment group to the marketing group at the 12 month period, we observe a standard deviation ratio of 3.13. At first glance, this suggests that the FNA might outperform balanced randomization, at least with respect to this specific outcome. This turns out

to be false once we investigate the heteroskedasticity in Δ *Tot. Bal. (6m)* and Δ *Tot. Bal. (12m)*. Quantiles of these variables are displayed in Table 4.4. Clearly, they have extremely fat right tails, which we confirm by computing the kurtosis, contained in Table 4.5. Fat tails worsen the performance of a variety of statistical techniques, including the FNA, as our analysis in Section 4.4 shows.

Table 4.3. Heteroskedasticity in Ashraf et al. (2006).

	Treat.	Cont.	Market.	Treat./Cont.	Treat./Market.
Δ Tot. Bal. (6m)	2347.60	2880.70	1335.98	1.76	0.81
Δ Tot. Bal. (12m)	6093.24	1945.00	2690.65	2.26	3.13
Δ Tot. Bal. $>$ 0% (12m)	0.47	0.45	0.42	1.11	1.05
Δ Tot. Bal. $>$ 20% (12m)	0.40	0.35	0.31	1.30	1.16

Table 4.4. Quantiles of Outcome Variables in Ashraf et al. (2006).

Variable	Group	1%	5%	10%	50%	90%	95%	99%	99.5%	99.9%
Δ Tot. Bal. (6m)	Treat.	-1100	-500	-300	0	500	1500	7200	13100	28900
	Market.	-1000	-600	-400	0	100	900	5600	12200	40800
	Cont.	-1600	-600	-500	0	0	600	2700	6100	18400
Δ Tot. Bal. (12m)	Treat.	-1300	-900	-500	0	500	1600	8500	18100	102300
	Market.	-1300	-900	-500	-100	300	1600	10500	15700	19900
	Cont.	-2000	-1200	-800	-100	100	900	6500	8100	34300

Table 4.5. Kurtosis of the Outcome Variables in Ashraf et al. (2006).

Variable	Treat.	Market.	Cont.
Δ Tot. Bal. (6m)	252.78	218.92	156.33
Δ Tot. Bal. (12m)	258.56	66.56	309.89

Researchers cannot estimate C_m if they only have access to a small pilot. However, using the full experiment data, we are able to estimate C_m for a range of m. Our results are displayed in Figure 4.4. Given the high kurtosis, there is a relatively large range of ratios

of standard deviations for which the balanced allocation is preferred to the FNA. We can further compute the m necessary before the FNA outperforms the balanced allocation. Our results are collected in Table 4.6. The first two rows, labelled "Exact", refer to intervals which are computed using the full experiment. Here, we see that the necessary pilot sizes are between 25–50% of the full experiment. For the comparison of Treatment and Marketing group at 6 months, the necessary pilot size exceeds $2,000$, falling outside the set of grid points we explored. To complete the analysis, we use our asymptotic result to obtain approximations of the necessary m. Comparing the asymptotic intervals to the exact ones, we see that the former is far too optimistic for the fat-tailed DGP in Ashraf et al. (2006). This is likely because the sub-Gaussian assumption is inappropriate for this data. Nonetheless, applying the asymptotic bounds to the case of Treatment vs Marketing group at 6 months, we find that a pilot size of 7,000 would be necessary for the FNA to outperform balanced randomization. This is nearly 4 times the size of the actual experiment.

Table 4.6. Necessary Pilot Sizes.

	Treat./Cont.	Treat./Market.
Exact 6m	930.70	–
Exact 12m	1014.94	475.29
Asympt. 6m	355.02	6857.59
Asympt. 12m	177.10	35.52

In sum, our analysis suggests that the FNA would have performed poorly in the context of Ashraf et al. (2006). Even though the outcomes of interest exhibit stronger heteroskedasticity, the fat tails of the outcome distributions also impedes the estimation

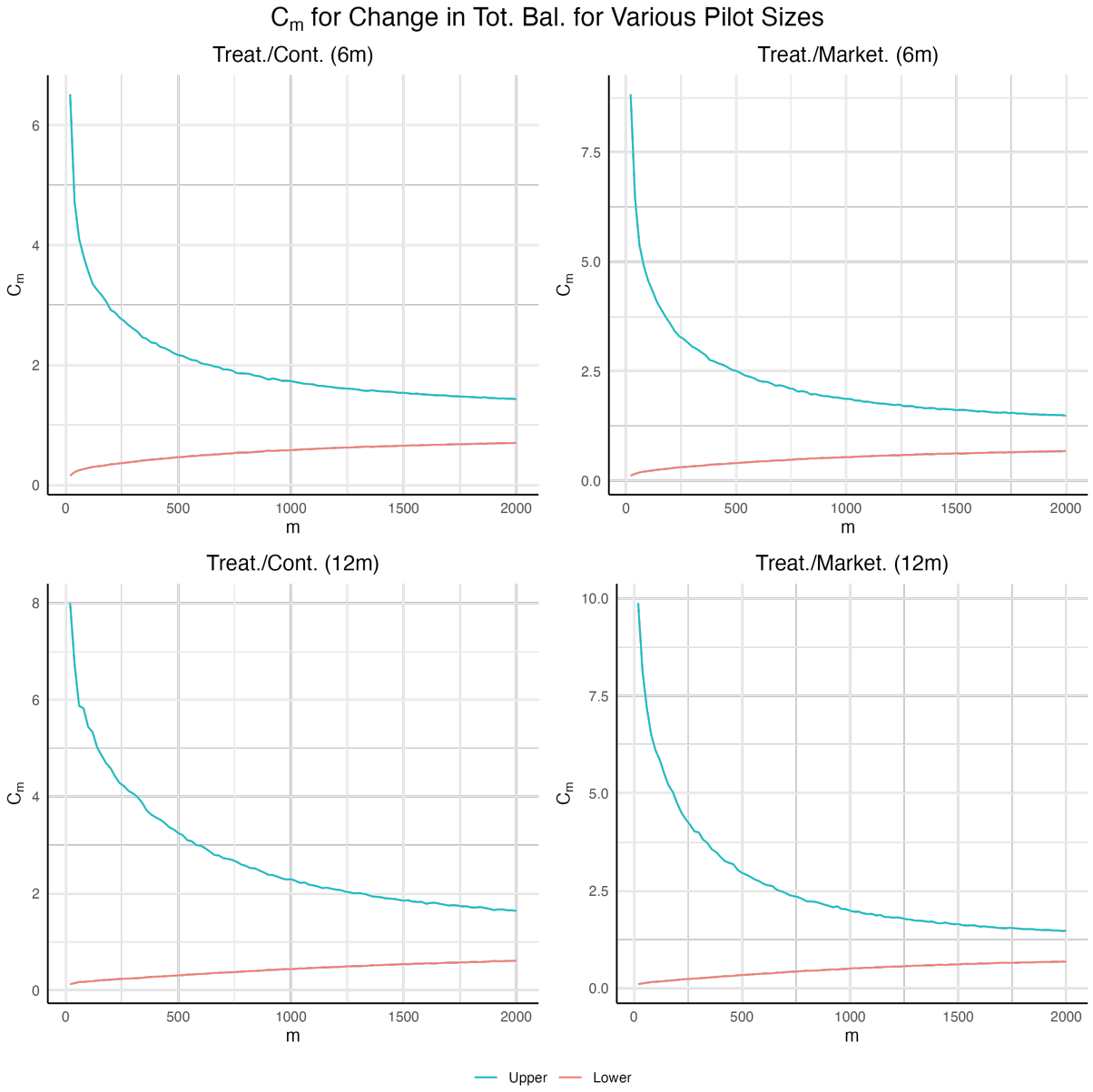

Figure 4.4. C_m for the outcomes of interest estimated using the full experiment data. Interval computed under the assumption that the pilot assigns half of the units to treatment and half to control.

of $\widetilde{p}$, so that ultimately, very large pilot sizes are needed for the FNA to outperform the balanced allocation in this example.

4.6. Conclusion

We study the properties of the Feasible Neyman Allocation (FNA) in a novel asymptotic framework for two-wave experiments that takes pilot size to be fixed as the size of the main wave tends to infinity. In this setting, the estimated allocation has error that is not negligible even in the limit. Our asymptotic model therefore corresponds more closely to the finite sample statistical problem when pilots are small and the optimal allocation may be poorly estimated. We establish the limiting distribution of the difference-in-means estimator under the FNA and characterize conditions under which it has larger asymptotic variance compared to balanced randomization, where half of the main wave is assigned to treatment and the remainder to control. This happens when the potential outcomes are relatively homoskedastic with respect to treatment status or exhibit high kurtosis – situations that may arise frequently in practice. This issue is likely to be exacerbated when observations exhibit cluster dependence or when researchers perform stratified randomization with many strata, so that the "effective observations" used for variance computations are fewer in number. Our results suggest that researchers should not use the FNA with small pilots, particularly when they believe any of the above occurs.

References

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge (2017). When Should You Adjust Standard Errors for Clustering?

Alt, J., R. Ducatez, and A. Knowles (2021a). Extremal eigenvalues of critical Erdős–Rényi graphs. *The Annals of Probability 49*(3), 1347–1401.

Alt, J., R. Ducatez, and A. Knowles (2021b). Poisson statistics and localization at the spectral edge of sparse Erdős–Rényi graphs. *arXiv preprint arXiv:2106.12519*.

Andrews, I., J. H. Stock, and L. Sun (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics 11*(1).

Angrist, J. D. and J.-S. Pischke (2008). *Mostly Harmless Econometrics*. Princeton University Press.

Ashraf, N., D. Karlan, and W. Yin (2006). Tying odysseus to the mast: Evidence from a commitment savings product in the philippines. *The Quarterly Journal of Economics 121*(2), 635–672.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association 116*(536), 1716–1730.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*(1), 347–365.

Avella-Medina, M., F. Parise, M. T. Schaub, and S. Segarra (2020). Centrality measures for graphons: Accounting for uncertainty in networks. *IEEE Transactions on Network Science and Engineering 7*(1), 520–537.

Avvisati, F., M. Gurgand, N. Guyon, and E. Maurin (2014). Getting parents involved: A field experiment in deprived schools. *Review of Economic Studies 81*(1), 57–83.

Azriel, D., M. Mandel, and Y. Rinott (2012). Optimal allocation to maximize the power of two-sample tests for binary response. *Biometrika 99*(1), 101–113.

Bai, Y. (2021). Why randomize? minimax optimality under permutation invariance. *Journal of Econometrics*.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013). The diffusion of microfinance. *Science 341*(6144), 1236498.

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2019). Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies 86*(6), 2453–2490.

Bell, R. and D. E. McCaffrey (2002). Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology, 28*(2), 169–181.

Benaych-Georges, F., C. Bordenave, and A. Knowles (2019). Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs. *The Annals of Probability 47*(3), 1653–1676.

Benaych-Georges, F., C. Bordenave, and A. Knowles (2020). Spectral radii of sparse random matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Volume 56, pp. 2141–2161. Institut Henri Poincaré.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Bertrand, M., E. Duflo, and M. Sendhil (2004). How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics, 119*, 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*(2), 137–151.

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences 106*(50), 21068–21073.

Bickel, P. J., A. Chen, and E. Levina (2011). The method of moments and degree distributions for network models. *The Annals of Statistics 39*(5), 2280–2301.

Bloch, F., M. O. Jackson, and P. Tebaldi (2021). Centrality measures in networks. *arXiv preprint arXiv:1608.05845*.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying (2014, 11). Does Working from Home Work? Evidence from a Chinese Experiment. *The Quarterly Journal of Economics 130*(1), 165–218.

Bollobás, B., S. Janson, and O. Riordan (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms 31*(1), 3–122.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology 2*(1), 113–120.

Borgatti, S. P., K. M. Carley, and D. Krackhardt (2006). On the robustness of centrality measures under conditions of imperfect data. *Social networks 28*(2), 124–136.

Borgs, C., J. T. Chayes, H. Cohn, and N. Holden (2018). Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research 18*, 1–71.

Bourlès, R., Y. Bramoullé, and E. Perez-Richet (2021). Altruism and risk sharing in networks. *Journal of the European Economic Association 19*(3), 1488–1521.

Bramoullé, Y. and G. Genicot (2018). Diffusion centrality: Foundations and extensions.

Breza, E., A. Chandrasekhar, B. Golub, and A. Parvathaneni (2019). Networks in economic development. *Oxford Review of Economic Policy 35*(4), 678–721.

Brittain, E. and J. J. Schlesselman (1982). Optimal allocation for the comparison of proportions. *Biometrics*, 1003–1009.

Bruhn, M. and D. McKenzie (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics 1*(4), 200–232.

Bryan, G., D. Karlan, and J. Zinman (2015, 08). Referrals: Peer screening and enforcement in a consumer credit field experiment. *American Economic Journal: Microeconomics 7*(3), 174–204.

Cai, J., D. Yang, W. Zhu, H. Shen, and L. Zhao (2021). Network regression and supervised centrality estimation. *Available at SSRN 3963523*.

Cai, Y., I. A. Canay, D. Kim, and A. M. Shaikh (2021). A User's Guide to Approximate Randomization Tests with a Small Number of Clusters. *Working Paper*.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008a). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics, 90*(3), 414–427.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008b). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Canay, I. A. and V. Kamat (2017, 10). Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design. *The Review of Economic Studies 85*(3), 1577–1608.

Canay, I. A. and V. Kamat (2018). Approximate permutation tests and induced order statistics in the regression discontinuity design. *The Review of Economic Studies 85*(3), 1577–1608.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017a). Randomization Tests under an Approximate Symmetry Assumption. *Econometrica, 85*(3), 1013–1030.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017b). Randomization tests under an approximate symmetry assumption. *Econometrica 85*(3), 1013–1030.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017c). Randomization tests under an approximate symmetry assumption. *Econometrica 85*(3), 1013–1030.

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017d). Supplement to 'Randomization tests under an approximate symmetry assumption'. *Econometrica Supplemental Material 85*. http://dx.doi.org/10.3982/ECTA13081.

Canay, I. A., A. Santos, and A. M. Shaikh (2019). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics (forthcoming)*.

Canay, I. A., A. Santos, and A. M. Shaikh (2021a, January). The wild bootstrap with a "small" number of "large" clusters. *The Review of Economics and Statistics 103*(2), 346–363.

Canay, I. A., A. Santos, and A. M. Shaikh (2021b). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics 103*(2), 346–363.

Candès, E. J. and T. Tao (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory 56*(5), 2053–2080.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017). Asymptotic behavior of at-test robust to cluster heterogeneity. *Review of Economics and Statistics 99*(4), 698–709.

Carvalho, V. M., M. Nirei, Y. U. Saito, and A. Tahbaz-Salehi (2021). Supply chain disruptions: Evidence from the great east japan earthquake. *The Quarterly Journal of Economics 136*(2), 1255–1321.

Carvalho, V. M. and A. Tahbaz-Salehi (2019). Production networks: A primer. *Annual Review of Economics 11*, 635–663.

Chandrasekhar, A. (2016). Econometrics of network formation. *The Oxford handbook of the economics of networks*, 303–357.

Chandrasekhar, A. and R. Lewis (2016). Econometrics of sampled networks. *Working Paper*.

Chandrasekhar, A. G., C. Kinnan, and H. Larreguy (2018). Social networks as contract enforcement: Evidence from a lab experiment in the field. *American Economic Journal: Applied Economics 10*(4), 43–78.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics 43*(1), 177–214.

Chen, L. H., L. Goldstein, and Q.-M. Shao (2011). *Normal approximation by Stein's method*, Volume 2. Springer.

Chen, Y., C. Cheng, and J. Fan (2021). Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *Annals of statistics 49*(1), 435.

Cheng, C., Y. Wei, and Y. Chen (2021). Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Transactions on Information Theory 67*(11), 7380–7419.

Chong, A., D. Karlan, J. Shapiro, and J. Zinman (2015). (ineffective) messages to encourage recycling: evidence from a randomized evaluation in peru. *The World Bank Economic Review 29*(1), 180–206.

Comola, M. and M. Fafchamps (2014). Testing unilateral and bilateral link formation. *The Economic Journal 124*(579), 954–976.

Comola, M. and M. Fafchamps (2017). The missing transfers: Estimating misreporting in dyadic data. *Economic Development and Cultural Change 65*(3), 549–582.

Conley, T., S. Gonçalves, and C. Hansen (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research 56*(4), 1139–1203.

Conley, T. G. and C. R. Taber (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics 93*(1), 113–125.

Costenbader, E. and T. W. Valente (2003). The stability of centrality measures when networks are sampled. *Social networks 25*(4), 283–307.

Cruz, C., J. Labonne, and P. Querubin (2017). Politician family networks and electoral outcomes: Evidence from the philippines. *American Economic Review 107*(10), 3006–37.

Cytrynbaum, M. (2021). Designing representative and balanced experiments by local randomization. *arXiv preprint arXiv:2111.08157*.

Dasaratha, K. (2020). Distributions of centrality on networks. *Games and Economic Behavior 122*, 1–27.

Davidson, R. and E. Flachaire (2008). The Wild Bootstrap, Tamed at Last. *Journal of Econometrics 146(1)*.

de Chaisemartin, C. and J. Ramirez-Cuellar (2019). At what level should one cluster standard errors in paired and small-strata experiments? *arXiv preprint arXiv:1906.00288*.

De Paula, A. (2017). Econometrics of network models. In *Advances in economics and econometrics: Theory and applications, eleventh world congress*, pp. 268–323. Cambridge University Press Cambridge.

De Paula, A., I. Rasul, and P. Souza (2020). Recovering social networks from panel data: identification, simulations and an application.

De Weerdt, J. and S. Dercon (2006). Risk-sharing networks and insurance against illness. *Journal of development Economics 81*(2), 337–356.

Dedecker, J., P. Doukhan, and G. Lang (2007). *Weak Dependence: With Examples and Applications*. Springer.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical association 69*(345), 118–121.

Deming, D. J., N. Yuchtman, A. Abulafi, C. Goldin, and L. F. Katz (2016, 03). The value of postsecondary credentials in the labor market: An experimental study. *American Economic Review 106*(3), 778–806.

Devroye, L. and N. Fraiman (2014). Connectivity of inhomogeneous random graphs. *Random Structures & Algorithms 45*(3), 408–420.

DiCiccio, C. J., T. J. DiCiccio, and J. P. Romano (2020). Exact tests via multiple data splitting. *Statistics & Probability Letters 166*, 108865.

Dillon, M. R., H. Kannan, J. T. Dean, E. S. Spelke, and E. Duflo (2017). Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. *Science 357*(6346), 47–55.

Djogbenou, A. A., J. G. MacKinnon, and M. Ø. Nielsen (2019). Asymptotic theory and wild bootstrap inference with clustered errors. *Journal of Econometrics 212*(2), 393–412.

Dong, H., N. Chen, and K. Wang (2020). Modeling and change detection for count-weighted multilayer networks. *Technometrics 62*(2), 184–195.

Doukhan, P. and S. Louhichi (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic processes and their applications 84*(2), 313–342.

Eagle, N., M. Macy, and R. Claxton (2010). Network diversity and economic development. *Science 328*(5981), 1029–1031.

Eldridge, J., M. Belkin, and Y. Wang (2018). Unperturbed: spectral analysis beyond davis-kahan. In *Algorithmic Learning Theory*, pp. 321–358. PMLR.

Fafchamps, M. and S. Lund (2003). Risk-sharing networks in rural Philippines. *Journal of development Economics 71*(2), 261–287.

Feige, U. and E. Ofek (2005). Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms 27*(2), 251–275.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012). The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics 127*(3), 1057–1106.

Galiani, S. and P. J. McEwan (2013). The heterogeneous impact of conditional cash transfers. *Journal of Public Economics 103*, 85–96.

Gneezy, U., J. A. List, J. A. Livingston, X. Qin, S. Sadoff, and Y. Xu (2019). Measuring Success in Education: The Role of Effort on the Test Itself. *American Economic Review: Insights 1*, 291–308.

Graham, B. S. (2020a). Network data. In *Handbook of Econometrics*, Volume 7, pp. 111–218. Elsevier.

Graham, B. S. (2020b). Sparse network asymptotics for logistic regression.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology 78*(6), 1360–1380.

Griffith, A. (2022). Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. *Journal of Labor Economics 40*(4), 779–805.

Hahn, J., K. Hirano, and D. Karlan (2011). Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics 29*(1), 96–108.

Hansen, B. E. and S. Lee (2019a). Asymptotic theory for clustered samples. *Journal of econometrics 210*(2), 268–290.

Hansen, B. E. and S. Lee (2019b). Asymptotic Theory for Clustered Samples. *Journal of Econometrics 210(2)*, 268–290.

Hochberg, Y. V., A. Ljungqvist, and Y. Lu (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance 62*(1), 251–301.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps. *Social networks 5*(2), 109–137.

Hu, A. and H. Spamann (2020). Inference with cluster imbalance: The case of state corporate laws. *AEA Law & Corporate Governance Working Paper*.

Hu, F. and W. F. Rosenberger (2003). Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association 98*(463), 671–678.

Hu, F. and W. F. Rosenberger (2006). *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons.

Ibragimov, R. and U. K. Müller (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics 28*(4), 453–468.

Ibragimov, R. and U. K. Müller (2016). Inference with Few heterogeneous Clusters. *The Review of Economics and Statistics, 98*(1), 83–96.

Ibragimov, R. and U. K. Müller (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics 98*(1), 83–96.

Imbens, G. W. and M. Kolesár (2016). Robust Standard Errors in Small Samples: Some Practical Advice. *The Review of Economics and Statistics, 98*(4), 701–712.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Jochmans, K. (2018). Semiparametric analysis of network formation. *Journal of Business & Economic Statistics 36*(4), 705–713.

Kingman, J. F. C. (1978). Uses of exchangeability. *The Annals of Probability 6*(2), 183 – 197.

Kinnan, C. and R. Townsend (2012). Kinship and financial networks, formal financial access, and risk reduction. *American Economic Review 102*(3), 289–93.

Le, C. M., E. Levina, and R. Vershynin (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms 51*(3), 538–561.

Le, C. M. and T. Li (2020). Linear regression and its inference on noisy network-linked data. *arXiv preprint arXiv:2007.00803*.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics 43*(1), 215–237.

Leider, S., M. M. Möbius, T. Rosenblat, and Q.-A. Do (2009). Directed altruism and enforced reciprocity in social networks. *The Quarterly Journal of Economics 124*(4), 1815–1851.

Levin, K. and E. Levina (2019). Bootstrapping networks with latent space structure. *arXiv preprint arXiv:1907.10821*.

Lewbel, A., X. Qu, X. Tang, et al. (2021). *Social Networks with Mismeasured Links*. Boston College.

Li, T., E. Levina, and J. Zhu (2020). Network cross-validation by edge sampling. *Biometrika 107*(2), 257–276.

Liang, K.-Y. and S. L. Zeger (1986a). Longitudinal Data Analysis for Generalized Linear Models. *Biometrika 73(1)*, 13–22.

Liang, K.-Y. and S. L. Zeger (1986b). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

Lovász, L. (2012). *Large networks and graph limits*, Volume 60. American Mathematical Soc.

MacKinnon, J. G., M. A. Nielsen, and M. D. Webb (2020). Testing for the Appropriate Level of Clustering in Linear Regression Models. *Queens Economics Department Working Paper No. 1428*.

Manresa, E. (2016). Estimating the structure of social interactions using panel data. *Working Paper*.

McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition. *American Economic Review 107*(8), 2278–2307.

Melfi, V. and C. Page (1998). Variablility in adaptive designs for estimation of success probabilities. *IMS Lecture Notes - Monograph Series*, 106–114.

Meng, X., N. Qian, and P. Yared (2015). The institutional causes of china's great famine, 1959–1961. *The Review of Economic Studies 82*(4), 1568–1611.

Motalebi, N., N. T. Stevens, and S. H. Steiner (2021). Hurdle blockmodels for sparse network modeling. *The American Statistician 75*(4), 383–393.

Munyo, I. and M. A. Rossi (2015). First-day criminal recidivism. *Journal of Public Economics 124*, 81–90.

Negahban, S. and M. J. Wainwright (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning*

Research 13(1), 1665–1697.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Journal of the Royal Statistical Society*, Volume 97, pp. 558–625.

Nze, P. A. and P. Doukhan (2004). Weak dependence: models and applications to econometrics. *Econometric Theory 20*(6), 995–1045.

Rajkumar, K., G. Saint-Jacques, I. Bojinov, E. Brynjolfsson, and S. Aral (2022). A causal test of the strength of weak ties. *Science 377*(6612), 1304–1310.

Reagans, R. and E. W. Zuckerman (2001). Networks, diversity, and productivity: The social capital of corporate r&d teams. *Organization science 12*(4), 502–517.

Rose, C. (2016). Identification of spillover effects using panel data. Technical report, Working Paper.

Rosenzweig, M. R. (1988). Risk, implicit contracts and the family in rural areas of low-income countries. *The Economic Journal 98*(393), 1148–1170.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics 2*, 881–935.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, Volume 3, pp. 249–277. Elsevier.

Schervish, M. (1995). *Theory of Statistics*. Springer Series in Statistics. Springer New York.

Segarra, S. and A. Ribeiro (2015). Stability and continuity of centrality measures in weighted graphs. *IEEE Transactions on Signal Processing 64*(3), 543–555.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Tabord-Meehan, M. (2021). Stratification trees for adaptive randomization in randomized controlled trials. *arXiv preprint arXiv:1806.05127*.

Tao, T. (2012). *Topics in random matrix theory*, Volume 132. American Mathematical Soc.

Thirkettle, M. (2019). Identification and estimation of network statistics with missing link data. *Working Paper*.

Udry, C. (1994). Risk and insurance in a rural credit market: An empirical investigation in northern nigeria. *The Review of Economic Studies 61*(3), 495–526.

van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

van der vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Veitch, V. and D. M. Roy (2019). Sampling and estimation for (sparse) exchangeable graphs. *The Annals of Statistics 47*(6), 3274–3299.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, Volume 47. Cambridge university press.

Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer Series in Operations Research and Financial Engineering. Springer New York.

Wigner, E. P. (1957). Characteristics vectors of bordered matrices with infinite dimensions ii. *Annals of Mathematics*, 203–207.

Xu, G. (2018). The costs of patronage: Evidence from the british empire. *American Economic Review 108*(11), 3170–98.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134(2)*, 557–598.

Young, S. J. and E. R. Scheinerman (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pp. 138–149. Springer.

APPENDIX A

Appendix to Chapter 1

A.1. Bias of Diffusion under noise ($\hat{\beta}^{(T)}$)

Tables A.1 and A.2 presents $b_T(t, \delta)$ used for calculating the bias estimator in Case (b) of Theorem 1.5. In practice, papers rarely compute $T > 5$. We provide these terms for $T \leq 10$. Functions for computing the bias terms for arbitrary T are available from the author's website.

Each row of Tables A.1 and A.2 provide the coefficients for δ^s in $b_T(t, \delta)$, for a particular T and t. To obtain the bias formula for a given T, sum across all t's for a given T. For example, when $T = 2$, the correction term is

$$\left(\delta^2 - 3\delta^3 + 3\delta^4\right) \iota_n' \hat{A} \iota_n + \left(3\delta^3 - 2\delta^4\right) \iota_n' \hat{A}^2 \iota_n + \left(2\delta^4\right) \iota_n' \hat{A}^3 \iota_n \ .$$

T	t	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7	δ^8	δ^9	δ^{10}
1	1	1								
	1	1	-3	3						
2	2		3	-2						
	3			2						
	1	1	-3	7	-4	-8				
	2		3	-4		10				
3	3			5	-2	-2				
	4				4	-1				
	5					2				
	1	1	-3	7	-13	-15	91	-182		
	2		3	-4	5	24	-94	160		
	3			5	-7	-1	36	-84		
4	4				8	-6	-2	27		
	5					7	-5			
	6						5	-4		
	7							3		
	1	1	-3	7	-13	-4	161	-500	952	-654
	2		3	-4	5	24	-178	450	-740	314
	3			5	-7	11	57	-222	456	-362
	4				8	-15	6	66	-225	317
5	5					12	-14	4	59	-148
	6						11	-13	6	32
	7							9	-12	6
	8								7	-8
	9									4

Table A.1. Coefficients for the bias of diffusion centrality, $b_T(t, \delta)$. Blanks indicate a coefficient of 0.

T	t	δ^2	δ^3	δ^4	δ^5	δ^6	δ^7	δ^8	δ^9	δ^{10}	δ^{11}	δ^{12}	δ^{13}	δ^{14}	δ^{15}	δ^{16}	δ^{17}	δ^{18}	δ^{19}	δ^{20}
	1	1	-3	7	-13	-4	184	-819	1869	-1935	-3737	15981								
	2		3	-4	5	24	-229	770	-1496	954	4740	-14604								
	3			5	-7	11	52	-346	869	-1053	-1526	7728								
	4				8	-15	28	81	-406	915	-674	-1883								
	5					12	-28	22	99	-432	814	-312								
6	6						17	-27	19	94	-359	485								
	7							16	-26	21	67	-209								
	8								14	-25	21	35								
	9									12	-21	15								
	10										9	-13								
	11											5								
	1	1	-3	7	-13	-4	184	-1097	3043	-3843	-6915	49578	-115459	80767						
	2		3	-4	5	24	-229	1088	-2552	2198	9142	-45912	90964	-37634						
	3			5	-7	11	52	-431	1363	-2000	-2751	23692	-59966	51305						
	4				8	-15	28	63	-565	1565	-1556	-5348	25674	-39288						
	5					12	-28	59	102	-692	1717	-1266	-5929	18469						
	6						17	-47	52	130	-739	1625	-718	-4616						
7	7							23	-46	48	127	-669	1263	-272						
	8								22	-45	50	100	-516	758						
	9									20	-44	50	68	-306						
	10										18	-40	44	33						
	11											15	-32	29						
	12												11	-19						
	13													6						
	1	1	-3	7	-13	-4	184	-1097	4525	-7293	-6944	83485	-256225	310955	520469	-2445004				
	2		3	-4	5	24	-229	1088	-3969	4898	11110	-79100	207636	-174066	-652454	2204430				
	3			5	-7	11	52	-431	1963	-3592	-2432	39251	-130367	185191	177779	-1142919				
	4				8	-15	28	63	-696	2366	-3194	-7525	52268	-127455	81923	320735				
	5					12	-28	59	58	-912	2822	-3094	-10153	55373	-107794	30033				
	6						17	-47	110	108	-1096	3070	-2701	-12050	51958	-73965				
	7							23	-73	102	146	-1166	2990	-1896	-11737	35122				
8	8								30	-72	97	145	-1099	2591	-1255	-7492				
	9									29	-71	99	118	-943	2064	-940				
	10										27	-70	99	86	-732	1305				
	11											25	-66	93	51	-426				
	12												22	-58	78	18				
	13													18	-45	49				
	14														13	-26				
	15															7				
	1	1	-3	7	-13	-4	184	-1097	4525	-12637	-1360	110968	-420423	695532	681922	-7068240	17664266	-13712687		
	2		3	-4	5	24	-229	1088	-3969	9368	8424	-109118	350830	-448472	-1023716	6464948	-13721396	6724528		
	3			5	-7	11	52	-431	1963	-6002	474	51064	-209648	389396	179724	-3281313	9018463	-8293240		
	4				8	-15	28	63	-696	3367	-5914	-6669	77446	-236871	246153	850584	-4132503	6454139		
	5					12	-28	59	58	-1079	4123	-6290	-11274	91953	-236476	139244	993350	-2918136		
	6						17	-47	110	18	-1366	4803	-6352	-15294	101721	-231751	105547	672714		
	7							23	-73	188	79	-1614	5172	-5880	-18707	102925	-204469	74412		
	8								30	-107	179	128	-1709	5108	-4789	-19615	88499	-132335		
9	9									38	-106	173	129	-1645	4670	-3937	-15825	53896		
	10										37	-105	175	102	-1486	4121	-3585	-8366		
	11											35	-104	175	70	-1274	3362	-2653		
	12												33	-100	169	35	-968	2058		
	13													30	-92	154	2	-542		
	14														26	-79	125	-16		
	15															21	-60	76		
	16																15	-34		
	17																	8		
	1	1	-3	7	-13	-4	184	-1097	4525	-12637	10890	125858	-597596	1235468	309432	-11044097	36311072	-45046667	-75501514	347810028
	2		3	-4	5	24	-229	1088	-3969	9368	211	-130730	514002	-866532	-979926	10410814	-28739014	23733098	95980642	-310601986
	3			5	-7	11	52	-431	1963	-6002	6504	56591	-292720	663931	-81390	-5051600	18395415	-26721605	-26472010	165085728
	4				8	-15	28	63	-696	3367	-9914	-1678	97369	-363531	535304	1071910	-8050856	19840070	-13232554	-47818679
	5					12	-28	59	58	-1079	5644	-11058	-7738	125398	-398107	394043	1745080	-8688965	16108742	-2368523
	6						17	-47	110	18	-1535	6719	-11878	-13503	148861	-422298	318022	1897993	-7599134	9772214
	7							23	-73	188	-84	-1901	7684	-12237	-18749	165197	-429048	278089	1667351	-4861956
	8								30	-107	301	-11	-2225	8199	-11710	-23857	171609	-407905	236540	1042204
	9									38	-150	291	50	-2347	8154	-10298	-26127	159555	-335522	149307
10	10										47	-149	284	53	-2286	7674	-9217	-22752	124908	-199857
	11											46	-148	286	26	-2124	7102	-8832	-15237	70335
	12												44	-147	286	-6	-1911	6343	-7913	-6786
	13													42	-143	280	-41	-1605	5040	-5264
	14														39	-135	265	-74	-1179	2982
	15															35	-122	236	-92	-637
	16																30	-103	187	-76
	17																	24	-77	111
	18																		17	-43
	19																			9

Table A.2. Coefficients for the bias of diffusion centrality, $b_T(t, \delta)$ (continued from Table A.1). Blanks indicate a coefficient of 0.

A.2. Additional Motivation for Econometric Framework

This section discusses the connection of our econometric framework to the "Weak Ties" theory of social networks. We also present other examples of network data that fit our framework.

A.2.1. "Weak Ties" Theory

In the seminal paper titled "The Strength of Weak Ties", Granovetter (1973) argues that lower intensity links, which constitute most of any given person's relationships, are the key drivers of many important social and economic outcomes. For example, in tracing the network of job referrals, the author finds that 83% of recent job changers in a Boston suburb found their new jobs through friends whom they saw fewer than twice a week, and who were only "marginally included in the current network of contacts". The author further notes: "It is remarkable that people receive crucial information from individuals whose very existence they have forgotten." A series of empirical work has found evidence in favor of the weak ties theory across diverse applications such as innovation (e.g. Reagans and Zuckerman 2001), economic development (e.g. Eagle et al. 2010) and job referrals (e.g. Rajkumar et al. 2022). This theory lends credence to our econometric model, in which an unobserved network of weak ties not only drives economic effects but also generates a sparse observed network.

A.2.2. Additional Examples

Example A.1. Carvalho et al. (2021) studies the propagation of shocks through production networks during the Great East Japanese Earthquake of 2011. In the ideal

production network, A_{ij} records the value of i's sales to j as a proportion of the value of i's total sales. In turn, A_{ij} depends on U_i and U_j, which might index the quality of a firm's product, with higher quality firms requiring more and higher quality inputs. However, these variables are not observed. Instead, the authors have access to data from a credit reporting agency which includes supplier and customer information for firms. The authors explicitly note two limitations in their data: "First, it only reports a binary measure of interfirm supplier-customer relations... we do not observe a yen measure associated with their transactions. Second, the forms used by [the credit agency] limit the number of suppliers and customers that firms can report to 24 each." Suppose firms only report suppliers from whom they receive delivery during the month in which the forms are filed. Then a supplier that sends fewer inputs are more likely to be omitted in any given month. Abstracting away from concerns about network censoring (see Griffith 2022), the conditional independence assumption would also be satisfied if the delivery schedules for suppliers are independent.

Example A.2. Xu (2018) studies how patronage affected the promotion and performance of bureaucrats in the Colonial Office of the British Empire. In the ideal network for measuring patronage, A_{ij} records intensity of the friendship between i and j. Here, U_i might index traits such as gregariousness, polo skills and drinking habits among others. Bureaucrats having more in common with their patrons may be more likely to be recommended for promotion. However, the link intensity between bureaucrats are not observed. Instead, the paper proxies for relationships using indicators for shared ancestry, membership of social groups (such as the aristocracy) or attendance of the same elite school or university. In this context, our data-generating process means that bureaucrats who are

closer are more likely to satisfy the above criteria for connection. The conditional independence assumption would be satisfied if the lack of observation are independent across agent pairs, e.g. if some university records were randomly lost.

A.3. Eigenvector Regularization

As our analysis in Section 1.3.1 shows, regression with eigenvector centrality is more sensitive to sparsity under measurement error than degree or diffusion centralities. In this section, we propose a regularization method that makes eigenvector centrality competitive with the alternatives.

Definition A.1 (Regularized Eigenvector Centrality)**.** Suppose p_n is known. Let

$$\lambda_i := \min\left\{\frac{2np_n}{\hat{C}_i^{(1)}}, 1\right\} .$$

Then, define $\hat{A}_\lambda$ to be the regularized version of $\hat{A}$, where

$$\left(\hat{A}_\lambda\right)_{ij} = \sqrt{\lambda_i\lambda_j}\hat{A}_{ij} .$$

Finally, define regularized eigenvector centrality and the corresponding OLS estimator to be:

$$\hat{C}_\lambda^{(\infty)} = a_n v_1\left(\hat{A}_\lambda\right) \quad , \quad \hat{\beta}_\lambda^{(\infty)} = \frac{Y'\hat{C}_\lambda^{(\infty)}}{\left(\hat{C}_\lambda^{(\infty)}\right)'\hat{C}_\lambda^{(\infty)}} .$$

Our proposed measure is the principal eigenvector of $\hat{A}_\lambda$, which is in turn a regularized version of the observed adjacency matrix $\hat{A}$. The regularization technique, proposed in Le et al. (2017), re-weights edges so that in $\hat{C}_{\lambda,i}^{(1)} \leq 2np_n$ for all $i \in [n]$. It is well-known that high-degree vertices interfere with concentration of random matrices and that their removal solves the problem (Feige and Ofek 2005). However, such a drastic procedure is not ideal: intuition suggests that high degree vertices are important in a network, forming hubs that connect many individuals. Le et al. (2017) shows that re-weighting the edges

of high-degree vertices is sufficient to enforce concentration. In turn, we have consistency of $\hat{\beta}_\lambda^{(\infty)}$ as our next theorem asserts.

Theorem A.1 (Consistency with Regularized Eigenvector Centrality)**.** Suppose Assumptions 1.1 and 1.2 hold. Suppose further that $E[\varepsilon_i|U_i] = 0$ and $E[\varepsilon_i^2] = \sigma^2 < \infty$ and $\lambda_1(f) > \lambda_2(f)$. Then, $a_n \to \infty$ and $p_n \succ n^{-1}$ implies that $\tilde{\beta}^{(\infty)} \xrightarrow{p} \beta^{(\infty)}$.

Our result shows that $\hat{\beta}_\lambda^{(\infty)}$ is able to accommodate as much sparsity as $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(T)}$. As such, when faced with sparse matrices, researchers could benefit from using regularized eigenvector centrality in their regression instead. One difficulty with using the method is that p_n is not known in practice. It is not possible to estimate p_n since the graphon f is unknown. Under a mild assumption, however, the following is possible:

Corollary A.1 (Estimated p_n)**.** Suppose $\int_{[0,1]^2} f(u,v)\,dudv \geq M > 0$. Let

$$\rho_n = p_n \int_{[0,1]^2} f(u,v)\,dudv \quad , \quad \hat{\rho}_n = \frac{\iota_n' A \iota_n}{n(n-1)} \ .$$

Next define

$$\hat{\lambda}_i = \min\left\{ \frac{3n\hat{\rho}_n}{M \cdot \hat{C}_i^{(1)}} \, , \, 1 \right\} \ .$$

Using $\hat{\lambda}$ in place of λ in Definition A.1 does not change the conclusions of Theorem A.1.

A.3.1. Proof of Theorem A.1

As in the Proof of Theorem 1.2, write

$$\hat{\beta}_\lambda^{(\infty)} = \beta^{(\infty)} + \beta^{(\infty)} \left(v_1(\hat{A}_\lambda) \right)' \left(v_1(A) - v_1(\hat{A}_\lambda) \right) + \frac{1}{\sqrt{a_n}} v_1(\hat{A}_\lambda)' \varepsilon^{(\infty)} \ .$$

By Theorem and Remark 2.2 of Le et al. (2017), with probability at least $1 - n^{-r}$,

$$\left\| A - \hat{A} \right\| \leq k r^{3/2} \sqrt{n p_n}$$

where k is a universal constant. Therefore, by the Davis-Kahan inequality (Theorem 4.5.5 in Vershynin 2018),

$$\left\| v_1(A) - v_1(\hat{A}) \right\| \leq \frac{\|\hat{A} - A\|}{n p_n (\lambda_1 - \lambda_2)} = O_p\left(\frac{1}{\sqrt{n p_n}} \right) = o_p(1) \ .$$

Again, note that

$$E\left[v_1(\hat{A})' \varepsilon^{(\infty)} \,\middle|\, U \right] \leq \|v_1(\hat{A})\| \bar{\sigma}^2 = \bar{\sigma}^2.$$

Since $a_n \to \infty$, conclude that $\frac{1}{\sqrt{a_n}} v_1(\hat{A})' \varepsilon^{(\infty)} \xrightarrow{p} 0$ and that $\hat{\beta}_\lambda^{(\infty)} \xrightarrow{p} \beta^{(\infty)}$.

A.3.2. Proof of Corollary A.1

We first note that $\hat{\rho}$ is a good estimator of ρ_n:

Theorem A.2 (Theorem 1, Bickel et al. 2011)**.** Under Assumption 1.1 and 1.2,

$$\sqrt{n} \left(\frac{\hat{\rho}_n}{\rho_n} - 1 \right) \xrightarrow{d} N(0, \sigma^2)$$

for some $\sigma^2 > 0$.

Next, noting that $M / \int f(u, v) \leq 1$

$$P\left(\frac{\hat{\rho}_n}{M} \geq \frac{2 M p_n}{3 \int f(u, v) \, , \, du dv} \right) \to 1$$

Setting

$$\hat{\lambda}_i = \min\left\{ \frac{3n\hat{\rho}_n}{M \cdot C_i^{(1)}} \, , \, 1 \right\}$$

ensures that w.p.a. 1,

$$\max_{i\in[n]} C_{\lambda,i}^{(1)} \leq \frac{3n\hat{\rho}_n}{M} \leq \frac{2\int f(u,v)\, dudv}{M} \cdot np_n \, .$$

By Remark 2.1 of Le et al. (2017), the oracle procedure re-weights edges adjacent to fewer than $10/(np_n)$ nodes. Since $\hat{\lambda}_i \geq \lambda_i$, re-weighting using $\hat{\lambda}$ therefore also alters edges adjacent to fewer than $10/(np_n)$ nodes. As such, by Theorem 2.1 of $\|\hat{A}_{\hat{\lambda}} - A\| = O(\sqrt{np_n})$ w.p.a. 1. The proof then proceeds as in that of Theorem A.1.

A.4. Proofs

Without loss of generality, let $f(u,v) = f(v,u)$ and the $f(u,u) = 0$. Further define:

$$W = \int_{[0,1]^2} f(u,v)\, dudv \ .$$

We state below a convenient lemma:

Lemma A.1 (Concentration in Spectral Norm)**.** Suppose Assumptions 1.1 and 1.2 hold. Let $\nu \in (0,1)$. Then with probability at least $1 - \exp\left(-n^2 p_n^2 \sqrt{k \frac{\log n}{\log\log n}}\right)$

$$\left\| A - \hat{A} \right\| \leq k \left(np_n\right)^{(1+\nu)/2} \left(\frac{\log n}{\log\log n}\right)^{(1-\nu)/4}$$

where k is a universal constant. In other words,

$$\left\| A - \hat{A} \right\| = O_p\left(\left(np_n\right)^{(1+\nu)/2} \left(\frac{\log n}{\log\log n}\right)^{(1-\nu)/4} \right) \ . \tag{A.1}$$

A.4.1. Proof of Theorem 1.1

In the setting with no measurement error, we write:

$$\tilde{\beta}^{(d)} = \frac{\sum_{i=1}^n Y_i C_i^{(d)}}{\sum_{i=1}^n \left(C_i^{(d)}\right)^2} = \beta + \frac{\sum_{i=1}^n C_i^{(d)} \varepsilon_i^{(d)}}{\sum_{i=1}^n \left(C_i^{(d)}\right)^2} \ .$$

We first show that OLS is consistent when the lower bounds in the Theorem obtains. Start with degree:

$$\begin{aligned}\sum_{i=1}^{n} C_i^{(1)}\varepsilon_i^{(1)} &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_n f(U_i,U_j)\varepsilon_i^{(1)} \\ &= \frac{p_n}{2}\binom{n}{2}\cdot\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n} f(U_i,U_j)\varepsilon_i^{(1)} + f(U_j,U_i)\varepsilon_j^{(1)} \\ &= O_p\left(n^{3/2}p_n\right) \ ,\end{aligned} \tag{A.2}$$

In the last equality, we use our assumption that $E\left[\varepsilon_i^{(d)}|U_i\right] = 0$ and $E\left[\left(\varepsilon_i^{(d)}\right)^2|U_i\right] \leq \bar{\sigma}^2 < \infty$, so that

$$\sqrt{n}\cdot\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n} f(U_i,U_j)\varepsilon_i^{(1)} + f(U_j,U_i)\varepsilon_j^{(1)} \xrightarrow{d} \mathrm{N}(0,\gamma)$$

for some $\gamma > 0$ by the standard CLT for U-statistics (e.g. Theorem 12.3 in Van der Vaart 2000). Similarly,

$$\begin{aligned}\sum_{i=1}^{n}\left(C_i^{(1)}\right)^2 &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} p_n f(U_i,U_j)\right)^2 \\ &= p_n^2\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} f(U_i,U_j)f(U_i,U_k) \\ &= p_n^2\binom{n}{3}\cdot\frac{1}{\binom{n}{3}}\sum_{k=1}^{n} f(U_i,U_j)f(U_i,U_k) = O_p(n^3p_n^2)\end{aligned} \tag{A.3}$$

By the LLN for U-statistics,

$$\frac{1}{\binom{n}{3}}\sum_{k=1}^{n} f(U_i,U_j)f(U_i,U_k) \xrightarrow{p} \gamma$$

Next, note that $\gamma > 0$. Let k and B_k be such that $f(u,v) > 1/k$ for all $(u,v) \in B_k$. Our assumption $W > 0$ ensures that there exists k such that $P_k := \int_{[0,1]^2} \mathbf{1}_{B_k}\, dudv > 0$. Then,

$$\begin{aligned}\gamma &= \int_{[0,1]^3} f(u_1,u_2)f(u_1,u_3)\, du_1du_2du_3 \\ &\geq \int_{\pi_1(B_k)\times\pi_2(B_k)\times\pi_2(B_k)} k^{-2}\, du_1du_2du_3 \geq P_k^2k^{-2} > 0\end{aligned}$$

where $\pi_j(B_k)$ denotes the projection of B_k onto the j^{th} coordinate. Hence, we have consistency if $n^{3/2}p_n \to \infty$. If $n^{3/2}p_n \approx 1$, the $\tilde{\beta}^{(1)} - \beta^{(1)}$ converges to a normal distribution. If $n^{3/2}p_n \prec 1$, $\tilde{\beta}^{(1)} - \beta^{(1)}$ diverges. Hence, we have consistency if and only if $n^{3/2}p_n \to \infty$

Next, consider diffusion centrality. Note that:

$$\sum_{i=1}^n C_i^{(T)}\varepsilon_i^{(T)} = \sum_{t=1}^T \delta^t \cdot \iota_n' A^t \varepsilon^{(T)} \quad , \quad \sum_{i=1}^n \left(C_i^{(T)}\right)^2 = \sum_{t=1}^T \delta^{2t} \cdot \iota_n' A^{2t} \iota_n \ .$$

We will identify the dominant terms in the numerator and denominator respectively in each regime of p_n. For $t \geq 2$, write:

$$\left[A^t\right]_{ij} = p_n^t \sum_{k_1=1}^n \sum_{k_2=1}^n \cdots \sum_{k_{t-1}=1}^n f(U_i,U_{k_1})f(U_{k_1},U_{k_2})\cdots f(U_{k_{t-1}},U_j) \ .$$

Applying the CLT for U-statistics as before, we have that

$$\iota_n' A^t \varepsilon^{(T)} = p_n^t \sum_{i=1}^n \sum_{j=1}^n \sum_{k_1=1}^n \sum_{k_2=1}^n \cdots \sum_{k_{t-1}=1}^n f(U_i,U_{k_1})f(U_{k_1},U_{k_2})\cdots f(U_{k_{t-1}},U_j)\varepsilon_j^{(T)} = O_p\left(p_n^t n^{t+1/2}\right) \ .$$

Similarly,

(A.4)

$$\iota_n' A^{2t} \iota_n = p_n^{2t} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_{2t-1}=1}^{n} f(U_i, U_{k_1}) f(U_{k_1}, U_{k_2}) \cdots f(U_{k_{2t-1}}, U_j) = O_p\left(p_n^{2t} n^{2t+1}\right) \ .$$

Next, suppose $np_n \succ 1$. Then the dominant terms in the numerator and denominator are of order $O(p_n^T n^{T+1/2})$ and $O(p_n^{2T} n^{2T+1})$ respectively. As such,

$$\tilde{\beta}^{(T)} - \beta^{(T)} = O_p(p_n^{-T} n^{-T-1/2}) = o_p(1) \ .$$

Suppose instead that $np_n \approx 1$. Then, all terms in the numerator are of the same order. The same is true for the denominator. As before, $\tilde{\beta}^{(T)} - \beta^{(T)} = O_p(n^{-1/2}) = o_p(1)$.

Finally, suppose $np_n \prec 1$. In this regime, diffusion is equivalent to degree to a first order. The dominant terms in the numerator and denominator are of order $O(p_n n^{3/2})$ and $O(p_n^2 n^3)$ respectively. Then, as before, we obtain consistency if and only if $n^{3/2} p_n \to \infty$.

Lastly, consider eigenvector centrality. Given our assumptions, $v_1(A)$ is well-defined with high probability. Next note that by construction, $\sum_{i=1}^{n} \left(C_i^{(\infty)}\right)^2 = a_n^2$. Furthermore, by our assumptions,

$$E\left[\sum_{i=1}^{n} C_i^{(\infty)} \varepsilon_i^{(\infty)} \,\middle|\, U\right] = \sum_{i=1}^{n} C_i^{(\infty)} E\left[\varepsilon_i^{(\infty)} \,\middle|\, U\right] = 0$$

$$\text{Var}\left[\sum_{i=1}^{n} C_i^{(\infty)} \varepsilon_i^{(\infty)} \,\middle|\, U\right] = \sum_{i=1}^{n} \left(C_i^{(d)}\right)^2 \text{Var}\left[\varepsilon_i^{(\infty)} \,\middle|\, U\right] \leq a_n^2 \bar{\sigma}^2 \ .$$

As such,

$$\text{Var}\left[\tilde{\beta}^{(\infty)} - \beta^{(\infty)}\right] \leq \frac{\bar{\sigma}^2}{a_n^2} \to 0 \text{ if } a_n \to \infty \ .$$

Thus, $a_n \to \infty$ implies that $\tilde{\beta}^{(\infty)} \xrightarrow{L_2} \beta^{(\infty)}$.

Necessity follows from the counterexample in our main text, reproduced here for completeness. Suppose $f = p_n \cdot 1$ so that $A = p_n \iota_n \iota_n'$. Then $C^{(\infty)}(A) = a_n \iota_n / \sqrt{n}$. Hence,

$$\tilde{\beta}^{(\infty)} = \frac{\sqrt{n}}{a_n} \cdot \frac{Y' \iota_n}{\iota_n' \iota_n} = \beta^{(\infty)} + \frac{1}{a_n \sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^{(\infty)} \ .$$

Under our assumptions, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^{(\infty)} \xrightarrow{d} N(0, \mathrm{Var}[\varepsilon_i^{(\infty)}])$. For $\tilde{\beta}^{(\infty)}$ to be consistent for $\beta^{(\infty)}$, it is therefore necessary for $a_n \to \infty$.

A.4.2. Proof of Theorem 1.2

We first write:

$$\hat{\beta}^{(d)} = \beta^{(d)} + \beta^{(d)} \frac{\left(\hat{C}^{(d)}\right)' \left(C^{(d)} - \hat{C}^{(d)}\right)}{\left(\hat{C}^{(d)}\right)' \hat{C}^{(d)}} + \frac{\left(\hat{C}^{(d)}\right)' \varepsilon^{(d)}}{\left(\hat{C}^{(d)}\right)' \hat{C}^{(d)}}$$

For convenience, denote

$$\hat{A}_{ij} = p_n f(U_i, U_j) + \xi_{ij} \quad , \quad E[\xi_{ij} \,|\, U_i, U_j] = 0 \ .$$

Also let $\xi_i = \sum_{j=1}^{n} \xi_{ij} = \sum_{j \neq i} \xi_{ij}$ and $\xi = (\xi_1, ..., \xi_n)'$. Finally, let $\boldsymbol{\xi}$ by the $n \times n$ matrix with $(i,j)^{\text{th}}$ entry ξ_{ij}. Note that $\xi = \boldsymbol{\xi} \iota_n$. By Assumption 1.2, $\xi_{ij} \perp\!\!\!\perp \varepsilon_k^{(d)} \,|\, U$ for all i, j, k and $d \in \{1, T, \infty\}$.

A.4.2.1. Degree. We first show that $np_n \succ 1$ is sufficient for consistency of $\hat{\beta}^{(1)}$. Using our new notation, the numerator is:

$$\left(\hat{C}^{(1)}\right)' \varepsilon^{(1)} = C^{(1)} \varepsilon^{(1)} + \xi' \varepsilon^{(1)} \ .$$

By conditional independence of ξ and $\varepsilon^{(1)}$, $E\left[\xi'\varepsilon^{(1)}\right] = 0$

$$\begin{aligned}
\mathrm{Var}\left[\xi'\varepsilon^{(1)} \,|\, U\right] &= \mathrm{Var}\left[2\sum_{i=1}^{n}\sum_{j>i}\xi_{ij}\varepsilon_i^{(1)} \,\middle|\, U\right] = 4\sum_{i=1}^{n}\mathrm{Var}\left[\varepsilon_i^{(1)}\sum_{j>i}\xi_{ij} \,\middle|\, U\right] \\
&= 4\sum_{i=1}^{n}\left(E\left[\left(\varepsilon_i^{(1)}\right)^2 \,|\, U\right]\sum_{j>i}E\left[\xi_{ij}^2 \,|\, U\right]\right) \\
&\leq 2\bar{\sigma}^2\sum_{i=1}^{n}\sum_{j>i}p_n f(U_i,U_j)\left(1-p_n f(U_i,U_j)\right) \\
&\leq 2\bar{\sigma}^2\sum_{i=1}^{n}\sum_{j>i}p_n f(U_i,U_j)
\end{aligned}$$

Taking expectations over U, we have that

$$\mathrm{Var}\left[\xi'\varepsilon^{(1)}\right] \leq 2\bar{\sigma}^2 n^2 p_n \cdot W \quad \Rightarrow \quad \xi'\varepsilon^{(1)} = O_p(np_n{}^{1/2})$$

Given Equation (A.2), $C^{(1)}\varepsilon^{(1)} = O_p(n^{3/2}p_n)$ is thus dominant in the numerator if $np_n \succ 1$.

Next, consider the denominator, which has the form:

$$\left(\hat{C}^{(1)}\right)'\hat{C}^{(1)} = \left(C^{(1)}\right)'C^{(1)} + 2\left(C^{(1)}\right)'\xi + \xi'\xi \ .$$

We bound the last term in L_1-norm. Observe that it has conditional expectation:

$$\begin{aligned}
E\left[\xi'\xi \,|\, U\right] &= E\left[\sum_{i=1}^{n}\left(\sum_{j\neq i}\xi_{ij}\right)^2 \,\middle|\, U\right] = \sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i}E\left[\xi_{ij}\xi_{ik} \,\middle|\, U\right] \\
&= \sum_{i=1}^{n}\sum_{j\neq i}E\left[\xi_{ij}^2 \,\middle|\, U\right] \leq \sum_{i=1}^{n}\sum_{j\neq i}p_n f(U_i,U_j) \ .
\end{aligned}$$

Taking expectations over U,

$$E\left[\xi'\xi\right] \leq n^2 p_n \cdot W \quad \Rightarrow \quad \xi'\xi = O_p(n^2 p_n)\ . \tag{A.5}$$

Next, consider the middle term, which we will bound in L_2-norm. Write

$$\begin{aligned} E\left[\left(\left(C^{(1)}\right)'\xi\right)^2 \,\middle|\, U\right] &= E\left[\sum_{i=1}^n \sum_{j=1}^n C_i^{(1)}\xi_i C_j^{(1)}\xi_j \,\middle|\, U\right] \\ &= E\left[\sum_{i=1}^n C_i^{(1)}\xi_i C_i^{(1)}\xi_i \,\middle|\, U\right] + E\left[\sum_{i=1}^n \sum_{j\neq i}^n C_i^{(1)}\xi_i C_j^{(1)}\xi_j \,\middle|\, U\right]\ . \end{aligned}$$

Note that

$$\begin{aligned} E\left[\sum_{i=1}^n \sum_{j\neq i}^n C_i^{(1)}\xi_i C_j^{(1)}\xi_j \,\middle|\, U\right] &= \sum_{i=1}^n \sum_{j\neq i}^n C_i^{(1)} C_j^{(1)} E\left[\xi_i\xi_j \,|\, U\right] = \sum_{i=1}^n \sum_{j\neq i}^n C_i^{(1)} C_j^{(1)} E\left[\xi_{ij}^2 \,|\, U\right] \\ &\leq \sum_{i=1}^n \sum_{j\neq i}^n C_i^{(1)} C_j^{(1)} p_n f(U_i, U_j) \\ &= \sum_{i=1}^n \sum_{k=1}^n \sum_{l=1}^n \sum_{j\neq i}^n p_n^3 f(U_i, U_k) f(U_i, U_l) f(U_i, U_j) \end{aligned}$$

The second equality above follows form the fact that when $i \neq j$, $E[\xi_{ik}\xi_{jl} \,|\, U] = 0$ unless $k = j$ and $l = i$. Furthermore,

$$\begin{aligned} E\left[\sum_{i=1}^n C_i^{(1)}\xi_i C_i^{(1)}\xi_i \,\middle|\, U\right] &\leq \sum_{i=1}^n \left(C_i^{(1)}\right)^2 E\left[\xi_i^2 \,|\, U\right] = \sum_{i=1}^n \left(C_i^{(1)}\right)^2 E\left[\sum_{j\neq i} \xi_{ij}^2 \,\middle|\, U\right] \\ &\leq \sum_{i=1}^n \left(C_i^{(1)}\right)^2 \sum_{j\neq i} p_n f(U_i, U_j) \\ &\leq \sum_{i=1}^n \sum_{k=1}^n \sum_{l=1}^n p_n^2 f(U_i, U_k) f(U_i, U_l) \sum_{j\neq i} p_n f(U_i, U_j) \end{aligned}$$

Taking expectation over two displays above,

$$E\left[\left(\left(C^{(1)}\right)'\xi\right)^2\right] = O(n^4 p_n^3) \quad \Rightarrow \quad \left(C^{(1)}\right)'\xi = O_p(n^2 p_n^{3/2}) \ . \tag{A.6}$$

By Equation (A.3), $\left(C^{(1)}\right)' C^{(1)} = O_p(n^3 p_n^2)$. Putting the rates we derived together, the denominator is

$$\left(\hat{C}^{(1)}\right)' \hat{C}^{(1)} = O_p(n^3 p_n^2) + O_p(n^2 p_n^{3/2}) + O_p(n^2 p_n) \ . \tag{A.7}$$

Hence, $np_n \succ 1$ implies that

$$\frac{\left(\hat{C}^{(1)}\right)' \varepsilon^{(1)}}{\left(\hat{C}^{(1)}\right)' \hat{C}^{(1)}} \xrightarrow{p} 0 \ .$$

It remains to note that

$$\left(\hat{C}^{(1)}\right)' \left(C^{(1)} - \hat{C}^{(1)}\right) = \left(C^{(1)}\right)'\xi + \xi'\xi$$

so that by the rates in (A.5), (A.6) and (A.7),

$$\beta^{(d)} \frac{\left(\hat{C}^{(d)}\right)' \left(C^{(d)} - \hat{C}^{(d)}\right)}{\left(\hat{C}^{(d)}\right)' \hat{C}^{(d)}} = O_p(n^{-1} p_n^{-1}) \xrightarrow{p} 0 \ .$$

We can loosely write the above results as

$$\hat{\beta}^{(1)} - \beta^{(1)} \approx \frac{n^2 p_n^{3/2} + n^2 p_n}{n^3 p_n^2 + n^2 p_n^{3/2} + n^2 p_n} + \frac{n^{3/2} p_n + n p_n^{1/2}}{n^3 p_n^2 + n^2 p_n^{3/2} + n^2 p_n}$$

As such, $\hat{\beta}^{(1)}$ is consistent for $\beta^{(1)}$ if $np_n \succ 1$.

Suppose instead that $n^{-2} \prec p_n \prec n^{-1}$. By our rate calculations, we can write

$$\hat{\beta}^{(1)} - \beta^{(1)} = -\beta^{(1)} \cdot \frac{\xi'\xi + o_p(n^2 p_n)}{\xi'\xi + o_p(n^2 p_n)} + o_p(1)$$

In other words, $\hat{\beta}^{(1)} \xrightarrow{p} 0$. Finally, if $p_n \prec n^{-2}$, we $\hat{\beta}^{(1)} - \beta^{(1)}$,

$$\hat{\beta}^{(1)} = \frac{\xi'\varepsilon^{(1)}}{\xi'\xi} + o_p(1) = O_p(n^{-2}p_n^{-1})$$

diverges in probability.

A.4.2.2. Diffusion Centrality. Diffusion centrality is comprised of terms of the form:

$$\hat{A}^t = (A + \boldsymbol{\xi})^t = \sum_{B \in \tilde{\mathbf{B}}} B$$

Here, $\tilde{\mathbf{B}} = \{A, \boldsymbol{\xi}\}^t$. B is a mixed product of A and $\boldsymbol{\xi}$, and will be the central object of our analysis. For convenience, define:

Definition A.2 (Mixed Product of Order t)**.** A mixed product of order t is a term of the form $B = \prod_{j=1}^{t} B_j$ where $B_j \in \{A, \boldsymbol{\xi}\}$. Suppose $B_j = \boldsymbol{\xi}$ for $\tau \geq 0$ number of j's. We will also say that the order of $\boldsymbol{\xi}$ in B is τ. Define $\mathcal{J} = \left\{ j \in [2t+1] \,|\, b_{k_j, k_{j+1}} = \xi_{k_j, k_{j+1}} \right\}$. Then, $\mathcal{J}$ indicate the locations of the $\boldsymbol{\xi}$ in the mixed product B. Let $p = (p_1, ..., p_r)'$ record lengths of the contiguous blocks in $\mathcal{J}$. If $p_1, ..., p_r$ are all even, we say that B is even.

The dependence of $\mathcal{J}$ and p on B is suppressed for convenience.

Example A.3. In the above notation,

$$B = A^2\boldsymbol{\xi}^3 A\boldsymbol{\xi}^2 A \quad \Rightarrow \quad \mathcal{J} = \{(3,4,5)\,,\,(7,8)\,,\,(13,14,15)\,,\,(17,18)\} \quad , \quad p = (3,2,3,2)\,.$$

First note that degree centrality is diffusion centrality with $T = 1$. Since $\hat{\beta}^{(1)}$ is inconsistent for $\beta^{(1)}$ when $np_n \prec 1$, consistency of diffusion centrality also requires that $np_n \succ 1$. We show that $\hat{\beta}^{(T)} \xrightarrow{p} \beta^{(T)}$ when $np_n \succ 1$. Write

$$\hat{C}^{(T)} = \left(\sum_{t=1}^{T} \delta^t \hat{A}^t\right)\iota_n = \left(\sum_{t=1}^{T} \delta^t \left(A + \boldsymbol{\xi}\right)^t\right)\iota_n$$

Expanding the products, we can write $\left(\hat{C}^{(T)}\right)' \hat{C}^{(T)}$ and $\left(\hat{C}^{(T)}\right)' \left(\hat{C}^{(T)} - C^{(T)}\right)$ as sums involving mixed products of A and $\boldsymbol{\xi}$.

We seek to bound $\iota_n' B \iota_n$ in L_2-norm. First note that if $B_j = A$ for all $j \in [t]$, then by Equation (A.4), $B = O_p(n^{t+1}p_n^t)$. Suppose that $B_j = \xi$ for at least one j. Then,

Lemma A.2. Suppose B is a order t mixed product of A and $\boldsymbol{\xi}$. Suppose that the order of $\boldsymbol{\xi}$ in B is $\tau \geq 1$. Then, there exists $\alpha, \beta \in \mathbb{N}$, $\alpha \geq \beta$ such that

$$\iota_n' B \iota_n = O_p\left(n^{t+1-\alpha/2} p_n^{t-\beta/2}\right) \preccurlyeq O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right)\,.$$

In particular,

$$\iota_n' A^t \iota_n = O_p\left(n^{2t+2} p_n^{2t}\right) \succ n^{2t+2-\alpha} p_n^{2t-\beta}\,.$$

Furthermore, suppose B is not even. Then,

$$\iota_n' B \iota_n = O_p\left(\frac{1}{\sqrt{n}}\right) \cdot O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right)\,.$$

If B is even, then

$$\iota_n' B \iota_n - E\left[\iota_n' B \iota_n \,|\, U\right] = O_p\left(\frac{1}{\sqrt{n}}\right) \cdot O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right) \ .$$

Taking expectations over U, we therefore have that

$$\iota_n' A^t \iota_n = O_p(n^{t+1} p_n^t) \succ \iota_n B \iota_n = O_p(n^{t+1-\alpha/2} p_n^{t-\beta/2})$$

under the assumption that $np_n \succ 1$, as long as B contains at least one $\boldsymbol{\xi}$. Now, we return to the nuisance term:

$$\beta^{(T)} \frac{\left(\hat{C}^{(T)}\right)' \left(C^{(T)} - \hat{C}^{(T)}\right)}{\left(\hat{C}^{(T)}\right)' \hat{C}^{(T)}} \ .$$

By our analysis, the dominant term in the denominator is $\iota_n A^{2T} \iota_n = O_p(n^{2T+2} p_n^{2T})$. Every term in the numerator has strictly smaller order. Hence, we conclude that the nuisance term is $o_p(1)$. It remains to show that

$$\frac{\left(\hat{C}^{(T)}\right)' \varepsilon^{(T)}}{\left(\hat{C}^{(T)}\right)' \hat{C}^{(T)}} \approx \frac{\iota_n A^T \varepsilon^{(T)}}{\iota_n A^{2T} \iota_n} \xrightarrow{p} 0 \ .$$

Note that the numerator is a U-statistic of order $T+1$. It also has mean 0 by our conditional mean independence assumption. Hence, by the U-statistic LLN, the numerator is of order $o_p(n^{T+1} p_n^T)$, which is again strictly smaller than that of the denominator. Conclude that $\hat{\beta}^{(T)} \xrightarrow{p} \beta^{(T)}$ if $np_n \succ 1$.

A.4.2.3. Eigenvector Centrality.

Inconsistency. We first provide a counterexample under the assumption that p_n satisfies Equation (1.5). Let $f = 1$, $\beta = 1$ and suppose $\varepsilon_i^{(\infty)} \perp\!\!\!\perp U_i$ (By Assumption 1.2, $\varepsilon_i^{(\infty)} \perp\!\!\!\perp \xi_{jk}$

for all $i, j, k \in [n]$). Theorem 1.7, Remark 1.4 and Remark 1.8 in Alt et al. (2021b) provides the following description of the $v_1(\hat{A})$. Let $i \in [n]$ be a vertex, $B_r(i)$ be the set of vertices which are in the r-neighbourhood of i. Let $S_r(i) = B_r(i) \backslash B_{r-1}(i)$ be the sphere of radius r around i. Let $\mathfrak{u} = \frac{\log n}{np_n \log\log n}$. Then, w.p.a. 1, there exists $\tilde{v}$ such that for any $\eta > 0$,

$$\|\tilde{v} - v_1(\hat{A})\| \leq \frac{1}{\mathfrak{u} \cdot np_n} + \frac{(np_n)^{-1/2+3\eta}}{\sqrt{\mathfrak{u}}} + \frac{1}{np_n} \ . \tag{A.8}$$

Furthermore, $\tilde{v}$ has the following structure:

$$\tilde{v} = \sum_{r=0}^{R} u_r s_r(i) \quad , \quad s_r(i) = \frac{\mathbf{1}_{S_r(i)}}{\|\mathbf{1}_{S_r(i)}\|}$$

where $R \prec \frac{np_n}{\log\log n}$ and

$$u_1 = \frac{1}{\sqrt{np_n}} u_0 \quad , \quad u_r \leq \left(\frac{2}{\sqrt{\mathfrak{u}}}\right)^{r-1} u_1$$

and u_0 is defined by the normalization $\|\tilde{v}\| = 1$. The result of Alt et al. (2021b) says that $v_1(\hat{A})$ is well approximated by an eigenvector that is exponentially localized around some vertex i. This vertex is in fact the one with the highest realized degree. Let us calculate a lower bound on u_0.

$$1 = \tilde{v}'\tilde{v} = \sum_{r=1}^{R} u_r^2 \leq u_0^2 + \frac{1}{np_n} u_0^2 \left(\sum_{r=1}^{\infty} \left(\frac{4}{\mathfrak{u}}\right)^{r-1} \right) \ .$$

The above inequality comes from using upper bounds for u_r and replacing R with ∞. Collecting the u_0's, we find that w.p.a. 1,

$$u_0^2 \geq \frac{1}{1 + \frac{1}{np_n}\frac{1}{1-4/\mathfrak{u}}} \ .$$

Since $np_n, \mathfrak{u} \to \infty$ when p_n satisfies Equation (1.5), we have that for n large enough, $u_0 \geq \frac{1}{\sqrt{2}}$ w.p.a. 1. Now, write

$$\begin{aligned}
\hat{\beta}^{(\infty)} &= \frac{a_n \left(v_1(\hat{A})\right)' Y}{a_n^2 \left(v_1(\hat{A})\right)' v_1(\hat{A})} = \frac{1}{a_n}\left(v_1(\hat{A})\right)' Y \\
&= \frac{1}{a_n}\left(v_1(\hat{A})\right)' \left(a_n \frac{\iota_n}{\sqrt{n}} + \varepsilon^{(\infty)}\right) \\
&= \frac{\left(v_1(\hat{A})\right)' \iota_n}{\sqrt{n}} + \frac{\left(v_1(\hat{A})\right)' \varepsilon^{(\infty)}}{a_n}
\end{aligned}$$

By independence of $\varepsilon^{(\infty)}$ and $(\boldsymbol{\xi}, U)$, we have that $\text{Var}\left[v_1(\hat{A})'\varepsilon^{(\infty)} \,\middle|\, U\right] = \|v_1(\hat{A})\|\sigma^2 = \sigma^2$. Hence, σ^2/a_n is a lower bound for the variance of $\hat{\beta}^{(\infty)}$. Hence, $a_n \to \infty$ is necessary for consistency.

Suppose $a_n \to \infty$. We have consistency if and only if

$$\frac{\left(v_1(\hat{A})\right)' \iota_n}{\sqrt{n}} \xrightarrow{p} 1 \ ,$$

in which case

$$\hat{\beta}^{(\infty)} = \frac{\left(v_1(\hat{A})\right)' \iota_n}{\sqrt{n}} + o_p(1) = \frac{\tilde{v}'\iota_n}{\sqrt{n}} + o_p(1)$$

Notice that the optimization problem:

$$\max_{v\in\mathbb{R}^n} v'\iota_n \quad \text{such that} \quad \|v\| = 1$$

has solution $v = \iota_n/\sqrt{n}$ and optimal value $\sqrt{n}$. We can also consider the constrained optimization problem:

$$\max_{v\in\mathbb{R}^n} v'\iota_n \quad \text{such that} \quad \|v\| = 1 \text{ and } v_1 \geq \frac{1}{\sqrt{2}} \ .$$

This problem has solution $v_1 = \frac{1}{\sqrt{2}}$ and $v_{-1} = \iota_{n-1}/\sqrt{2n}$ and optimal value

$$\gamma^* := \frac{1}{\sqrt{2}} + \frac{n-1}{\sqrt{2n}}$$

The constrained maximization problem corresponds to the best case allocation of $\left(v_1(\hat{A})\right)_{-i}$ that makes $\left(v_1(\hat{A})\right)'\iota$ as close to $\sqrt{n}$ as possible, subject to the requirement that $\left(v_1(\hat{A})\right)_i \geq 1/\sqrt{2}$. As such, w.p.a. 1, we have that

$$\hat{\beta}^{(\infty)} \leq \frac{\gamma^*}{\sqrt{n}} = \frac{1}{\sqrt{2n}} + \frac{n-1}{n\sqrt{2}} \to \frac{1}{\sqrt{2}} \ .$$

Hence, $\hat{\beta}^{(\infty)}$ is bounded away from $\beta^{(\infty)} = 1$ in probability. Conclude that the estimator is inconsistent.

Consistency. We next show that $\hat{\beta}^{(\infty)} \xrightarrow{p} \beta^{(\infty)}$ when $np_n \succ \sqrt{\frac{\log n}{\log\log n}}$. Write

$$\hat{\beta}^{(\infty)} = \beta^{(\infty)} + \beta^{(\infty)} \frac{\left(\hat{C}^{(\infty)}\right)'\left(C^{(\infty)} - \hat{C}^{(\infty)}\right)}{\left(\hat{C}^{(\infty)}\right)'\hat{C}^{(\infty)}} + \frac{\left(\hat{C}^{(\infty)}\right)'\varepsilon^{(\infty)}}{\left(\hat{C}^{(\infty)}\right)'\hat{C}^{(\infty)}} \tag{A.9}$$

$$= \beta^{(\infty)} + \beta^{(\infty)}\left(v_1(\hat{A})\right)'\left(v_1(A) - v_1(\hat{A})\right) + \frac{1}{a_n} v_1(\hat{A})'\varepsilon^{(\infty)} \tag{A.10}$$

since $\left(\hat{C}^{(\infty)}\right)' \hat{C}^{(\infty)} = a_n^2$ by construction. Therefore, by Lemma A.1 and the Davis-Kahan inequality (e.g. Theorem 4.5.5 in Vershynin 2018), for any $\nu \in (0,1)$,

$$\left\| v_1(A) - v_1(\hat{A}) \right\| \leq \frac{\|\hat{A} - A\|}{np_n(\lambda_1 - \lambda_2)} = O_p\left(\left(\sqrt{\frac{\log n}{\log\log n}} \Big/ np_n \right)^{(1-\nu)/2} \right) = o_p(1)$$

where first equality follows from Equation (A.1) and the second from our assumption on np_n.

Finally, note that

$$E\left[v_1(\hat{A})' \varepsilon^{(\infty)} \,\middle|\, U \right] \leq \|v_1(\hat{A})\| \bar{\sigma}^2 = \bar{\sigma}^2.$$

Since $a_n \to \infty$, conclude that $\frac{1}{a_n} v_1(\hat{A})' \varepsilon^{(\infty)} \xrightarrow{p} 0$ and that $\hat{\beta}^{(\infty)} \xrightarrow{p} \beta^{(\infty)}$.

A.4.3. Proof of Theorem 1.5

A.4.3.1. Case (a). Although case (b) specializes to (a), we will prove (a) separately because

(1) The proof for our plug-in estimator for case (a) is also the base case for an induction argument in the proof of case (b)

(2) Case (c), by Lemma A.3 equivalent to case (a) to a first order.

To prove (a) first recall our analysis in the proof of Theorem 1.2, which yields:

$$\begin{aligned} \hat{\beta}^{(1)} &= \beta^{(1)} + \beta^{(1)} \frac{\left(\hat{C}^{(1)}\right)' \left(C^{(1)} - \hat{C}^{(1)}\right)}{\left(\hat{C}^{(1)}\right)' \hat{C}^{(1)}} + \frac{\left(\hat{C}^{(1)}\right)' \varepsilon^{(\infty)}}{\left(\hat{C}^{(1)}\right)' \hat{C}^{(1)}} \\ &= \beta^{(1)} + \beta^{(1)} \frac{\iota_n' A \boldsymbol{\xi} \iota_n + \iota_n' \boldsymbol{\xi}^2 \iota_n}{\iota_n A^2 \iota_n + o_p\left(\iota_n A^2 \iota_n'\right)} + \frac{O_p\left(n^{3/2} p_n\right)}{\iota_n A^2 \iota_n + o_p\left(\iota_n' A^2 \iota_n\right)} \,. \end{aligned} \tag{A.11}$$

Recall also that,

$$E\left[\iota_n'\boldsymbol{\xi}\iota_n \,|\, U\right] = \sum_{i=1}^{n}\sum_{j\neq i} E\left[\xi_{ij}^2 \,|\, U\right] = \sum_{i=1}^{n}\sum_{j\neq i} p_n f(U_i, U_j)\left(1 - p_n f(U_i, U_j)\right)$$

so that the unconditional expectation is

$$E\left[\iota_n'\boldsymbol{\xi}\iota_n\right] = \Omega\left(n^2 p_n\right) \ .$$

To obtain our desired result, we will show that $\iota_n' A\boldsymbol{\xi}\iota_n$ converges to a normal distribution asymptotically once suitable scaled, and that it dominates $\left(\iota_n'\boldsymbol{\xi}^2\iota_n - E\left[\iota_n'\boldsymbol{\xi}\iota_n \,|\, U\right]\right)$. We then show that the population quantities in the CLT can be estimated at a sufficiently fast rate.

First observe that by Assumption 1.2, $E\left[\iota_n' A\boldsymbol{\xi}\iota_n \,|\, U\right] = 0$. Next, define

$$\begin{aligned} V_*^{(1)}(U) &:= E\left[\left(\iota_n' A\boldsymbol{\xi}\iota_n\right)^2 \,\middle|\, U\right] \\ &= \sum_{j<k} p_n f(U_j, U_k)\left(1 - p_n f(U_j, U_k)\right)\left(\sum_{i\neq j} p_n f(U_i, U_j) + \sum_{i\neq k} p_n f(U_i, U_k)\right)^2 \ . \end{aligned}$$

Then, by the U-statistics LLN,

$$\frac{1}{n^4 p_n^3} V_*^{(1)}(U) \xrightarrow{p} \int f(U_1, U_2) f(U_1, U_3) f(U_1, U_4)\, dU + \frac{1}{2}\int f(U_1, U_2) f(U_2, U_3) f(U_3, U_4)\, dU > 0 \, .$$

Next, define the event $\Upsilon^{(1)} := \left\{V^{(1)}(U) > k n^4 p_n^3\right\}$, where k is chosen to be $1/2$ the magnitude of the limit above. We will apply the Berry-Esseen inequality conditional on

$U \in \Upsilon^{(1)}$. Note that by Assumption 1.2, ξ_{ij}'s continue to be independent after conditioning. By Theorem 3.7 in Chen et al. (2011),

$$\sup_{z\in\mathbf{R}}\left|P\left(\frac{\iota_n' A\boldsymbol{\xi}\iota_n}{\sqrt{V_*^{(1)}(U)}}\leq z\,\middle|\,U\right)-\Phi(z)\right|\leq 10\gamma\ .$$

Next, we evaluate the third moments of the summands:

$$\begin{aligned}E\left[\left|\xi_{jk}\sum_{i\neq j}p_n f(U_i,U_j)\right|^3\,\middle|\,U\right]&=\left|\sum_{i\neq j}p_n f(U_i,U_j)\right|^3 E\left[\left|\xi_{jk}\right|^3\,\middle|\,U\right]\\&\leq n^3p_n^3\cdot p_n\end{aligned}$$

As such, on $\Upsilon^{(1)}$,

$$\gamma\leq\sum_{j<k}\frac{n^3p_n^4}{\left(kn^4p_n^3\right)^{3/2}}\approx\frac{n^5p_n^4}{n^6p_n^{9/2}}=\frac{1}{np^{1/2}}\to 0$$

where the above bound is independent of U. Furthermore, $P(\Upsilon^{(1)})\to 1$. Conclude that

$$\frac{\iota_n' A\boldsymbol{\xi}\iota_n}{\sqrt{V_*^{(1)}(U)}}\xrightarrow{d}\mathrm{N}\left(0,1\right)\ .$$

It remains to show that $\iota_n'\boldsymbol{\xi}^2\iota_n-E\left[\iota_n'\boldsymbol{\xi}^2\iota_n\,|\,U\right]=o_p\left(n^2p_n^{3/2}\right)$. Write

$$\Gamma^{(1)}:=\sum_{i_1,\ldots,i_6}E\left[\left(\xi_{i_1,i_2}\xi_{i_2,i_3}-E\left[\xi_{i_1,i_2}\xi_{i_2,i_3}\,|\,U\right]\right)\left(\xi_{i_4,i_5}\xi_{i_5,i_6}-E\left[\xi_{i_4,i_5}\xi_{i_5,i_6}\,|\,U\right]\right)\,\middle|\,U\right]\ .$$

Note that

$$E\left[\xi_{i_1i_2}\xi_{i_2i_3}\xi_{i_4i_5}\xi_{i_5i_6}\,\middle|\,U\right]=0\Rightarrow E\left[\left(\xi_{i_1,i_2}\xi_{i_2,i_3}-E\left[\xi_{i_1,i_2}\xi_{i_2,i_3}\,|\,U\right]\right)\left(\xi_{i_4,i_5}\xi_{i_5,i_6}-E\left[\xi_{i_4,i_5}\xi_{i_5,i_6}\,|\,U\right]\right)\,\middle|\,U\right]=0\,.$$

This is because for the former to hold, we must have an edge (i_k, i_{k+1}) that is of multiplicity 1, which is sufficient for making the latter conditional expectation 0. Figure A.1 shows all possible configurations of indices that will lead to $E\left[\xi_{i_1 i_2}\xi_{i_2 i_3}\xi_{i_4 i_5}\xi_{i_5 i_6} \,\middle|\, U\right] \neq 0$. Table A.3 records the frequency of their appearance.

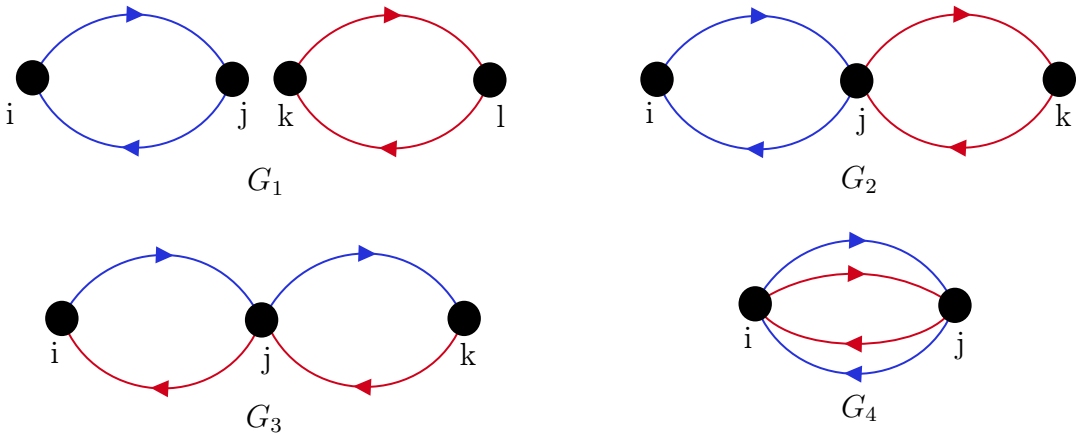

Figure A.1. The possible configurations of indices that will lead to $E\left[\xi_{ij}\xi_{jk}\xi_{i'j'}\xi_{j'k'} \,\middle|\, U\right]$ being non-zero. These are the only graphs that can be formed using 2 walks of length 2 and in which each edge has multiplicity at least 2.

Graph	Number of Instances	$E\left[\xi_{ij}\xi_{jk}\xi_{i'j'}\xi_{j'k'} \,\middle|\, U\right]$
G_1	$n(n-1)(n-2)(n-3)$	$p_n^2 f(U_i, U_j) f(U_k, U_l) + O_p(p_n^3)$
G_2	$n(n-1)(n-2)$	$p_n^2 f(U_i, U_j) f(U_k, U_l) + O_p(p_n^3)$
G_3	$n(n-1)(n-2)$	$p_n^2 f(U_i, U_j) f(U_k, U_l) + O_p(p_n^3)$
G_4	$n(n-1)$	$p_n f(U_i, U_j) + O_p(p_n^2)$

Table A.3. The number of instances of each graph, as well as the value of their conditional expectations, up to the leading term.

Observe that

$$E\left[\left(\xi_{i_1,i_2}\xi_{i_2,i_3} - E\left[\xi_{i_1,i_2}\xi_{i_2,i_3}\,|\,U\right]\right)\left(\xi_{i_4,i_5}\xi_{i_5,i_6} - E\left[\xi_{i_4,i_5}\xi_{i_5,i_6}\,|\,U\right]\right)\,\middle|\,U\right]$$
$$\neq E\left[\xi_{i_1,i_2}\xi_{i_2,i_3} - E\left[\xi_{i_1,i_2}\xi_{i_2,i_3}\,|\,U\right]\,\middle|\,U\right] E\left[\xi_{i_4,i_5}\xi_{i_5,i_6} - E\left[\xi_{i_4,i_5}\xi_{i_5,i_6}\,|\,U\right]\,\middle|\,U\right]$$

only if there is an edge that is common to both of the above multiplicands. In particular, G_1 and G_2 will not contribute to $\Gamma^{(1)}$. As such, by Table A.3, $\Gamma^{(1)} = O_p\left(n^3 p_n^2\right)$. Conclude that

$$\iota_n'\boldsymbol{\xi}^2\iota_n - E\left[\iota_n'\boldsymbol{\xi}^2\iota_n\,|\,U\right] = O_p\left(n^{3/2}p_n\right) = o_p\left(n^2 p_n^{3/2}\right)\ .$$

Using the above results, we can rewrite Equation (A.11) as

$$\hat{\beta}^{(1)} = \beta^{(1)} + \beta^{(1)}\frac{\iota_n' A\boldsymbol{\xi}\iota_n - E\left[\iota_n'\boldsymbol{\xi}^2\iota_n\,|\,U\right] + O_p\left(n^{3/2}p_n\right)}{\left(\hat{C}^{(1)}\right)\hat{C}^{(1)}}\ .$$

Consequently,

$$\frac{\hat{\beta}^{(1)} - \beta^{(1)}\left(1 - B^{(1)}\right)}{\beta^{(1)}\sqrt{V^{(1)}}} = \frac{\iota_n' A\boldsymbol{\xi}\iota_n}{\sqrt{V_*^{(1)}}} + \frac{O_p\left(n^{3/2}p_n\right)}{\Omega_p\left(n^2 p_n^{3/2}\right)} \xrightarrow{d} \mathrm{N}(0,1)\ .$$

where

$$B^{(1)} = \left(\left(\hat{C}^{(1)}\right)\hat{C}^{(1)}\right)^{-1} E\left[\iota_n'\boldsymbol{\xi}^2\iota_n\,|\,U\right]\ ,$$
$$V^{(1)} = \left(\left(\hat{C}^{(1)}\right)\hat{C}^{(1)}\right)^{-2} V_*^{(1)}\ .$$

Plug-in Estimation. Finally, we show that $\hat{B}^{(1)}$ and $\hat{V}^{(1)}$ estimate $B^{(1)}$ and $V^{(1)}$ at appropriate rates. Define $\hat{V}_*^{(1)} = \left(\left(\hat{C}^{(1)}\right)\hat{C}^{(1)}\right)^{-2}\hat{V}^{(1)}$. We will show that

$$\frac{\hat{B}^{(1)} - B^{(1)}}{\sqrt{V^{(1)}}} \xrightarrow{p} 0 \quad , \quad \frac{\hat{V}^{(1)}}{V^{(1)}} = \frac{\hat{V}_*^{(1)}}{V_*^{(1)}} \xrightarrow{p} 1 \ .$$

The first statement above is straightforward:

$$\begin{aligned}\frac{\hat{B}^{(1)} - B^{(1)}}{\sqrt{V^{(1)}}} &= \frac{1}{\sqrt{V_*^{(1)}}}\sum_{i\neq j} p_n f(U_i,U_j) + \xi_{ij} - p_n f(U_i,U_j)\left(1 - p_n f(U_i,U_j)\right) \\ &= O_p\left(\frac{1}{n^2 p_n^{3/2}}\right)\cdot\Bigg(\underbrace{\sum_{i\neq j}\xi_{ij}}_{\substack{=O_p(np_n^{1/2}) \\ \text{by ((A.5))}}} + \underbrace{\sum_{i\neq j} p_n^2 f^2(U_i,U_j)}_{=O_p(n^2p_n^2)}\Bigg) = O_p\left(\frac{1}{np_n} + p_n^{1/2}\right) = o_p(1) \ .\end{aligned}$$

Next, consider:

$$
\begin{aligned}
&2\left(\hat{V}_*^{(1)} - V_*^{(1)}\right) \\
&= \sum_{j\neq k} \hat{A}_{jk}\left(\hat{C}_j^{(1)} + \hat{C}_k^{(1)}\right)^2 - p_n f(U_j,U_k)\left(1 - p_n f(U_j,U_k)\right)\left(\sum_{i\neq j} p_n f(U_i,U_j) + \sum_{i\neq k} p_n f(U_i,U_k)\right)^2 \\
&= \sum_{j\neq k} \hat{A}_{jk}\left(\hat{C}_j^{(1)} + \hat{C}_k^{(1)}\right)^2 - p_n f(U_j,U_k)\left(\sum_{i\neq j} p_n f(U_i,U_j) + \sum_{i\neq k} p_n f(U_i,U_k)\right)^2 + O_p\left(n^4 p_n^4\right) \\
&= \underbrace{2\sum_{j\neq k} A_{jk}\left(C_j^{(1)} + C_k^{(1)}\right)(\xi_j + \xi_k)}_{=:\Gamma_1^{(1)}} + \underbrace{\sum_{j\neq k} A_{jk}\left(\xi_j + \xi_k\right)^2}_{=:\Gamma_2^{(1)}} \\
&\quad + \underbrace{\sum_{j\neq k} \xi_{jk}\left(C_j^{(1)} + C_k^{(1)}\right)^2}_{=:\Gamma_3^{(1)}} + \underbrace{2\sum_{j\neq k} \xi_{jk}\left(C_j^{(1)} + C_k^{(1)}\right)(\xi_j + \xi_k)}_{=:\Gamma_4^{(1)}} + \underbrace{\sum_{j\neq k} \xi_{jk}\left(\xi_j + \xi_k\right)^2}_{=:\Gamma_5^{(1)}}
\end{aligned}
$$

Recall that $V_*^1 = \Omega_p\left(n^4 p_n^3\right)$. We will show that $\Gamma_a^{(1)} = o_p\left(n^4 p_n^3\right)$ for $a \in [5]$.

$$
\begin{aligned}
\Gamma_1^{(1)} &= \sum_{j,k} A_{jk}\left(C_j^{(1)} + C_k^{(1)}\right)\sum_{i\neq j}\xi_{ij} + \sum_{j,k} A_{jk}\left(C_j^{(1)} + C_k^{(1)}\right)\sum_{i\neq k}\xi_{ik} \\
&= 2\sum_{i,j}\xi_{ij}\sum_k A_{jk}\left(C_j^{(1)} + C_k^{(1)}\right) \qquad \text{by symmetry}
\end{aligned}
$$

Taking conditional expectations,

$$\begin{aligned} E\left[\Gamma_1^{(1)} \,|\, U\right] &= E\left[\left(4\sum_{i<j}\xi_{ij}\sum_k A_{jk}\left(C_j^{(1)}+C_k^{(1)}\right)\right)^2 \,\Big|\, U\right] \\ &= 16\sum_{i<j} E[\xi_{ij}^2 \,|\, U]\left(\sum_k A_{jk}\left(C_j^{(1)}+C_k^{(1)}\right)\right)^2 \\ &\leq 16n^2p_n\cdot(2np_n)^4 \quad \text{since } C_j^{(1)} \leq np_n \text{ for all } j\in[n] \\ &= O_p\left(n^6p_n^5\right) \end{aligned}$$

Hence, $\Gamma_1^{(1)} = O_p\left(n^3p_n^{5/3}\right)$.

Next,

$$\Gamma_2^{(1)} = 2\sum_{j,k}A_{jk}\left(\sum_{i\neq j}\xi_{ij}\right)^2 + \sum_{j,k}A_{jk}\left(\sum_{i\neq j}\xi_{ij}\right)\left(\sum_{i\neq k}\xi_{ik}\right) \ .$$

First note that

$$\sum_{j,k}A_{jk}\left(\sum_{i\neq j}\xi_{ij}\right)\left(\sum_{i\neq k}\xi_{ik}\right) = \iota_n'\boldsymbol{\xi}A\boldsymbol{\xi}\iota_n = O_p\left(n^{7/2}p_n^{5/2}\right) \quad \text{by Lemma A.2.}$$

Secondly, we have that

$$\begin{aligned} E\left[\left(\sum_{j,k}A_{jk}\left(\sum_{i\neq j}\xi_{ij}\right)^2\right)^2 \,\Big|\, U\right] &= E\left[\left(\sum_{i,j,l}\xi_{ij}\xi_{jl}\sum_k A_{jk}\right)^2 \,\Big|\, U\right] \\ &\leq n^2p_n^2E\left[\left(\iota_n'\boldsymbol{\xi}^2\iota_n\right)^2 \,|\, U\right] \leq n^2p_n^2\cdot n^4p_n^2 \end{aligned}$$

Noting that the bound above does not depend on U, we have

$$\Gamma_2^{(1)} = O_p\left(n^3p_n^2\right) + O_p\left(n^{7/2}p_n^{5/2}\right) \ .$$

Now,

$$E\left[\left(\Gamma_3^{(1)}\right)^2 \,|\, U\right] = \sum_{j,k} E\left[\xi_{jk}^2 \,|\, U\right]\left(C_j^{(1)} + C_k^{(1)}\right)^4 \leq n^2 p_n \cdot (2np_n)^4$$

As such, $\Gamma_3^{(1)} = O_p\left(n^3 p_n^{5/2}\right)$.

By a similar argument to above, we also have that

$$\Gamma_4^{(1)} = 2\sum_{j,k,l} \xi_{jk}\xi_{jl}\left(C_j^{(1)} + C_k^{(1)}\right) = O_p\left(np_n\right) \cdot O_p\left(\iota_n \boldsymbol{\xi}^2 \iota_n\right) = O_p\left(n^3 p_n^2\right) \ .$$

Finally,

$$\Gamma_5^{(1)} = 2\sum_{j,k} \xi_{jk}\left(\sum_{i\neq j} \xi_{ij}\right)^2 + \sum_{j,k} \xi_{jk} \sum_{i\neq j} \xi_{ij} \sum_{l\neq k} \xi_{kl}$$

First observe that

$$\sum_{j,k} \xi_{jk} \sum_{i\neq j} \xi_{ij} \sum_{l\neq k} \xi_{kl} = \iota_n' \boldsymbol{\xi}^3 \iota_n = O_p\left(n^2 p_n^{3/2}\right) \qquad \text{by Lemma A.2.}$$

Now,

$$E\left[\left(\sum_{j,k} \xi_{jk}\left(\sum_{i\neq j} \xi_{ij}\right)^2\right)^2 \,|\, U\right] = \sum_{i_1,\ldots i_8} E\left[\xi_{i_1 i_2}\xi_{i_1 i_3}\xi_{i_1 i_4} \cdot \xi_{i_5 i_6}\xi_{i_5 i_7}\xi_{i_5 i_8} \,|\, U\right]$$

Relative to Lemma 2, here we are counting the contributions made by two three-pointed stars. The graphs that contribute the above expectation are displayed in Figure A.2. Their frequencies and magnitudes are recorded in Table A.4. Summing up the contribution of each graph, we have that the above display is $O_p\left(n^4 p_n^2\right)$. Hence, $\Gamma_5^{(1)} = O_p\left(n^2 p_n\right) + O_p\left(n^2 p_n^{3/2}\right) = O_p\left(n^2 p_n\right)$.

Putting all our results together, we have that

$$\frac{\hat{V}_*^1 - V_*^{(1)}}{V_*^{(1)}} = o_p(1) \ ,$$

which together with our central limit therom and result on $\hat{B}^{(1)}$ implies our desired result:

$$\hat{S}^{(1)} := \frac{\hat{\beta}^{(1)} - \beta^{(1)}\left(1 - \hat{B}^{(1)}\right)}{\beta^{(1)}\sqrt{\hat{V}^{(1)}}} \xrightarrow{d} \mathrm{N}(0,1) \ .$$

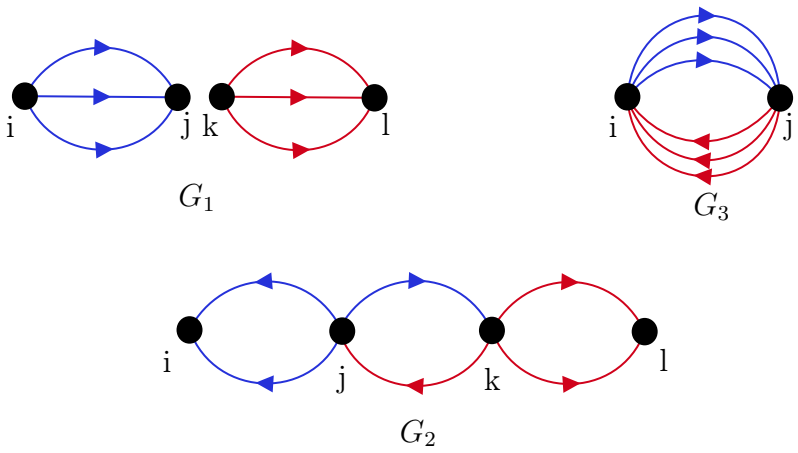

Figure A.2. The possible configurations of indices that will lead to $E\left[\xi_{i_1 i_2}\xi_{i_1 i_3}\xi_{i_1 i_4} \cdot \xi_{i_5 i_6}\xi_{i_5 i_7}\xi_{i_5 i_8} \,|\, U\right]$ being non-zero. These are the only graphs that can be formed using 2 3-pointed stars and in which each edge has multiplicity at least 2.

A.4.3.2. Case (b). As with case (a), our strategy is to remove the bias coming from $\boldsymbol{\xi}^2$ and to obtain a central limit theorem on the leading term of the remainder. Write

$$\hat{\beta}^{(T)} = \beta^{(T)} + \beta^{(T)}\frac{\left(\hat{C}^{(T)}\right)'\left(C^{(T)} - \hat{C}^{(T)}\right)}{\left(\hat{C}^{(T)}\right)'\hat{C}^{(T)}} + \frac{\left(\hat{C}^{(T)}\right)'\varepsilon^{(T)}}{\left(\hat{C}^{(T)}\right)'\hat{C}^{(T)}} \ . \tag{A.12}$$

Graph	Number of Instances	$E\left[\xi_{ij}\xi_{jk}\xi_{i'j'}\xi_{j'k'} \,\middle|\, U\right]$
G_1	$n(n-1)(n-2)(n-3)$	$p_n^2 f(U_i, U_j) f(U_k, U_l) + O_p(p_n^3)$
G_2	$n(n-1)(n-2)(n-3)$	$p_n^3 f(U_i, U_j) f(U_j, U_k) f(U_k, U_l) + O_p(p_n^4)$
G_3	$n(n-1)$	$p_n f(U_i, U_j) + O_p(p_n^2)$

Table A.4. The number of instances of each graph, as well as the value of their conditional expectations, up to the leading term. Note that we can consider G_1 with $j = k$, though the contribution of this term is strictly smaller than the contribution of G_1.

As before, $\left(\hat{C}^{(T)}\right)' \left(C^{(T)} - \hat{C}^{(T)}\right)$ comprises mixed products of A and $\boldsymbol{\xi}$, whose order with respect to $\boldsymbol{\xi}$ is at least 1. Let B be such a term. By Lemma A.2, if B is even,

$$\iota_n' B \iota_n - E\left[\iota_n' B \iota_n \,|\, U\right] = O_p\left(\frac{n^{t+1} p_n^t}{\sqrt{n}\left(\sqrt{n p_n}\right)^\tau}\right) .$$

Otherwise,

$$\iota_n' B \iota_n = O_p\left(\frac{n^{t+1} p_n^t}{\sqrt{n}\left(\sqrt{n p_n}\right)^\tau}\right) .$$

In other words, once the even terms are centered, the dominant terms in $\left(\hat{C}^{(T)}\right)' \left(C^{(T)} - \hat{C}^{(T)}\right)$ are of order $2T$ overall, and have order 1 with respect to $\boldsymbol{\xi}$. Such terms are dominant provided that they attain the stated upper bounds. There are T of these, taking the form

below:

$$
\begin{aligned}
&\iota_n' \sum_{t=1}^{T} A^{T+t-1} \boldsymbol{\xi} A^{T-t} \iota_n \\
&= \sum_{j,k} \xi_{jk} \sum_{t=1}^{T} \sum_{i} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_{T+t-1},j} \cdot A_{k,i_{T+t+2}} \cdots A_{i_{2T}.i_{2T+1}} \\
&= \sum_{j<k} \xi_{jk} \sum_{t=1}^{T} \Bigg(\sum_{i} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_{T+t-1},j} \cdot A_{k,i_{T+t+2}} \cdots A_{i_{2T}.i_{2T+1}} \\
&\qquad\qquad + \sum_{i} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_{T+t-1},k} \cdot A_{j,i_{T-t+2}} \cdots A_{i_{2T}.i_{2T+1}} \Bigg) \\
&= \sum_{j<k} \xi_{jk} \sum_{t=1}^{T} \Bigg(\sum_{i} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_{T+t-1},j} \cdot A_{k,i_{T+t+2}} \cdots A_{i_{2T}.i_{2T+1}} \\
&\qquad\qquad + \sum_{i} A_{i_{2T+1},i_{2T}} A_{i_{2T},i_{2T-1}} \cdots A_{i_{T+t+2},k} \cdot A_{j,i_{T+t-1}} \cdots A_{i_2.i_1} \Bigg) \text{ by symmetry} \\
&= \sum_{j<k} \xi_{jk} \sum_{t=1}^{2T} \left(\sum_{i} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_{t-1},j} \cdot A_{k,i_{t+2}} \cdots A_{i_{2T}.i_{2T+1}} \right) \text{ by change of index.}
\end{aligned}
$$

In the above display, summation over i is understood to exclude i_{T+t} and i_{T+t+1}, which have been replaced by j and k. Now define

$$
\begin{aligned}
V_*^{(T)}(U) &:= \delta^{4T} E \left[\left(\iota_n' \sum_{t=1}^{T} A^{T+t-1} \boldsymbol{\xi} A^{T-t} \iota_n \right)^2 \,\middle|\, U \right] \\
&= \frac{1}{2} \delta^{4T} \sum_{j,k} A_{jk}(1 - A_{jk}) \left(\sum_{t=1}^{2T} \sum_{i} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_{t-1},j} \cdot A_{k,i_{t+2}} \cdots A_{i_{2T}.i_{2T+1}} \right)^2 .
\end{aligned}
$$

We can get an intuition for the above term by considering binary A, in which case the variance counts the number of ways two paths of length $2T+1$ have at least one overlapping edge. The archetypal motif is displayed in Figure A.3.

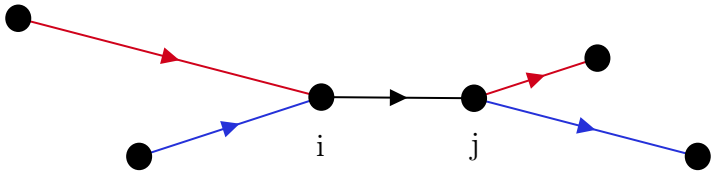

Figure A.3. When A is binary, $V_*^{(T)}(U)$ counts motifs like the one displayed. Here, the red path and the blue path have the same length of $2T+1$ and overlap on the edge (i,j).

By the U-statistic LLN,

$$\frac{1}{n^{4T}p_n^{4T-1}}V_*^{(T)} \xrightarrow{p} \delta^{4T}\sum_{t=1}^{2T}\sum_{s=1}^{2T}\frac{1}{2}\int f(U_1,U_2)\cdot[f(U_3,U_4)\cdots f(U_{t+1},U_1)]\cdot[f(U_2,U_{t+2})\cdots f(U_{2T},U_{2T+1})]$$

(A.13)

$$\cdot[f(U_{2T+3},U_{2T+4})\cdots f(U_{2T+1+s},U_1)]$$

$$\cdot[f(U_2,U_{2T+2+s})\cdots f(U_{4T-s-1},U_{4T-s})]\ dU\ .$$

Notice that

$$\sqrt{n^{4T}p_n^{4T-1}} = \frac{n^{2T+1}p_n^{2T}}{\sqrt{n}\sqrt{np_n}}\ ,$$

so that our conjectured leading term in fact strictly dominates all the other terms. Now, let $\mathbf{B}$ be the set of even mixed products in $\left(\hat{C}^{(T)}\right)'\left(C^{(T)}-\hat{C}^{(T)}\right)$. Furthermore, define

$$B^{(T)} = \left(\left(\hat{C}^{(T)}\right)'\hat{C}^{(T)}\right)^{-1}\sum_{B\in\mathbf{B}}E\left[\iota_n'B\iota_n\,|\,U\right]$$

$$V^{(T)} = \left(\left(\hat{C}^{(T)}\right)'\hat{C}^{(T)}\right)^{-2}V_*^{(T)}\ .$$

Then,

$$\frac{\left(\hat{\beta}^{(T)} - \beta^{(T)} - B^{(T)}\right)}{\beta^{(T)}\sqrt{V^{(T)}}} = \frac{\left(\hat{C}^{(T)}\right)'\left(C^{(T)} - \hat{C}^{(T)}\right) - \sum_{B\in\mathbf{B}} E\left[\iota_n' B \iota_n \,|\, U\right]}{\sqrt{V_*^{(T)}}}$$
$$= \frac{\iota_n' \sum_{t=1}^{T} A^{T+t-1} \boldsymbol{\xi} A^{T-t} \iota_n}{\sqrt{V_*^{(T)}}} + o_p(1) \ .$$

The proof then proceeds similarly to Case (a). Define the event $\Upsilon^{(T)} := \left\{V_*^{(T)} > k n^{4T} p_n^{4T-1}\right\}$, where k is chosen to be $1/2$ the magnitude of the limit in Equation (A.13). Applying the Berry-Eseen inequality of Chen et al. (2011) conditioning on $U \in \Upsilon^{(T)}$ and noting that $P\left(\Upsilon^{(T)}\right) \to 1$, we obtain:

$$\frac{\iota_n' \sum_{t=1}^{T} A^{T+t-1} \boldsymbol{\xi} A^{T-t} \iota_n}{\sqrt{V_*^{(T)}}} \xrightarrow{d} \mathrm{N}(0,1)$$

which yields the desired result.

Plug-in Estimation. We need to estimate the bias for $B \in \mathbf{B}$, as well as $V_*^{(T)}$. We estimate $V_*^{(T)}$, as in case (a), we replace A_{jk} with $\hat{A}_{jk}$, and $(1 - A_{jk})$ with 1. The proof is largely similar to that in Case (a). It is tedious but straightforward since there are no rate requirements on the estimation of $V_*^{(T)}$. We discuss them in turn.

The main challenge in inference for $\beta^{(T)}$ arises because the standard deviation of $\hat{\beta}^{(d)}$ is larger than it's bias. In order for the resulting de-biased inference method to improve mean square error, bias estimation must occur at a sufficiently fast rate.

Our strategy is as follows. Let $B \in \mathbf{B}$ be order t and have order τ with respect to $\boldsymbol{\xi}$. It's block structure is described by $p = (p_1, ..., p_r)$, where each component is even. We

first claim is that there exists a function $\tilde{\gamma}(t, A)$ taking the form:

$$\tilde{\gamma}(t, A) = \tilde{\gamma}_1(t) \cdot A + \tilde{\gamma}_2(t) \cdot A^2 + \cdots + \tilde{\gamma}_{t-1}(t) \cdot A^{t-1}$$

such that

$$E\left[\iota_n' B \iota_n \,|\, U\right] - \iota_n' \left(A^{t-\tau} \prod_{j=1}^{r} \tilde{\gamma}(p_j, A) \right) \iota_n = O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right) \ .$$

In words, the bias of B is "close to" $\tilde{\gamma}$, a polynomial of the unobserved adjacency matrix A. Then provided that we have good estimators of $\iota_n A^t \iota_n$, we will be able to estimate $E\left[\iota_n' B \iota_n \,|\, U\right]$ by substituting them into $\tilde{\gamma}$.

To obtain the $\tilde{\gamma}(t, A)$, write:

$$E\left[\iota_n' B \iota_n \,|\, U\right] = \sum_{i_1, \ldots i_{t+1}} E\left[B_{i_1 i_2} \cdots B_{i_t i_{t+1}} \,|\, U\right] \ . \tag{A.14}$$

We are interested in graphs induced by relationships on the nodes $[n]$ that will lead to non-zero contributions to the above sum. Only nodes corresponding to $\boldsymbol{\xi}$ matters, i.e. i_j s.t $j \in \mathcal{J} \cup (\mathcal{J} + 1)$. Hence, we are interested in graphs formed by overlaying r walks, each of length $p_1, ..., p_r$. For a given graph G, write its order as

$$\sum_{i \in r_G} E\left[B_{i_1 i_2} \cdots B_{i_t i_{t+1}} \,|\, U\right] = O_p(n^{\alpha+1} p_n^{\beta}).$$

where $\alpha, \beta \in \mathbb{N}$ and $\alpha \geq \beta$. By Lemma 2, we know that

$$\iota_n' B \iota_n = O_p\left(n^{t+1-\alpha/2} p_n^{t-\beta/2}\right) \preccurlyeq O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right) \ .$$

Since $\tau \geq 2$, any graph for which $\alpha > \beta$ satisfies our criteria without bias correction. To achieve the $\frac{1}{\sqrt{n}}$ term in (A.4.3.2), we only need to deal with graphs for which $\alpha = \beta$. We called these the best case graphs in proof of 1.2, and they have the same characteristics as before. Namely, every edge must have multiplicity 2 greater than 2, and each edge must involve only one walk, since requiring an edge to be traversed by more than one walk increases α but not β. Thus, it is sufficient to consider the walks separately.

For a given p_j, we need to characterize walks for which $\alpha = \beta$. When $\alpha = \beta$, the order on p_n (i.e. the number of unique edges) is exactly 1 less than the order of n (i.e. the number of nodes). That is, the only graphs that matter are paths.

Let G be a walk with length p_j. Suppose that after removing duplicate edges, G is a path of length s. Then for deterministic vectors x and y,

$$\sum_{i \in r_G} x_{i_1} E\left[\xi_{i_1 i_2} \cdots \xi_{i_t i_{t+1}} \,|\, U\right] y_{i_{s+1}} = (1 + o_p(1)) \cdot \sum_{i} x_{i_1} A_{i_1 i_2} \cdots A_{i_s i_{s+1}} y_{i_{s+1}} \ .$$

The indices on the right hand side are unrestricted. The above assertion arises by the following injective mapping from $r_G \to [n]^s$. By definition, G is a walk of length p_j which traverses $s+1$ unique nodes. Let $j_1, ..., j_s, j_{s+1}$ be the steps at which G reaches a new unique node. Then our injective map is $i \mapsto (i_{j_1}, ..., i_{j_s+1})$.

Example A.4. Consider the walk $i_1 \to i_2 \to i_1 \to i_3$, where all the nodes are distinct. Then $j = (1, 2, 4)$. Suppose $i = (5, 10, 5, 4)$. Then $i \mapsto (5, 10, 4)$.

There is a small error term arising from two sources. Firstly, we are only capturing the first order term of the $E\left[\xi_{i_1 i_2} \cdots \xi_{i_t i_{t+1}} \,|\, U\right]$. There are higher order terms whose magnitude are at most of p_n times this term that we omit. Second, there are paths on

the right hand side which is not in the range of our injective map. These are in turn i's on which a given node appears more than once. There cannot be more than n^{s-1} such paths. Hence, these paths are at most $O_p\left(\frac{1}{n}\right)$ of the right hand side term.

Noting that $\mathbf{G}$ is finite, the previous display allows to write

$$E\left[\iota_n \boldsymbol{\xi}^t \iota_n \,|\, U\right] = \sum_{G\in\mathbf{G}} \sum_{i\in r_G} x_{i_1} E\left[\xi_{i_1 i_2} \cdots \xi_{i_t i_{t+1}} \,|\, U\right] y_{i_{s+1}} \tag{A.15}$$

$$= \left(1 + O_p\left(p_n \frac{1}{n}\right)\right) \cdot \sum_{G\in\mathbf{G}} x' A^{s(G)} y \tag{A.16}$$

$$= \left(1 + O_p\left(p_n + \frac{1}{n}\right)\right) \cdot \sum_{s=1}^{t-1} \tilde{\gamma}_s(t) \cdot x' A^{s(G)} y \tag{A.17}$$

In the second equation, $s(G)$ is the number of unique edges in G. Note that $s(G) \leq t/2$ since every edge must have multiplicity at least 2. In the last equation, we collected the powers of A and defined $\tilde{\gamma}_s(t)$ to be the number of walks of length t with s unique nodes.

Let us now return to the arbitrary block B. As discussed previously, it is sufficient to consider graphs in which each walk forms a component that is disconnected from all others. On those graphs, each path is independent from the others. Equation (A.15) therefore allows us to write

$$\begin{aligned} E\left[\iota_n' B \iota_n \,|\, U\right] - \iota_n' \left(A^{t-\tau} \prod_{j=1}^{r} \tilde{\gamma}(p_j, A)\right) \iota_n &= O_p\left(p_n + \frac{1}{n}\right) E\left[\iota_n' B \iota_n \,|\, U\right] \\ &= O_p\left(p_n + \frac{1}{n}\right) O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right) \\ &= O_p\left(\sqrt{p_n} + \frac{1}{\sqrt{n}}\right) O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right) . \end{aligned}$$

The last equality above uses the fact that $np_n \to \infty$, yielding the desired bound. It is difficult to provide closed-form expression for $\tilde{\gamma}_s(t)$ since they are highly combinatorial. However, walks of length t are easy to enumerate on the computer for moderate t. Before proceeding, let us rewrite the $\tilde{\gamma}$ function in a more convenient form. Define $\gamma(B, A)$ such that

$$\left(A^{t-\tau} \prod_{j=1}^{r} \tilde{\gamma}(p_j, A) \right) = \gamma_1(B)A + ... + \gamma_{t-1}(B)A^{t-1} \tag{A.18}$$

Here, $\gamma_t(B) = 0$ because $\tau \geq 2$, and for each block of $\boldsymbol{\xi}$ in B, we have that every constituent $s(G)$ satisfies $s(G) \leq t/2$.

At this point, we have found a good estimator for $E\left[\iota_n' B \iota_n \,|\, U\right]$ in terms of the unobserved matrix A. In order to estimate $E\left[\iota_n' B \iota_n \,|\, U\right]$ at a good rate, we need good estimators for $\iota_n A^t \iota_n$. Let $\tilde{g}(t)$ be our estimator for $\iota_n A^t \iota_n$. Then we seek:

$$\iota_n' A^t \iota_n - \iota_n' \tilde{g}(t) \iota_n = o_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(n^{t+1} p_n^t\right) \ . \tag{A.19}$$

Suppose we estimate $\iota_n A^t \iota_n$ with the naive estimator: $\iota_n \hat{A}^t \iota_n$. Our proofs, in particular Lemma A.2, yields that the estimator is consistent at the following rate

$$\frac{\iota_n A^t \iota_n}{\iota_n \hat{A}^t \iota_n} = 1 + O_p\left(\frac{1}{np_n}\right) \ ,$$

so that the error term is too large relative to the variance.

Next write

$$\iota_n' \hat{A}^t \iota_n = \iota_n' A^t \iota_n + \sum_{B \in \mathbf{B}} \iota_n' B \iota_n \ .$$

Suppose for now that we have access to $\tilde{g}(1), ..., \tilde{g}(t-1)$ satisfying Equation (A.19). We can then consider defining

$$\tilde{g}(t) := \iota_n' \hat{A}^t \iota_n - \sum_{B \in \mathbf{B}} \gamma(B, g)$$

where with some abuse of notation, we define

$$\gamma(B, g) := \gamma_1(B)\tilde{g}(1) + ... + \gamma_{t-1}(B)\tilde{g}(t-1) \ .$$

Since $\tilde{g}(1), ..., \tilde{g}(t-1)$ satisfy Equation (A.19), and noting that $\mathbf{B}$ is finite,

$$\iota_n' \hat{A}^t \iota_n - \sum_{B \in \mathbf{B}} \gamma(B, g) = \iota_n' A^t \iota_n + O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(n^{t+1/2} p_n^{t-1/2}\right) \ .$$

which satisfies Equation (A.19) since $\frac{1}{\sqrt{np_n}} \to 0$. As such, we can recursively construct $\tilde{g}(t)$ from $\tilde{g}(1), ..., \tilde{g}(t-1)$. However, as our proofs in Case (a) shows, $\tilde{g}(1) = \hat{A}$ is valid. Rewrite $\tilde{g}$ such that $\tilde{g} = g$ and

$$g(t) = g_1(t)\hat{A} + \cdots + g_t(t)\hat{A}^t \ .$$

The coefficients of $g(t)$ are presented in Table A.5.

With $\gamma(\cdot, g)$ in hand, we are able to estimate $E\left[\iota_n B \iota_n \,|\, U\right]$ for arbitrary B. Recall that the bias of $\hat{\beta}^{(T)}$ is

$$B_*^{(T)} = \sum_{B \in \mathbf{B}} \delta^{t(B)} \iota_n' B \iota_n \ .$$

As before, $\mathbf{B}$ is the set of even Bs that are generated. Note that this set is "asymmetric" in that A^t appears as a product from the left but not the right. $t(B)$ is the function giving

the order of B. Debiasing by our estimators, we obtain that

$$\sum_{B\in\mathbf{B}} \delta^{t(B)} \iota_n' \left(B - \gamma(B,g)\right) \iota_n = O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(n^{2T} p_n^{2T-1}\right) \ .$$

This is because B is even and of order at most $2T$. As such, since $\tau \geq 2$,

$$\iota_n' B - \gamma(B,g)\iota_n = O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(n^{2T+1-\tau/2} p_n^{2T-\tau/2}\right) = O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(n^{2T} p_n^{2T-1}\right)$$

Now, since $\sqrt{V_*^{(T)}} = \Omega_p\left(n^{2T} p_n^{2T-1/2}\right)$, we conclude that

$$\frac{\hat{B}_*^{(T)} - B_*^{(T)}}{\sqrt{V_*^{(T)}}} = o_p(1) \ .$$

Substituting our computed values of $\tilde{\gamma}_s(t)$ and $g(t)$ yields the formula given in Appendix A.1.

A.4.3.3. Case (c). Suppose $\beta^{(d)} = 0$ for $d \in \{1, T\}$. We can write

$$\hat{\beta}^{(T)} = \frac{\left(\hat{C}^{(T)}\right)' \varepsilon^{(T)}}{\left(\hat{C}^{(T)}\right)' \hat{C}^{(T)}} = \frac{\iota_n \left(\sum_{t=1}^T \delta^t \hat{A}^t\right) \varepsilon^{(T)}}{\iota_n \left(\sum_{t=1}^T \delta^t \hat{A}^t\right)^2 \iota_n}$$

Let B be a mixed product of order t, and let it have order $\tau \geq 0$ with respect to $\boldsymbol{\xi}$. Then,

$$E\left[\left(\left(\varepsilon^{(T)}\right)' B \iota_n\right)^2 \,|\, U, \boldsymbol{\xi}\right]$$
$$= \sum_{i_1=1}^n \sum_{i_2=1}^n \cdots \sum_{i_{2t+2}=1}^n E\left[\varepsilon_{i_1}^{(T)} \varepsilon_{i_{t+2}}^{(T)} \,|\, U, \boldsymbol{\xi}\right] B_{i_1,i_2} B_{i_2,i_3} \cdots B_{i_t,i_{t+1}} \cdot B_{i_{t+2},i_{t+3}} B_{i_{t+3},i_{t+4}} \cdots B_{i_{2t+1},k_{2t+2}}$$

r \ t	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	-1	2	-5	12	-20	-12	295	-1584	5623	-12530	-1806	186702
2		1	-2	4	-8	8	42	-340	1510	-4712	8408	13088	-194318
3			1	-3	7	-14	10	96	-655	2552	-6190	1068	83832
4				1	-4	11	-24	22	142	-1043	4078	-9444	-2150
5					1	-5	16	-39	48	176	-1558	6542	-16554
6						1	-6	22	-60	94	178	-2170	10028
7							1	-7	29	-88	167	122	-2836
8								1	-8	37	-124	275	-26
9									1	-9	46	-169	427
10										1	-10	56	-224
11											1	-11	67
12												1	-12
13													1

r \ t	14	15	16	17	18	19	20
1	-1101323	3938488	-7533897	-13585642	198008994	-999517964	3021609795
2	981200	-3101066	4292162	20354680	-188470026	832916330	-2145039932
3	-530446	2005368	-4310942	-4074647	91205574	-496007668	1614224856
4	151068	-879116	3034670	-4907736	-17574745	186419358	-871382472
5	4548	213314	-1337608	4785512	-8228118	-25081260	283591630
6	-28178	22194	281946	-2019280	7855844	-16309132	-23702626
7	14700	-46038	58866	333648	-2899960	12404253	-30152117
8	-3480	20662	-72062	124800	337020	-3955392	18806973
9	-309	-3984	27916	-108304	232853	246742	-5122509
10	633	-780	-4178	36312	-156828	398862	-1830
11	-290	904	-1503	-3829	45486	-219512	641615
12	79	-368	1252	-2554	-2629	54786	-297780
13	-13	92	-459	1690	-4022	-182	63185
14	1	-14	106	-564	2232	-6010	4010
15		1	-15	121	-684	2893	-8636
16			1	-16	137	-820	3689
17				1	-17	154	-973
18					1	-18	172
19						1	-19
20							1

Table A.5. The coefficients $g_r(t)$ for $t \leq 20$.

Now, $E\left[\varepsilon_{i_1}^{(T)}\varepsilon_{i_{t+2}}^{(T)} \,|\, U, \boldsymbol{\xi}\right] = 0$ unless $i_1 = i_{t+2}$. Hence, we only need to consider sequences i where $i_1 = i_{t+2}$. Under this restriction,

$$E\left[\left(\left(\varepsilon^{(T)}\right)' B\iota_n\right)^2 \,|\, U, \boldsymbol{\xi}\right] \leq \bar{\sigma}^2 \iota_n' \tilde{B} \iota_n \ .$$

where $\tilde{B}$ is of order $2t+1$ unconditionally, and of order 2τ with respect to $\boldsymbol{\xi}$. Conclude by Lemma A.2 that

$$\left(\varepsilon^{(T)}\right)' B\iota_n = O_p\left(\sqrt{n^{2t+3/2-\tau} p_n^{2t+1-\tau}}\right) = O_p\left(n^{t+3/4-\tau/2} p_n^{t+1/2-\tau/2}\right) \ . \tag{A.20}$$

Next, write

$$\begin{aligned}
\frac{1}{p_n^T} \iota_n' A^T \varepsilon^{(T)} &= \sum_{i_1,\ldots,i_{T+1}} f(U_{i_1}, U_{i_2}) \cdots f(U_{i_T}, U_{i_{T+1}}) \varepsilon_{i_{T+1}}^{(T)} \\
&= \sum_{i \in I_{T+1}} \sum_{\pi \in \Pi_{T+1}} f(U_{i_{\pi(1)}}, U_{i_{\pi(2)}}) \cdots f(U_{i_{\pi(T)}}, U_{i_{\pi(T+1)}}) \varepsilon_{(i_{\pi(T+1)})}^{(T)} \ .
\end{aligned}$$

where I_{T+1} comprises all unordered subsets of $T+1$ integers chosen from $[n]$ and Π_{T+1} is the set of permutations on $[T+1]$. We can hence define the following symmetric U-statistic kernel of order $T+1$:

$$h\left(\left(U_{i_1}, \varepsilon_{i_1}^{(T)}\right), \cdots, \left(U_{i_{T+1}}, \varepsilon_{i_{T+1}}^{(T)}\right)\right) = \sum_{\pi \in \Pi_{T+1}} f(U_{i_{\pi(1)}}, U_{i_{\pi(2)}}) \cdots f(U_{i_{\pi(T)}}, U_{i_{\pi(T+1)}}) \varepsilon_{i_{\pi(T+1)}}^{(T)} \ .$$

Since f is bounded and $\varepsilon_i^{(T)}$ has uniformly bounded conditional expectations, $E[h^2] < \infty$. By the U-statistic CLT (Theorem 12.3 in Van der Vaart 2000),

(A.21)

$$\sqrt{n}\frac{1}{\binom{n}{T+1}}\sum_{i\in I_{T+1}}\sum_{\pi\in\Pi_{T+1}} f(U_{i_{\pi(1)}},U_{i_{\pi(2)}})\cdots f(U_{i_{\pi(T)}},U_{i_{\pi(T+1)}})\varepsilon^{(T)}_{i_{\pi(T+1)}} \xrightarrow{d} \mathrm{N}\left(0\,,\,(T+1)^2\zeta_1\right)\ ,$$

where

$$\begin{aligned}
\zeta_1 &= E\left[h\left(\left(U_1,\varepsilon_1^{(T)}\right),\left(U_2,\varepsilon_2^{(T)}\right),\cdots,\left(U_{T+1},\varepsilon_{T+1}^{(T)}\right)\right)h\left(\left(U_1,\varepsilon_1^{(T)}\right),\left(U_{T+2},\varepsilon_{T+2}^{(T)}\right),\cdots,\left(U_{2T+1},\varepsilon_{2T+1}^{(T)}\right)\right)\right] \\
&= E\left[\sum_{\pi\in\Pi_T}\sum_{\pi'\in\Pi'_T} f(U_{\pi(1)},U_{\pi(2)})\cdots f(U_{\pi(T)},U_{T+1})\cdot f(U_{\pi'(T+1)},U_{\pi'(T+2)})\cdots f(U_{\pi'(2T)},U_{T+1})\left(\varepsilon_{T+1}^{(T)}\right)^2\right] \\
&= (T!)^2\,E\left[f(U_1,U_2)\cdots f(U_T,U_{T+1})\cdot f(U_1,U_{T+2})\cdots f(U_{2T},U_{2T+1})\left(\varepsilon_1^{(T)}\right)^2\right] \neq 0 \text{ by assumption.}
\end{aligned}$$

Here, Π'_T is the set of permutations on $\{T+1,...,2T\}$. As such,

$$\iota_n' A^T \varepsilon^{(T)} = O_p\left(\frac{1}{\sqrt{n}} n^{T+1} p_n^T\right)\ .$$

Together with the bound in Equation (A.20), this implies that $\iota_n' A^T \varepsilon^{(T)}$ is the dominant term in the $\left(\hat{C}^{(T)}\right)' \varepsilon^{(T)}$ Next, note that by the U-statistic LLN,

$$\begin{aligned}
&\frac{1}{n^{2T+1}}\sum_{j=1}^{n}\left(\iota_n\left(A^T\right)_{\cdot,j}\right)^2\left(\varepsilon_j^{(T)}\right)^2 \\
&= \frac{1}{n^{2T+1}}\sum_{i_1,\ldots,i_{2T+1}} f(U_{i_1},U_{i_2})\cdots f(U_{i_T},U_{i_{T+1}})\cdot f(U_{i_1},U_{i_{T+2}})\cdots f(U_{i_{2T}},U_{i_{2T+1}})\left(\varepsilon_{i_{2T+1}}^{(T)}\right)^2 \\
&\xrightarrow{p} \frac{1}{(T!)^2}\zeta_1\ .
\end{aligned}$$

By the usual plug-in arguments, we have that

$$\begin{aligned}
\frac{1}{n^{2T+1}}\left(\left(\hat{C}^{(T)}\right)'\hat{C}^{(T)}\right)^2\hat{V}_0^{(T)} &= \frac{1}{n^{2T+1}}\sum_{j=1}^{n}\left(\hat{C}_j^{(T)}\right)^2\left(\hat{\varepsilon}_j^{(T)}\right)^2 \\
&= \frac{1}{n^{2T+1}}\sum_{j=1}^{n}\left(C_j^{(T)}\right)^2\left(\varepsilon_j^{(T)}\right)2 + o_p(1) \quad \text{by the consistency of } \hat{\beta}^{(T)} \\
&= \frac{1}{n^{2T+1}}\sum_{j=1}^{n}\delta^{2T}\left(\iota_n\left(A^T\right)_{\cdot,j}\right)^2\left(\varepsilon_j^{(T)}\right)^2 + o_p(1) \quad \text{by the dominance of } A^T
\end{aligned}$$

As such, the robust/heteroskedasticity consistent t-statistic

$$\begin{aligned}
\frac{\hat{\beta}^{(T)}}{\sqrt{\hat{V}_0^{(T)}}} &= \frac{\left(\hat{C}^{(T)}\right)'\varepsilon^{(T)}}{\sqrt{\left(\left(\hat{C}^{(T)}\right)'\hat{C}^{(T)}\right)^2\hat{V}_0^{(T)}}} \\
&= \frac{n^{-(T+1/2)}\delta^T\iota_n'A^T\varepsilon^{(T)}}{\sqrt{\frac{1}{n^{2T+1}}\sum_{j=1}^{n}\delta^{2T}\left(\iota_n\left(A^T\right)_{\cdot,j}\right)^2\left(\varepsilon_j^{(T)}\right)^2}} + o_p(1) \\
&= \frac{\frac{\sqrt{n}}{\binom{n}{T+1}}\sum_{i\in I_{T+1}}\sum_{\pi\in\Pi_{T+1}}f(U_{i_{\pi(1)}},U_{i_{\pi(2)}})\cdots f(U_{i_{\pi(T)}},U_{i_{\pi(T+1)}})\varepsilon_{i_{\pi(T+1)}}^{(T)}}{(T+1)!\sqrt{\frac{1}{(T!)^2}\zeta_1}} + o_p(1) \\
&\xrightarrow{d} \mathrm{N}(0,1) \quad \text{by Equation (A.21)}
\end{aligned}$$

As such, the robust/heteroskedasticity consistent t-statistic is appropriate for inference under the null hypothesis that $\beta^{(T)} = 0$.

A.4.4. Proof of Theorem 1.6

We start by writing

$$\hat{\beta}^{(\infty)} = \frac{Y'\left(a_n v_1(\hat{A})\right)}{\left(a_n v_1(\hat{A})\right)'\left(a_n v_1(\hat{A})\right)} = \beta^{(\infty)}\left(v_1(A)\right)' v_1(\hat{A}) + \frac{1}{a_n}\left(\varepsilon^{(\infty)}\right)' v_1(\hat{A}) \ .$$

The main tool we use to study the above term is the following:

Lemma A.3. Suppose Assumption E2 holds and that p_n satisfies Equation (1.10). Then,

$$\left(v_1(A)\right)' v_1(\hat{A}) = \left(v_1(A)\right)' v_1(A) + \frac{\left(v_1(A)\right)'\boldsymbol{\xi} v_1(A)}{\lambda_1(A)} + \frac{\left(v_1(A)\right)'\boldsymbol{\xi}^2 v_1(A)}{\left(\lambda_1(A)\right)^2} + o_p\left(\frac{1}{(np_n)^3}\right)$$
$$+ \sum_{r=2}^{R} \frac{\lambda_r(A)}{\lambda_1(A)} \frac{v_r(A)' v_1(\hat{A})}{v_r(A)' v_r(\hat{A})} \cdot O_p\left(\frac{1}{np_n}\right)$$

and

$$\left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' v_1(\hat{A}) = \left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' v_1(A) + o_p\left(\frac{1}{a_n}\right) \ .$$

Now, by the analogous arguments as in Lemma A.2 and proof of Lemma A.3,

$$\frac{\left(v_1(A)\right)'\boldsymbol{\xi} v_1(A)}{\lambda_1(A)} = O_p\left(\frac{1}{\sqrt{n}\sqrt{np_n}}\right) \quad , \quad \frac{\left(v_1(A)\right)'\boldsymbol{\xi}^2 v_1(A)}{\left(\lambda_1(A)\right)^2} = O_p\left(\frac{1}{np_n}\right) \quad , \tag{A.22}$$

$$\frac{\left(v_1(A)\right)'\boldsymbol{\xi}^2 v_1(A)}{\left(\lambda_1(A)\right)^2} - \frac{E\left[\left(v_1(A)\right)'\boldsymbol{\xi}^2 v_1(A) \,|\, U\right]}{\left(\lambda_1(A)\right)^2} = O_p\left(\frac{1}{\sqrt{n}\,(np_n)}\right) \tag{A.23}$$

Furthermore, by the Davis-Kahan Inequality,

$$\left|\frac{1}{\lambda_s(A)} v_r(A)' v_s(\hat{A})\right| \leq \left|\frac{1}{\lambda_s(A)} v_r(A)' v_s(A)\right| + \|v_r(A)\| \cdot \left\|\frac{v_s(\hat{A}) - \theta v_s(A)}{\lambda_s(A)}\right\| \ . \tag{A.24}$$

A.4.4.1. Case (a). Let us now consider individual cases in (a). starting with (iii). Suppose we require only that $p_n \succ n^{-1}\left(\frac{\log n}{\log\log n}\right)^{1/2+\eta}$. Then by Lemma A.1, we can claim that

$$\left\| \frac{v_s(\hat{A}) - \theta v_s(A)}{\lambda_s(A)} \right\| = o_p(1)$$

but cannot control the rate of convergence. Nonetheless, if $a_n \succ np_n$, Equations (A.22) and (A.24), together with Lemma A.3 implies that

$$\hat{\beta}^{(\infty)} = \beta^{(\infty)} + \left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' v_1(A) + o_p\left(\frac{1}{a_n}\right) . \tag{A.25}$$

Note that bias correction is irrelevant in this regime since bias is of smaller order than a_n.

Suppose instead, as in Case (a) (ii) that $p_n \succ n^{-1}\log n$. Then, by Theorem 1.1 in the Supplementary Appendix to Lei and Rinaldo (2015), we can claim that

$$\left\| \frac{v_s(\hat{A}) - \theta v_s(A)}{\lambda_s(A)} \right\| = O_p\left(\frac{1}{\sqrt{np_n}}\right)$$

Then, provided that $a_n \prec (np_n)^{3/2}$,

$$\hat{\beta}^{(\infty)} - \beta^{(\infty)} - \beta^{(\infty)} \frac{E\left[(v_1(A))' \boldsymbol{\xi}^2 v_1(A) \,|\, U\right]}{(\lambda_1(A))^2} = \left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' v_1(A) + o_p\left(\frac{1}{a_n}\right) .$$

This time the bias correction is required.

Finally, consider Case (a) (i), when $\beta^{(\infty)} = 0$. Then it is immediate that Equation (A.25) obtains. In all three cases, the asymptotic distribution of the estimator depends on $\left(\varepsilon^{(\infty)}\right)' v_1(A)/a_n$. Next, notice that

$$\mathrm{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right) = E\left[\left(\left(\varepsilon^{(\infty)}\right)' v_1(A)\right)^2 \,|\, U\right] = \sum_{i=1}^{n} [v_1(A)]_i^2 \, E\left[\left(\varepsilon_i^{(\infty)}\right)^2 \,|\, U\right] .$$

As such, by Assumption 1.3,

$$\underline{\sigma}^2 \leq \text{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right) \leq \bar{\sigma}^2 \ .$$

Now, as in the proof of Lemma A.3, let Υ be the event that $|\frac{1}{\sqrt{n}} \sum_{i=1}^{r} \phi_r(U_i)\phi_s(U_i) < 1/R^2$. This happens with probability approaching 1 since R is finite. On this event, $\|v\| = 1$ implies that $|v_r| < 2$ for all r. Furthermore, observe that since $\|f\|_\infty \leq 1$, $\|\phi_r\|_\infty \leq 1$. As such, on Υ, $|v_r(A)\|_\infty \leq 2R/\sqrt{n}$. As such,

$$\sum_{i=1}^{n} \frac{[v_1(A)]_i^3 \, E\left[\left|\varepsilon_i^{(\infty)}\right|^3 \,|\, U\right]}{\text{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right)} \leq \frac{\bar{\kappa}_3}{\underline{\sigma}^2} \frac{8R^3}{\sqrt{n}} \to 0 \ .$$

Note that the bound on the right-hand side does not depend on U. Putting the above ingredients together, we have that on Υ, the Berry-Esseen Inequality of Chen et al. (2011) yields

$$\sup_{z \in \mathbf{R}} \left| P\left(\frac{\left(\varepsilon^{(\infty)}\right)' v_1(A)}{\sqrt{\text{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right)}} \leq z \right) - \Phi(z) \right| \leq 10 \cdot \frac{\bar{\kappa}_3}{\underline{\sigma}^2} \frac{8R^3}{\sqrt{n}} \ .$$

Then since $P(\Upsilon) \to 1$, we have that

$$\frac{\left(\varepsilon^{(\infty)}\right)' v_1(A)}{\sqrt{\text{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right)}} \xrightarrow{d} \text{N}(0,1) \ .$$

Define

$$B^{(\infty)} = \frac{E\left[(v_1(A))' \boldsymbol{\xi}^2 v_1(A) \,|\, U\right]}{\left(\lambda_1(A)\right)^2} \ ,$$

$$V_0^{(\infty)} = \text{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right) \ ,$$

Then we have that

$$\frac{\hat{\beta}^{(\infty)} - \beta^{(\infty)}\left(1 - B^{(\infty)}\right)}{\sqrt{V_0^{(\infty)}}} = \frac{\varepsilon' v_1(A)}{\sqrt{\mathrm{Var}\left(\left(\varepsilon^{(\infty)}\right)' v_1(A) \,|\, U\right)}} + o_p(1) \xrightarrow{d} \mathrm{N}(0,1) \ .$$

The validity of plug-in estimation follows from arguments that are essentially identical to Section A.4.3.3.

A.4.4.2. Case (b). Suppose $a_n \succ n\sqrt{p_n}$. By Lemma A.3, we have that

$$\hat{\beta}^{(\infty)} = \beta^{(\infty)} + \beta^{(\infty)}\Bigg[(v_1(A))' v_1(A) + \frac{(v_1(A))' \boldsymbol{\xi} v_1(A)}{\lambda_1(A)} + \frac{(v_1(A))' \boldsymbol{\xi}^2 v_1(A)}{(\lambda_1(A))^2} + o_p\left(\frac{1}{(np_n)^3}\right) + \sum_{r=2}^{R} \frac{\lambda_r(A)}{\lambda_1(A)} \frac{v_r(A)' v_1(\hat{A})}{v_r(A)' v_r(\hat{A})} \cdot O_p\left(\frac{1}{np_n}\right) + o_p\left(\frac{1}{a_n}\right)\Bigg] \ .$$

Furthermore, since $p_n \succ n^{-1} \log n$, the Davis-Kahan Inequality (Theorem 4.5.5 in Vershynin 2018), together with Theorem 1.1 in the Supplymentary Material to Lei and Rinaldo (2015) gives us that

$$\sum_{r=2}^{R} \frac{\lambda_r(A)}{\lambda_1(A)} \frac{v_r(A)' v_1(\hat{A})}{v_r(A)' v_r(\hat{A})} = O_p\left(\frac{1}{\sqrt{np_n}}\right) \ .$$

As such,

$$\begin{aligned} \hat{\beta}^{(\infty)} - \beta^{(\infty)}\left(1 - B^{(\infty)}\right) &= \beta^{(\infty)} \frac{(v_1(A))' \boldsymbol{\xi} v_1(A)}{\lambda_1(A)} + O_p\left(\frac{1}{(np_n)^{3/2}}\right) \\ &= \beta^{(\infty)} \frac{(v_1(A))' \boldsymbol{\xi} v_1(A)}{\lambda_1(A)} + O_p\left(\frac{1}{n\sqrt{p_n}}\right) \ , \end{aligned} \tag{A.26}$$

since $p_n \succ 1/\sqrt{n}$. Now note that

$$\begin{aligned}
&\operatorname{Var}\left(v_1(A)'\boldsymbol{\xi} v_1(A) \,|\, U\right) \\
&= E\left[\left(v_1(A)'\boldsymbol{\xi} v_1(A)\right)\right] \\
&= E\left[\left(2\sum_{i<j}\left[v_1(A)\right]_i\left[v_1(A)\right]_j\xi_{ij}\right)^2 \,\middle|\, U\right] \\
&= \sum_{i<j}\left[v_1(A)\right]_i^2\left[v_1(A)\right]_j^2 p_n f(U_i,U_j)\left(1-p_n f(U_i,U_j)\right) \\
&= 4\left(1+O_p(p_n)\right)\cdot\sum_{i<j}\left[v_1(A)\right]_i^2\left[v_1(A)\right]_j^2 p_n f(U_i,U_j)
\end{aligned}$$

Recall that from the proof of Lemma A.3 that

$$A = \sum_{r=1}^{R}\tilde{\lambda} n\left(\frac{\phi_r(U)}{\sqrt{n}}\right)\frac{\phi_r(U)}{\sqrt{n}} \quad \Rightarrow \quad v_1(A) = \sum_{r=1}^{R}\alpha_r\frac{\phi_r(U)}{\sqrt{n}}\ .$$

so that $|\alpha_r| \le 2R$ for all $r \in [R]$ w.p.a. 1. We now argue that $\alpha_1 \xrightarrow{p} 1$ and $\alpha_r \to 0$ for $r \ge 2$. Note that we can write

$$\begin{aligned}
Av_1(A) &= \left(\sum_{r=1}^{R}\tilde{\lambda}_r n\left(\frac{\phi_r(U)}{\sqrt{n}}\right)\frac{\phi_r(U)}{\sqrt{n}}\right)\left(\sum_{r=1}^{R}\alpha_r\frac{\phi_r(U)}{\sqrt{n}}\right) \\
&= \sum_{r=1}^{R}\tilde{\lambda}_r n\cdot\alpha_r\phi_r\cdot\left(\frac{\phi_r(U)}{\sqrt{n}}\right)'\frac{\phi_r(U)}{\sqrt{n}} + \sum_{r\neq s}\tilde{\lambda}_r\alpha_s\left(\frac{\phi_r(U)}{\sqrt{n}}\right)\frac{\phi_s(U)}{\sqrt{n}}\ .
\end{aligned}$$

Consequently,

$$
\begin{aligned}
(v_1(A))' A v_1(A) = & \sum_{r=1}^{R} \tilde{\lambda}_r n \alpha_r^2 \left(\left(\frac{\phi_r(U)}{\sqrt{n}} \right)' \frac{\phi_r(U)}{\sqrt{n}} \right)^2 \\
& + \sum_{r=1}^{R} \sum_{s=1}^{R} \tilde{\lambda}_r n \alpha_r \alpha_r \left(\frac{\phi_r(U)}{\sqrt{n}} \right) \frac{\phi_s(U)}{\sqrt{n}} \left(\frac{\phi_r(U)}{\sqrt{n}} \right)' \left(\frac{\phi_r(U)}{\sqrt{n}} \right) \\
& + \sum_{r=1}^{R} \sum_{s \neq t} \tilde{\lambda}_s n \alpha_t \alpha_r \left(\frac{\phi_r(U)}{\sqrt{n}} \right) \frac{\phi_s(U)}{\sqrt{n}} \left(\frac{\phi_r(U)}{\sqrt{n}} \right) \frac{\phi_t(U)}{\sqrt{n}}
\end{aligned}
$$

Now,

$$
\left(\frac{\phi_r(U)}{\sqrt{n}} \right)' \frac{\phi_r(U)}{\sqrt{n}} = \frac{1}{n} \sum_{i=1}^{n} \phi_r(U_i) \phi_s(U_i) = \begin{cases} 1 + O_p\left(\frac{1}{\sqrt{n}}\right) & \text{if } r = s \\ O_p\left(\frac{1}{\sqrt{n}}\right) & \text{if } r \neq s \end{cases}
$$

Since $|\alpha_r| \leq 2R$ w.p.a 1,

$$
(v_1(A))' A v_1(A) = \sum_{r=1}^{R} \tilde{\lambda}_r n \alpha_r^2 + o_p(1)
$$

Since $|\tilde{\lambda}_1| > |\tilde{\lambda}_r|$ for $r \geq 2$,

$$
\frac{(v_1(A))' A v_1(A)}{\tilde{\lambda}_1 n} \xrightarrow{p} 1 \quad \Rightarrow \quad \alpha_1 \xrightarrow{p} 1 \ .
$$

Doing a similar expansion for $\|v\|^2$, we arrive at

$$
1 = \|v_1(A)\|^2 = \sum_{r=1}^{R} \alpha_r^2 + O_p\left(\frac{1}{\sqrt{n}}\right) \ .
$$

Since $\alpha_1 \xrightarrow{p} 1$,

$$\alpha_r \xrightarrow{p} 0 \quad \text{for all } r \geq 2 \; . \tag{A.27}$$

Since $|\phi_r(U_i)| \leq 1$, the above analysis also implies that

$$|[V_1(A)]_i| = \left| \sum_{r=1}^{R} (\alpha_r + o_p(1)) \frac{\phi_r(U_i)}{\sqrt{n}} \right| \leq \frac{1}{\sqrt{ns}} \sum_{r=1}^{R} |\alpha_r| + o_p(1)$$

where the bound on the right hand side depends on the convergence of α_r and does not vary across i. This yields $\|v_1(A)\|_\infty \leq 2/\sqrt{n}$ w.p.a. 1.

As such,

$$\mathrm{Var}\left(n \cdot v_1(A)' \boldsymbol{\xi} v_1(A) \,|\, U\right) = 4\left(1 + o_p(1)\right) \sum_{i<j} \phi_1(U_i)^2 \phi_1(U_j)^2 p_n f(U_i, U_j) \; .$$

Conclude by the U-statistics LLN that

$$\frac{1}{n^2 p_n} \mathrm{Var}\left(n \cdot v_1(A)' \boldsymbol{\xi} v_1(A) \,|\, U\right) \xrightarrow{p} 2E\left[\phi_1(U_1)^2 \phi_1(U_2)^2 f(U_1, U_2)\right] > 0 \; .$$

As such, if we define $\Upsilon^{(\infty)}$ to be the event on which

$$\mathrm{Var}\left(n \cdot v_1(A)' \boldsymbol{\xi} v_1(A) \,|U\right) > n^2 p_n \cdot E\left[\phi_1(U_1)^2 \phi_1(U_2)^2 f(U_1, U_2)\right]$$

and $\|v_1(A)\|_\infty \leq 2/\sqrt{n}$. By the Berry-Esseen Inequality,

$$\sup_{x \in \mathbf{R}} \left| P\left(\frac{\left(n \cdot v_1(A)\right) \boldsymbol{\xi} v_1(A)}{\sqrt{\mathrm{Var}\left(n \cdot v_1(A)' \boldsymbol{\xi} v_1(A) \,|\, U\right)}} \leq x \,\middle|\, U, \Upsilon^{(\infty)} \right) - \Phi(x) \right| \leq 10\gamma$$

where

$$\gamma = \sum_{i<j} \left[\sqrt{n} \cdot v_1(A)\right]_i^3 \left[\sqrt{n} v_1(A)\right]_j^3 \frac{E\left[\xi_{ij}^3 \,|\, U\right]}{\left(\mathrm{Var}\left(n \cdot v_1(A)' \boldsymbol{\xi} v_1(A) \,|\, U\right)\right)^{3/2}}$$
$$\leq 64R^4 \frac{n^2 p_n}{\left(n^2 p_n E\left[\phi_1(U_1)^2 \phi_1(U_2)^2 f(U_1, U_2)\right]\right)^{3/2}} \to 0$$

where the last bound follows because we are on the event Υ and does not depend on U otherwise. Substituting this into Equation (A.26), we have that

$$\frac{\lambda_1(A)\left(\hat{\beta}^{(\infty)} - \beta^{(\infty)}\left(1 - B^{(\infty)}\right)\right)}{\beta^{(\infty)}\sqrt{V^{(\infty)}}} = \frac{\left(n \cdot v_1(A)\right) \boldsymbol{\xi} v_1(A)}{\sqrt{\mathrm{Var}\left(n \cdot v_1(A)' \boldsymbol{\xi} v_1(A) \,|\, U\right)}} + o_p(1) \xrightarrow{d} \mathrm{N}(0,1) \ ,$$

where the estimate on the error follows because $\lambda_1(A)/\sqrt{V^{(\infty)}} = O_p\left(n^{-1} p_n^{-1/2}\right)$.

The validity of plug-in estimation follows from arguments that are essentially identical to Section A.4.3.1.

A.4.4.3. Proof of Corollary 1.5. By our analysis of $v_1(\hat{A})$ above, we have that

$$n \cdot a_n^{-2} = n\left(\frac{1}{n}\sum_{i=1}^n [v_1(\hat{A})]_i^2 - \left(\frac{1}{n}\sum_{i=1}^n [v_1(\hat{A})]_i\right)^2\right)$$
$$= 1 - \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n [v_1(\hat{A})]_i\right)^2 \text{ since } \|v_1(\hat{A})\| = 1$$
$$= 1 - \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\phi_1(U_i)}{\sqrt{n}} + o_p(1)\right)^2 \text{ by Equation (A.27)}$$
$$\xrightarrow{p} 1 - E\left[\phi_1(U_1)\right]^2 \ .$$

A.4.5. Proofs of Auxillary Lemmas

A.4.5.1. Proof of Lemma A.1. Suppose $np_n \succcurlyeq \log^2 n$. By Theorem 1.1 in the Supplementary Material to Lei and Rinaldo (2015), noting that the constants in their bounds are uniform in P_{ij} we have that with probability at least $1 - n^{-r}$, where r can be chosen independently of

$$\left\|A - \hat{A}\right\| \leq k(r)\sqrt{np_n}$$

where $k(r)$ is a constant that depends only on r.

Suppose instead that $\sqrt{\frac{\log n}{\log\log n}} \prec np_n \prec \log^2 n$. Our set up satisfies the requirements for Corollary 3.3 in Benaych-Georges et al. (2020). Setting their $\varepsilon^2 = \left(\sqrt{\frac{\log n}{\log\log n}}/(np_n)\right)^{1-\nu}$ and noting that their d is our np_n, we have that with probability at least $1-\exp\left(-(np_n)^{2+\nu}\left(\frac{\log n}{\log\log n}\right)^{(1-\nu)/}\right.$

$$\left\|A - \hat{A}\right\| \leq k\,(np_n)^{(1+\nu)/2}\left(\frac{\log n}{\log\log n}\right)^{(1-\nu)/4}$$

where k is a universal constant.

Combining these two inequalities yields the desired results.

A.4.5.2. Proof of Lemma A.2. Let B be a mixed product as in Definition A.2. Suppose $B_j = \boldsymbol{\xi}$ for at least one $j \in [t]$ and write:

$$\begin{aligned}(\iota_n' B \iota_n)^2 &= \left(\sum_{i=1}^{n}\sum_{k_1=1}^{n}\cdots\sum_{k_{t-1}=1}^{n}\sum_{j=1}^{n} b_{i,k_1}b_{k_1,k_2}\cdots b_{k_{t-1},j}\right)\cdot\left(\sum_{i'=1}^{n}\sum_{k_1'=1}^{n}\cdots\sum_{k_{t-1}'=1}^{n}\sum_{j'=1}^{n} b_{i',k_1'}b_{k_1',k_2'}\cdots b_{k_{t-1}',j'}\right)\\ &= \sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\cdots\sum_{k_{2t+2}=1}^{n} b_{k_1,k_2}b_{k_2,k_3}\cdots b_{k_t,k_t+1}\cdot b_{k_{t+2},k_{t+3}}b_{k_{t+3},k_{t+4}}\cdots b_{k_{2t+1},k_{2t+2}}\end{aligned}$$

In the second line we relabel the indices of summation. Each term in the above sum is a product of $2t$ terms. Note that the term $b_{k_{t+1},k_{t+2}}$ does not exist. Next, we take conditional

expectations:

$$E\left[(\iota_n' B \iota_n)^2 \mid U\right] = \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_{2t+2}=1}^{n} E\left[b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} \cdot b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \cdots b_{k_{2t+1},k_{2t+2}} \mid U\right] .$$

Notice that each summand is non-zero if an only if for each $b_{k_j,k_{j+1}} = \xi_{k_j,k_{j+1}}$, there is some j' such that

$$b_{k_{j'},k_{j'+1}} = \xi_{k_{j'},k_{j'+1}} = \begin{cases} \xi_{k_j,k_{j+1}} & \text{or,} \\ \xi_{k_{j+1},k_j} \end{cases}$$

In other words, each ξ_{ij} that appears in the summand appears at least twice, either as ξ_{ij} or ξ_{ji}. This property depends on the positions of the A's to the extent that they break up $\boldsymbol{\xi}$: neighbouring ξ_{ij} and ξ_{jk} share an index so that setting $i = k$ is sufficient for the conditional expectation of their product to be non-zero. If ξ_{ij} and ξ_{kl} are separated by at least one A_{pq}, we need to set $k = i$ and $l = j$. The number of restrictions on the indices that are needed for the terms to be non-zero therefore depend on $\mathcal{J}$ and p. In turn, these restrictions determine the order of magnitude of the conditional expectation.

We are interested in relations on $k_1, ..., k_{2t+2}$ which will make

$$E\left[b_{k_1,k_2} \cdots b_{k_t,k_{t+1}} \cdot b_{k_{t+2},k_{t+3}} \cdots b_{k_{2t+1},k_{2t+2}} \mid U\right] \neq 0 .$$

We represent this relation with the multi-graph G on nodes $[n]$ with each ξ_{ij} in the summand corresponding to an edge from i to j. If G is the multi-graph induced by a given relationship, we write that $k_1, ..., k_{2t+2} \in r_G$. Let the contribution of r_G to the our

overall sum be:

$$\sum_{(k_1,\ldots,k_{2t+2})\in r_G} E\left[b_{k_1,k_2}\cdots b_{k_t,k_t+1}\cdot b_{k_{t+2},k_{t+3}}\cdots b_{k_{2t+1},k_{2t+2}}\,|\,U\right] =: S_G$$

For S_G to be non-zero, every edge in G must have multiplicity at least 2. Furthermore, each G is constructed by performing a walk of length p_1, followed by p_2, and so on, until p_r.

The walks relate G to S_G in the following way. Initially, we are given a budget of $n^{2t+2}p_n^{2t}$. The budget on n is the number of times the graph G occurs, corresponding to "degree of freedom". The budget on p_n is the number of unique ξ_{ij} in the term. Given any initial vertex, start the first walk of length p_1. Add one to the multiplicity of each edge taken. In the j^{th} step, incrementing multiplicity from 0 to 1 is free: this corresponds to not restricting k_j and k_{j+1}. Incrementing multiplicity from a to $a+1$ for $a \geq 1$ costs np_n. This is because such a step corresponds to the restriction $k_j = k_{j'}$ where $k_{j'}$ denotes the other end point of the edge whose multiplicity is being incremented. Furthermore, we "lose" p_n when we restrict $\xi_{k_j,k_{j+1}}$ to be equal to an existing edge since there are now fewer unique ξ_{ij}'s. Having completed the first walk, start the second walk. If the first edge of the second walk increments the multiplicity of an edge from 0 to 1, it is free. However, incrementing multiplicity from a to $a+1$ for $a \geq 1$ costs n^2p_n. This is because placing the first edge of a new walk corresponds the two restrictions: $k_j = k_{j'}$, $k_{j+1} = k_{j'+1}$. However, the moments decrease only by p_n since we only lose one unique edge. Continue in the following way until all walks are completed. At the end of the walks, suppose cost is $n^\alpha p_n^\beta$. If every edge in G has multiplicity at least 2, $S_G = n^{2t+2-\alpha}p_n^{2t-\beta}$ by construction. Otherwise, $S_G = 0$.

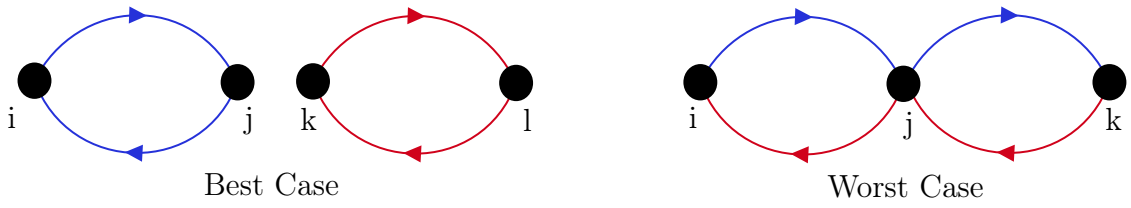

Figure A.4. Potential G for $p = (2, 2)$. Red indicates the first walk. Blue indicates the second walk. In the best case (highest order), $S_G = n^4 p_n^2$. In the worst case (lowest order), $S_G = n^3 p_n^2$.

Suppose $|\mathcal{J}| = 2\tau$ for some $t \geq 1$. Then at least a edges have multiplicity 2. The minimum cost of such a graph is $n^\tau p_n^\tau$, so that $\beta \geq \tau$. Note also that each edge costs weakly more n than p_n. As such, $\alpha \geq \beta$ so that $2t + 2 - \alpha - (2t - \beta) \leq 2$. Taking expectations over U, which preserves the order of the terms and then taking square root gives us the order of $\iota_n' B \iota_n$.

Tightened Bounds. Finally, we discuss when the bounds can be tightened by $\frac{1}{\sqrt{n}}$.

Note that given our discussion on costs, we know that the ideal least costly graph have edges of multiplicity exactly 2. Furthermore, all multiplicities of a given edge must belong to the same walk. In particular, the best case is attained only if $p_1, ..., p_r$ are all even. On the other hand, the worst case cost is $n^{2a} p_n^a$, which is attained if the second edge are all the initial edges of a new path. For an example, see Figure A.4.

Violation of the above "optimality" conditions will result in $\alpha \geq \beta + 1$. This is sufficient to yielding the $1/\sqrt{n}$ improvement. As such, if at least one of $p_1, ..., p_r$ is odd,

$$\iota_n' B \iota_n = O_p\left(\frac{1}{\sqrt{n}}\right) \cdot O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right) \ .$$

Next, suppose that $p_1, ...p_r$ are all even. Write

$$
\begin{aligned}
&(\iota_n' B \iota_n - E\left[\iota_n' B \iota_n \,|\, U\right])^2 \\
&= \sum_{k_1=1}^{n} \sum_{k_2=1}^{n} \cdots \sum_{k_{2t+2}=1}^{n} \left(b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} - E\left[b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} \,|\, U\right]\right) \\
&\qquad \cdot \left(b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \cdots b_{k_{2t+1},k_{2t+2}} - E\left[b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \,|\, U\right]\right) \\
&= \sum_{G} \sum_{k \in r_G} \left(b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} - E\left[b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} \,|\, U\right]\right) \\
&\qquad \cdot \left(b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \cdots b_{k_{2t+1},k_{2t+2}} - E\left[b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \,|\, U\right]\right)
\end{aligned}
$$

In the above display, we are summing over all G. However, as before, the set of relevant G can be substantially restricted. Define

$$
\begin{aligned}
S_G' := E\Bigg[&\sum_{k \in r_G} \left(b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} - E\left[b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} \,|\, U\right]\right) \\
&\cdot \left(b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \cdots b_{k_{2t+1},k_{2t+2}} - E\left[b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \,|\, U\right]\right) \,|\, U\Bigg]
\end{aligned}
$$

Note that $S_G = 0 \Rightarrow S_G' = 0$. This is because $S_G = 0$ only if there G has at least one edge with multiplicity exactly 1, which will also set $S_G' = 0$.

We now show that if G is optimal, then $S_G' = 0$. Suppose G attains the optimal rate. Then every edge has multiplicity exactly 2 formed from the same walk. For such a G, $b_{k_i,k_{i+1}}$ for $(i, i+1)$ in the first walk is independent from $b_{k_{i'},k_{i'+1}}$ where $(i', i'+1)$ is in the

second walk. As such,

$$S'_G := \sum_{k \in r_G} E\left[\left(b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} - E\left[b_{k_1,k_2} b_{k_2,k_3} \cdots b_{k_t,k_t+1} \,|\, U\right]\right)\right]$$
$$\cdot E\left[\left(b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \cdots b_{k_{2t+1},k_{2t+2}} - E\left[b_{k_{t+2},k_{t+3}} b_{k_{t+3},k_{t+4}} \,|\, U\right]\right) \,|\, U\right] = 0.$$

Next, note that by the Cauchy-Schwarz inequality, that for any G', $S'_{G'} \prec S_{G'}$. Let G' be a suboptimal graph. By our study on costs of walks, $S_{G'} \prec \frac{1}{\sqrt{n}} S_G$. Now let $\mathbf{G} = \{G \,|\, S_G \neq 0\}$. Conclude that

$$E\left[\left(\iota_n' B \iota_n - E\left[\iota_n' B \iota_n \,|\, U\right]\right)^2 \,\middle|\, U\right] = \sum_{G \in \mathbf{G}} S'_G = O_p\left(\frac{1}{\sqrt{n}}\right) O_p\left(S_G\right) = O_p\left(\frac{1}{\sqrt{n}}\right) \cdot O_p\left(n^{t+1-\tau/2} p_n^{t-\tau/2}\right).$$

A.4.5.3. Proof of Lemma A.3. The proof of this lemma is based on the "Neumann trick" (see for instance, Eldridge et al. 2018, or Theorem 2 of Chen et al. 2021). We use the formulation by Cheng et al. (2021). By their Lemma 1, we have that

(A.28)
$$\frac{\lambda_1(\hat{A})}{\lambda_1(A)\left(v_1(A)' v_1(\hat{A})\right)} w' v_1(\hat{A}) = w' v_1(A) + \underbrace{\frac{w' \boldsymbol{\xi} v_1(A)}{\lambda_1(\hat{A})}}_{=:\Lambda_1} + \frac{w' \boldsymbol{\xi}^2 v_1(A)}{\left(\lambda_1(\hat{A})\right)^2}$$
$$+ \underbrace{\sum_{t=3}^{\infty} \frac{w' \boldsymbol{\xi}^t v_1(A)}{\left(\lambda_1(\hat{A})\right)^t}}_{=:\Lambda_1} + \underbrace{\sum_{r=2}^{R} \frac{\lambda_r(A)}{\lambda_1(A)} \frac{v_r(A)' v_1(\hat{A})}{v_r(A)' v_r(\hat{A})} \left\{ \sum_{t=0}^{\infty} \frac{w' \boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t} \right\}}_{=:\Lambda_3}.$$

where we used the fact that f is rank R. In the remainder of the proof, we bound Λ_1, Λ_2 and Λ_3 for $w \in \{v_1(A), \varepsilon^{(\infty)}/a_n\}$.

Bounds for $v_1(A)$. Suppose $w = v_1(A)$. We start by bounding Λ_1. For a given $\nu \in (0,1)$, choose T such that $T(1-\nu) > 4 + 2/\eta$. Then,

$$(np_n)^{-(T(1-\nu)-4)} \left(\frac{\log n}{\log\log n}\right)^{T(1-\nu)/2} \to 0 \ . \tag{A.29}$$

This is because the above condition is equivalent to

$$p_n \succ n^{-1} \left(\frac{\log n}{\log\log n}\right)^{\frac{1}{2}+\frac{2}{T(1-\nu)-4}} \ ,$$

which follows by our choice of T since p_n satisfies Equation (1.10). Observe that by Weyl's Inequality (e.g. Theorem 4.5.3 in Vershynin 2018),

$$\left\|\lambda_r(\hat{A}) - \lambda_r(A)\right\| \leq \|\boldsymbol{\xi}\| = o_p\left(np_n\right) \ , \tag{A.30}$$

the rate estimate follows from Lemma A.1. Next, note that $\frac{1}{p_n}A$ is a weighted graph obtained by sampling U on the dense graphon f. As such, by Lemma 10.16 of Lovász (2012), $\frac{\lambda_r(A)}{np_n} = \lambda_r\left(\frac{1}{p_n}A\right)/n \xrightarrow{p} \tilde{\lambda}$. In other words, w.p.a. 1, we have that $\lambda_r(\hat{A}) \geq \tilde{\lambda} np_n/2 > 0$.

Next write

$$\begin{aligned}
\left|\sum_{t=T+1}^{\infty} \frac{w'\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right| &\leq \sum_{t=T+1}^{\infty} \|w\| \cdot \left(\frac{\|\boldsymbol{\xi}\|}{\tilde{\lambda}_r np_n/2}\right)^t \cdot \|v_r(A)\| \quad \text{w.p.a. 1} \\
&= O_p\left(\|w\| \left(\sqrt{\frac{\log n}{\log\log n}} \Big/ (np_n)\right)^{T(1-\nu)/2}\right) \text{ by Lemma A.1} \\
&= O_p\left(\frac{\|w\|}{(np_n)^2}\right) \text{ by Equation (A.29).} \qquad \text{(A.31)}
\end{aligned}$$

Meanwhile,

$$E\left[\left|\sum_{t=2}^{T}\frac{w'\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right|\right] \leq E\left[\left|\sum_{t=2}^{T}\frac{w'\boldsymbol{\xi}^t v_r(A)}{\left(\tilde{\lambda}_r np_n/2\right)^t}\right|\right] \leq \sum_{t=2}^{T} E\left[\frac{\left(w'\boldsymbol{\xi}^t v_r(A)\right)^2}{\left(\tilde{\lambda}_r np_n/2\right)^{2t}}\right]^{1/2} \tag{A.32}$$

where the last inequality above follows by an application of the triangle and Cauchy-Schwarz inequalities. Since T is finite, it suffices to bound each term individually. The next part is similar to the arguments in the proof Lemma A.2.

Next note that if v is an eigenvector of A with eigenvalue λ, it must satisfy:

$$\begin{aligned}\lambda v = Mv &= \sum_{r=1}^{R}\tilde{\lambda}_r \frac{\phi_r(U)}{\sqrt{n}}\frac{\phi_r(U)'}{\sqrt{n}}v \\ &= \sum_{r=1}^{R}\tilde{\lambda}_r v_r \frac{\phi_r(U)}{\sqrt{n}}\ .\end{aligned}$$

Hence, v is a linear combination of $\phi_r(U)/\sqrt{n}$'s. By convergence of the spectrum, we know that $\limsup \lambda \leq \tilde{\lambda}_1$. Now, let Υ be the event that $|\frac{1}{\sqrt{n}}\sum_{i=1}^{r}\phi_r(U_i)\phi_s(U_i) < 1/R^2$. This happens with probability approaching 1 since R is finite. On this event, $\|v\| = 1$ implies that $|v_r| < 2$ for all r. Furthermore, observe that since $\|f\|_\infty \leq 1$, $\|\phi_r\|_\infty \leq 1$. As such, on Υ, $|v_r(A)\|_\infty \leq 2R/\sqrt{n}$.

Now, for a $U \in \Upsilon$,

$$
\begin{aligned}
&E\left[\left(v_1(A)'\boldsymbol{\xi} v_r(A)\right)^2 \,\middle|\, U\right] \\
&= \frac{1}{n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\cdots\sum_{k_{2t+2}=1}^{n}\Big\{ \left[v_1(A)\right]_{k_1}\left[v_1(A)\right]_{k_{t+1}}\left[v_r(A)\right]_{k_{t+1}}\left[v_r(A)\right]_{k_{2t+2}}\cdot \\
&\qquad\qquad E\left[\xi_{k_1,k_2}\xi_{k_2,k_3}\cdots\xi_{k_t,k_t+1}\cdot\xi_{k_{t+2},k_{t+3}}\xi_{k_{t+3},k_{t+4}}\cdots\xi_{k_{2t+1},k_{2t+2}}\,|\,U\right]\Big\} \\
&\leq \frac{16R^4}{n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\cdots\sum_{k_{2t+2}=1}^{n} E\left[\xi_{k_1,k_2}\xi_{k_2,k_3}\cdots\xi_{k_t,k_t+1}\cdot\xi_{k_{t+2},k_{t+3}}\xi_{k_{t+3},k_{t+4}}\cdots\xi_{k_{2t+1},k_{2t+2}}\,|\,U\right]
\end{aligned}
$$

where the final inequality follows from our bound on $\|v_r(A)\|_\infty$ and the fact that for all $t \leq T$,

$$E\left[\xi_{ij}^t\,|\,U\right] = p_n f(U_i,U_j)\left(1-p_n f(U_i,U_j)\right)\cdots\left(1-t\cdot p_n f(U_i,U_j)\right) \geq 0 \text{ if } p_n \leq 1/T\ .$$

By Lemma A.2,

$$\frac{1}{n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\cdots\sum_{k_{2t+2}=1}^{n} E\left[\xi_{k_1,k_2}\xi_{k_2,k_3}\cdots\xi_{k_t,k_t+1}\cdot\xi_{k_{t+2},k_{t+3}}\xi_{k_{t+3},k_{t+4}}\cdots\xi_{k_{2t+1},k_{2t+2}}\right] = \frac{1}{n^2}O\left(n^{t+2}p_n^t\right)\ .$$

As such, for $t \geq 2$,

$$E\left[\frac{\left(v_1(A)'\boldsymbol{\xi}^t v_r(A)\right)^2}{\left(\tilde{\lambda}_r n p_n/2\right)^{2t}}\right]^{1/2} = O\left(\frac{1}{\left(\sqrt{np_n}\right)^t}\right)\ . \tag{A.33}$$

Next, suppose $t=0$. Then $v_1(A)'v_r(A)=0$ since $r \neq 1$. Suppose $t=1$.

$$v_1(A)'\boldsymbol{\xi} v_r(A) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[v_1(A)\right]_i\xi_{ij}\left[v_1(A)\right]_j$$

As such,

$$E\left[\left(v_1(A)'\boldsymbol{\xi}v_r(A)\right)^2 \mid U\right] \leq \frac{16R^4}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_n f(U_i,U_j)\left(1-p_n f(U_i,U_j)\right)$$

$$= O_p\left(p_n\right) \ .$$

Together with the fact that $P(\Upsilon) \to 1$, this yields

$$\frac{v_1(A)'\boldsymbol{\xi}v_r(A)}{\tilde{\lambda}_r n p_n/2} = O_p\left(\frac{1}{n\sqrt{p_n}}\right) \ . \tag{A.34}$$

Noting that $\left(v_1(A)\right)' v_r(A) = 0$, Equations (A.31), (A.33) and (A.34) yield

$$\sum_{t=0}^{\infty}\frac{w'\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t} = O_p\left(\frac{1}{np_n}\right) \quad , \quad \sum_{t=3}^{\infty}\frac{w'\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t} = o_p\left(\frac{1}{np_n}\right) \tag{A.35}$$

Substituting Equations and (A.35) into (A.28) yields the desired bound for $v_1(A)$.

Bounds for $\varepsilon^{(\infty)}/a_n$. Next, suppose $w = \varepsilon^{(\infty)}/a_n$. The proof follows that for $v_1(A)$ up to Equation (A.32). To proceed, recall that conditional on U, $\boldsymbol{\xi} \perp\!\!\!\perp \varepsilon^{(\infty)}$. Conditioning again on the event Υ,

$$E\left[\left(\left(\frac{\varepsilon^{(\infty)}}{a_n}\right)'\boldsymbol{\xi}^t v_r(A)\right)^2 \mid U,\boldsymbol{\xi}\right]$$

$$= \frac{1}{n\cdot a_n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\cdots\sum_{k_{2t+2}=1}^{n}\left\{E\left[\varepsilon_{k_1}^{(\infty)}\varepsilon_{k_{t+2}}^{(\infty)} \mid U,\boldsymbol{\xi}\right]\left[v_r(A)\right]_{k_{t+1}}\left[v_r(A)\right]_{k_{2t+2}}\cdot\right.$$

$$\left.\xi_{k_1,k_2}\xi_{k_2,k_3}\cdots\xi_{k_t,k_t+1}\cdot\xi_{k_{t+2},k_{t+3}}\xi_{k_{t+3},k_{t+4}}\cdots\xi_{k_{2t+1},k_{2t+2}}\right\}$$

$$\leq \frac{4R^2}{n\cdot a_n^2}\sum_{k_1=1}^{n}\sum_{k_2=1}^{n}\cdots\sum_{k_{2t+2}=1}^{n}E\left[\varepsilon_{k_1}^{(\infty)}\varepsilon_{k_{t+2}}^{(\infty)} \mid U,\boldsymbol{\xi}\right]\xi_{k_1,k_2}\xi_{k_2,k_3}\cdots\xi_{k_t,k_t+1}\cdot\xi_{k_{t+2},k_{t+3}}\xi_{k_{t+3},k_{t+4}}\cdots\xi_{k_{2t+1},k_{2t+2}}$$

Recall that conditional on U, $\boldsymbol{\xi} \perp\!\!\!\perp \varepsilon^{(\infty)}$. If $k_1 \neq k_{t+2}$, we can write:

$$\begin{aligned} E\left[\varepsilon_{k_1}^{(\infty)}\varepsilon_{k_{t+2}}^{(\infty)} \,|\, U, \boldsymbol{\xi}\right] &= E\left[E\left[\varepsilon_{k_1}^{(\infty)} \,|\, \varepsilon_{k_2}^{(\infty)}, U, \boldsymbol{\xi}\right] \varepsilon_{k_{t+2}}^{(\infty)} \,|\, U, \boldsymbol{\xi}\right] \\ &= E\left[E\left[\varepsilon_{k_1}^{(\infty)} \,|\, U_{k_1}\right] \varepsilon_{k_{t+2}}^{(\infty)} \,|\, U, \boldsymbol{\xi}\right] \\ &= E\left[\varepsilon_{k_1}^{(\infty)} \,|\, U_{k_1}\right] E\left[\varepsilon_{k_{t+2}}^{(\infty)} \,|\, U_{k_{t+2}}\right] = 0 \ . \end{aligned}$$

Hence, we only need to consider sequences where $k_{t+2} = k_1$, so that

$$\begin{aligned} &E\left[\left(\left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' \boldsymbol{\xi}^t v_r(A)\right)^2 \,|\, U, \boldsymbol{\xi}\right] \\ &\leq \frac{\bar{\sigma}^2}{n \cdot a_n} \sum_{k_1=1}^{n} \cdots \sum_{k_{t+1}=1}^{n} \sum_{k_{t+3}=1}^{n} \cdots \sum_{k_{2t+2}=1}^{n} \xi_{k_1,k_2}\xi_{k_2,k_3}\cdots\xi_{k_t,k_t+1} \cdot \xi_{k_1,k_{t+3}}\xi_{k_{t+3},k_{t+4}}\cdots\xi_{k_{2t+1},k_{2t+2}} \\ &= \frac{\bar{\sigma}^2}{n \cdot a_n} \iota \boldsymbol{\xi}^{2t+1} \iota \end{aligned}$$

Taking expectations over U and $\boldsymbol{\xi}$, we have that

$$E\left[\left(\left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' \boldsymbol{\xi}^t v_r(A)\right)^2\right] = \frac{\bar{\sigma}^2}{n \cdot a_n^2} O\left(n^{t+1} p_n^{t+1/2}\right)$$

where the rate estimates again follow from Lemma A.2. The order on n is smaller than the "best case" by $n^{1/2}$ due to the fact that $2t+1$ is odd, so that at least one edge will not be optimally paired. As such, for $t \geq 1$,

$$E\left[\frac{\left((\varepsilon^{(\infty)}/a_n)' \boldsymbol{\xi}^t v_r(A)\right)^2}{\left(\tilde{\lambda}_r n p_n/2\right)^{2t}}\right]^{1/2} = O\left(\frac{p_n^{1/4}}{a_n\left(\sqrt{np_n}\right)^t}\right) = o\left(\frac{1}{a_n}\right) \ . \tag{A.36}$$

Note that when $t = 1$, the above bound implies that $\Lambda_1 = o_p\left((a_n)^{-1}\right)$ for $w = \varepsilon^{(\infty)}/a_n$. If $t = 0$,

$$\left(\frac{\varepsilon^{(\infty)}}{a_n}\right)' v_r(A) = \frac{1}{a_n\sqrt{n}} \sum_{i=1}^{n} \phi_r(U_i)\varepsilon_i^{(\infty)} = O_p\left(\frac{1}{a_n}\right) \tag{A.37}$$

Combining our last two estimates, we have that

$$\sum_{t=0}^{T} \frac{\left(\varepsilon^{(\infty)}/a_n\right)' \boldsymbol{\xi}^t v_r(A)}{\left(\tilde{\lambda}_r n p_n/2\right)^t} = O_p\left(\frac{1}{a_n}\right) \tag{A.38}$$

Let $\check{\Upsilon}$ be the event that $\lambda_r(\hat{A}) \geq \tilde{\lambda}_r n p_n/2$ and

$$\|\boldsymbol{\xi}\| \leq \sqrt{np_n}\left(\frac{k\log n}{\log\log n}\right)^{1/4},$$

where k is the constant in Lemma A.1. Then $P(\check{\Upsilon}) \to 1$ by Lemma A.1 and Equation (A.30). Furthermore,

$$\begin{aligned}
P\left(\left|\sum_{t=T+1}^{\infty} \frac{\left(\varepsilon^{(\infty)}\right)' \boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right| \leq x \,\middle|\, U, \boldsymbol{\xi}, \check{\Upsilon}\right) &\leq \frac{1}{x^2} E\left[\left(\sum_{t=T+1}^{\infty} \frac{\left(\varepsilon^{(\infty)}\right)' \boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right)^2 \,\middle|\, U, \boldsymbol{\xi}, \check{\Upsilon}\right] \\
&\leq \frac{1}{x^2}\bar{\sigma}^2 \left\|\sum_{t=T+1}^{\infty} \frac{\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right\|^2 \leq \sum_{t=T+1}^{\infty} \|\boldsymbol{\xi}\|^{2t} \\
&\leq \frac{\tilde{k}}{np_n} \text{ by Equation (A.29), on the event } \check{\Upsilon}.
\end{aligned}$$

The bound on the right hand side does not depend on U and $\boldsymbol{\xi}$ once we condition on $\check{\Upsilon}$. Hence,

$$P\left(\left|\sum_{t=T+1}^{\infty}\frac{\left(\varepsilon^{(\infty)}\right)'\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right|\leq x\right)\leq P\left(\check{\Upsilon}\right)P\left(\left|\sum_{t=T+1}^{\infty}\frac{\left(\varepsilon^{(\infty)}\right)\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}\right|\leq x\,\middle|\,\check{\Upsilon}\right)+1-P\left(\check{\Upsilon}\right)$$
$$\leq\frac{\tilde{k}}{np_n}+1-P\left(\check{\Upsilon}\right)\to 0\ .$$

Hence, we conclude that

$$\sum_{t=T+1}^{\infty}\frac{\left(\varepsilon^{(\infty)}/a_n\right)'\boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}=\frac{1}{a_n}o_p(1)=o_p\left(\frac{1}{a_n}\right)\ . \tag{A.39}$$

Next, note that

$$\begin{aligned}\left|\frac{1}{\lambda_s(A)}v_r(A)'v_s(\hat{A})\right|&\leq\left|\frac{1}{\lambda_s(A)}v_r(A)'v_s(A)\right|+\|v_r(A)\|\cdot\left\|\frac{v_s(\hat{A})-\theta v_s(A)}{2\tilde{\lambda}_s np_n}\right\|+o_p(1)\\&=\left|\frac{1}{\lambda_s(A)}v_r(A)'v_s(A)\right|+o_p(1)\ ,\end{aligned} \tag{A.40}$$

where the last equation follows because by the Davis-Kahan Inequality (Theorem 4.5.5 in Vershynin (2018)),

$$\|v_s(\hat{A})-\theta v_s(A)\|\leq\frac{\left\|\hat{A}-A\right\|}{\Delta_{\min}\cdot np_n}=o_p(1)\quad\text{ by Lemma A.1.}$$

As before, since $v_r(A)'v_1(A) = 0$ for all $r \geq 2$, we can write Λ_3

(A.41)
$$\sum_{r=2}^{R} \underbrace{\frac{\lambda_r(A)}{\lambda_1(A)} \frac{v_r(A)'v_1(\hat{A})}{v_r(A)'v_r(\hat{A})}}_{= o_p(1) \text{ by Eq. (A.40)}} \left\{ \underbrace{\sum_{t=0}^{T} \frac{\left(\varepsilon^{(\infty)}/a_n\right)' \boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}}_{= O_p\left(a_n^{-1}\right) \text{ by Eq. (A.38)}} + \underbrace{\sum_{t=T+1}^{\infty} \frac{\left(\varepsilon^{(\infty)}/a_n\right)' \boldsymbol{\xi}^t v_r(A)}{\left(\lambda_r(\hat{A})\right)^t}}_{= o_p\left(a_n^{-1}\right) \text{ by Eq. (A.39)}} \right\} = o_p\left(\frac{1}{a_n}\right) .$$

Finally, we note that by arguments identical to the above,

$$\sum_{t=1}^{\infty} \frac{\left(\varepsilon/a_n\right)' \boldsymbol{\xi}^t v_1(A)}{\left(\lambda_1(\hat{A})\right)^t} = \underbrace{\sum_{t=1}^{T} \frac{\left(\varepsilon/a_n\right)' \boldsymbol{\xi}^t v_1(A)}{\left(\lambda_1(\hat{A})\right)^t}}_{= o_p\left(a_n^{-1}\right) \text{ by Eq. (A.36)}} + \underbrace{\sum_{t=T+1}^{\infty} \frac{w'\boldsymbol{\xi}^t v_1(A)}{\left(\lambda_1(\hat{A})\right)^t}}_{= o_p\left(a_n^{-1}\right) \text{ by Eq. (A.39)}} = o_p\left(\frac{1}{a_n}\right) . \tag{A.42}$$

We conclude by remarking that from Equation (A.40),

$$\frac{\lambda_1(\hat{A})}{\lambda_1(A)\left(v_1(A)'v_1(\hat{A})\right)} = \frac{\lambda_1(A)}{\lambda_1(A)\left(v_1(A)'v_1(A)\right)} + o_p(1) ,$$

so that the LHS of Equation (A.28) also converges in probability to $w'v_1(\hat{A})$.

APPENDIX B

Appendix to Chapter 2

B.1. Proof for Theorem 2.1

We first write:

$$\begin{aligned}
&\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i - W_i\hat{\Pi}_j\right)\hat{U}_i \\
=&\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i - W_i\hat{\Pi}_j\right)U_i && \text{(B.1)} \\
&-\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i - W_i\hat{\Pi}_j\right)^2\left(\hat{\beta}-\beta\right) && \text{(B.2)} \\
&-\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i - W_i'\hat{\Pi}_j\right)\left(W_i'\hat{\Pi}_j(\hat{\beta}-\beta)+W_i'(\hat{\gamma}-\gamma)\right) . && \text{(B.3)}
\end{aligned}$$

We analyse parts (B.1) and (B.3) in turn. Part (B.2) is handled using our worst case bound. Starting with (B.1):

$$\begin{aligned}
&\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i - W_i\hat{\Pi}_j\right)U_i \\
&= \frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}\left(X_i - W_i\Pi_j\right)U_i - \left(\hat{\Pi}_j - \Pi_j\right)\frac{1}{\sqrt{n_j}}\sum_{i\in\mathcal{I}_j}W_iU_i \\
&= \mathrm{N}\left(0,\sigma_j^2\right) + o_p(1)O_p(1) ,
\end{aligned}$$

where the last equation follows from assumption 2.2. Next for term (B.3), note that:

$$\begin{aligned}
&\frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} W_i \left(X_i - W_i' \hat{\Pi}_j \right) \\
&= \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} W_i \left(X_i - W_i \Pi_j \right) - \sqrt{n_j} \left(\hat{\Pi}_j - \Pi_j \right) \frac{1}{n_j} \sum_{i \in \mathcal{I}_j} W_i W_i' = O_p(1) \ .
\end{aligned}$$

As such, by assumption (2.2),

$$\left(\hat{\Pi}_j' (\hat{\beta} - \beta) + (\hat{\gamma} - \gamma) \right)' \frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} W_i \left(X_i - W_i' \hat{\Pi}_j \right) = o_p(1) \ . \tag{B.4}$$

Our analysis of terms (B.1) and (B.3) show that setting $\lambda = \hat{\beta} - \beta$, we have that

$$\hat{S}_n (\hat{\beta} - \beta) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \Sigma) \ .$$

Furthermore, under the null hypothesis, Σ is diagonal since

$$\frac{1}{\sqrt{n_j}} \sum_{i \in \mathcal{I}_j} \left(X_i - W_i \Pi_j \right) U_i$$

is independent across sub-clusters.

Maintaining the assumption that $\hat{\beta} - \beta$ is known, we show that the requirements for Theorem 3.1 in Canay et al. (2017a) are met.

i. $\hat{S}_n (\hat{\beta} - \beta) \xrightarrow{d} S$ by the analysis above.

ii. By symmetry of S about 0 and the fact that Σ is diagonal, it is immediate that gS has the same distribution as S under the null hypothesis.

iii. For all $g \neq g'$, $P(T(gS) \neq T(g'S)) = 1$. This is because for a given component j on which $g_j \neq g'_j$, $S_j = -S_j$ if and only if $S_j = 0$, which occurs with probability 0.

Hence, we have that:

$$E\left[\mathbf{1}\left\{p(\hat{S}_n(\hat{\beta} - \beta)) \leq \alpha\right\}\right] \to \alpha \ .$$

This test is conservative since we break ties in favour of not rejecting the null-hypothesis. It is then immediate that:

$$\limsup_{n\to\infty} E\left[\mathbf{1}\left\{\sup_{\lambda\in\mathbf{R}} p(\hat{S}_n(\lambda)) \leq \alpha\right\}\right] \leq \limsup_{n\to\infty} E\left[\mathbf{1}\left\{p(\hat{S}_n(\hat{\beta} - \beta)) \leq \alpha\right\}\right] = \alpha \ . \qquad \square$$

B.2. Inference with Unnecessarily Coarse Clusters

In this section, we demonstrate by simulation the problems with using unnecessarily coarse clusters for inference. Consider the model:

$$Y_{t,j,k} = \beta + \frac{1}{\sqrt{1-\phi^2}} U_{t,j,k} \ ,$$

$$U_{t,j,k} = \phi U_{t-1,j,k} + \varepsilon_{t,j,k}, \quad \varepsilon_{t,j,k} \overset{\text{iid}}{\sim} N(0,1), \quad U_{1,j,k} \overset{\text{iid}}{\sim} N(0,1)$$

where t is an observation from fine cluster j in coarse cluster k. Here, individuals in the same fine cluster j are dependent due to $U_{t,j,k}$, but individuals in different fine clusters are independent. Clustering at both the fine and coarse levels are therefore valid, though as we show, unncessarily coarse clustering will lead to issues. Suppose there are 4 coarse clusters, 12 fine clusters in each coarse clusters, and 100 observations per fine cluster. Further set $\phi = 0.25$ and $\beta = 1$. Suppose want to test the hypothesis $\beta = \beta_0$ at 5% level

of significance. The most popular options are CCE-based tests, ARTs, or wild bootstrap based tests, all of which can be implemented with clustering at either the k level, or at the j level.

Figure B.1 presents the rejection rates from each of these tests as we vary β_0 from 1 to 0 (equivalently, as $1 - \beta_0$ varies from 0 to 1). The left panel pertains to CCE-based tests. Under the null hypothesis, when $\beta_0 = 1$ (i.e when $1 - \beta_0 = 0$), the test using coarse clustering rejects over 12% of the time, more than twice the nominal size. However, same test controls size well with fine clustering. The middle panel presents results from conservative ARTs that do not perform random tie-breaking. The test controls size and has good power with fine clustering. With coarse clustering, however, the test never rejects. This is because conservative ARTs can reject only when the size is smaller than 2^{q-1}, where q is the number of clusters. This is an extreme example of how using coarse clusters could dramatically reduce power. The right panel presents results from wild bootstrap-based tests (Cameron et al. 2008a) commonly used when the number of clusters is small. Our setting satisfies the requirements of Canay et al. (2019), so that we expect tests based on either level of clustering to control size, as they indeed do. However, the test with coarse clustering has much lower power than with fine clustering. All in all, our simulation shows that inference using unnecessarily coarse levels of clustering leads to problems.

B.3. Residualized Null Hypothesis

In this section, we explain why a researcher conducting inference on β only has to test the residualized hypothesis in equation (2.4). As in standard notation, let W be

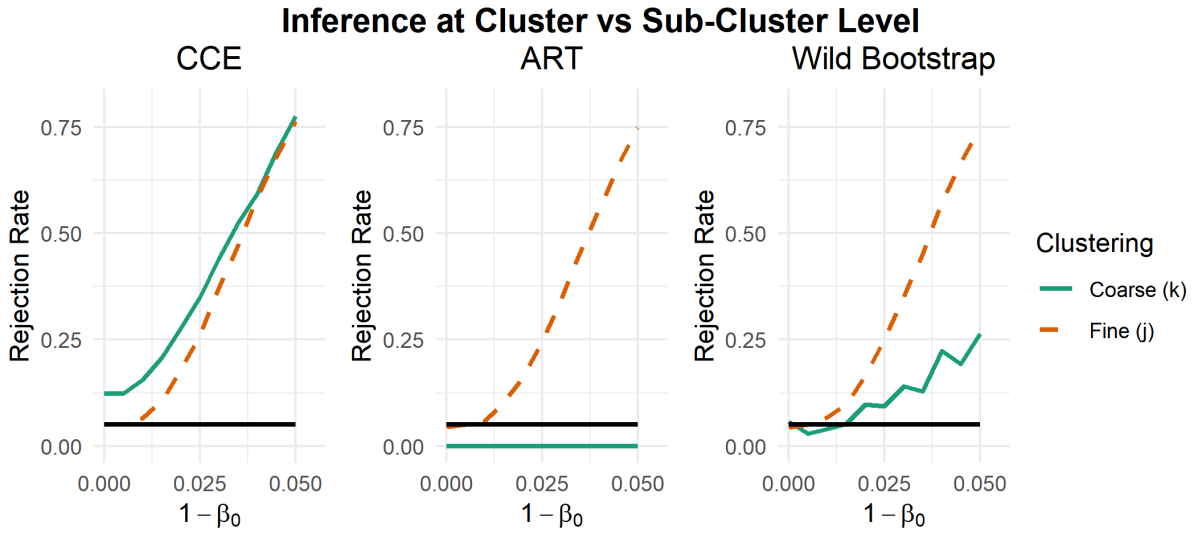

Figure B.1. Rejection rates from implementing CCE-based tests, ARTs or wild bootstrap-based tests in a simple model (see text). We assume 4 coarse clusters, 12 fine clusters per coarse clusters and 100 observations per fine cluster. Simulation are repeated 1000 times. The black line indicates the nominal size of the test (5%).

the matrix with W_i' in its i^{th} row. In the following we assume that W has full rank for convenience.[1] Write:

$$\begin{aligned} P_W Y &= W(W'W)^{-1}W'Y \\ &= W(W'W)^{-1}W'\left(X\hat{\beta} + W'\hat{\gamma} + \hat{U}\right) \\ &= W\hat{\gamma} + W\hat{\Pi}\hat{\beta}\ , \end{aligned}$$

where $\hat{\beta}$ and $\hat{\gamma}$ are full sample OLS estimates from equation (2.1) and $\hat{U}$ are the residuals from the same regression. $\hat{\Pi} = (W'W)^{-1}W'X$ is one of the possible consistent estimators

[1]The case where W is rank deficient is similar except $\hat{\Pi}$ might not have an explicit formula. However, this does not change the properties of the projection residuals.

for Π when W has full rank. By the Frisch-Waugh-Lovell theorem, we can then write:

$$\begin{aligned}
\hat{\beta} &= \left((X - P_W X)'(X - P_W X)\right)^{-1}(X - P_W X)'(Y - P_W Y) \\
&= \frac{\sum_{i=1}^{n}\left(X_i - W_i\hat{\Pi}\right)\left(Y_i - W_i\hat{\gamma} - W_i\hat{\Pi}\hat{\beta}\right)}{\sum_{i=1}^{n}\left(X_i - W_i\hat{\Pi}\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(X_i - W_i\hat{\Pi}\right)\left((X_i - W_i\hat{\Pi})\beta + W_i\hat{\Pi}\beta + W_i\gamma + U_i - W_i\hat{\gamma} - W_i\hat{\Pi}\hat{\beta}\right)}{\sum_{i=1}^{n}\left(X_i - W_i\hat{\Pi}\right)^2} \\
&= \beta + \frac{\sum_{i=1}^{n}\left(X_i - W_i\hat{\Pi}\right)\left(U_i - W_i\hat{\Pi}(\hat{\beta} - \beta) - W_i(\hat{\gamma} - \gamma)\right)}{\sum_{i=1}^{n}\left(X_i - W_i\hat{\Pi}\right)^2} .
\end{aligned}$$

Then, by the same argument that leads to equation (B.4), we have that under consistency of $\hat{\gamma}$, $\hat{\beta}$ and $\sqrt{n}$-consistency of $\hat{\Pi}$:

$$\sqrt{n}\left(\hat{\beta} - \beta\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - W_i\Pi\right)U_i}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - W_i\Pi\right)^2} + o_p(1) .$$

Or, if we are estimating $\hat{\beta}$ cluster-by-cluster,

$$\sqrt{n_j}\left(\hat{\beta}_j - \beta\right) = \frac{\frac{1}{\sqrt{n_j}}\sum_{i=1}^{n_j}\left(X_i - W_i\Pi_j\right)U_i}{\frac{1}{n_j}\sum_{i=1}^{n}\left(X_i - W_i\Pi_j\right)^2} + o_p(1) .$$

As such, the asymptotic distribution of the $\hat{\beta}/\hat{\beta}_j$'s depend only on $Z_i = (X_i - W_i\Pi)U_i$. If there is no dependence in Z_i across the sub-clusters, approximate randomization test at the sub-cluster level yields valid inference.

Our asymptotic framework takes the number of sub-clusters as fixed. However, if there is no dependence across sub-clusters in the Z_i's, then provided that the usual regularity

conditions hold (e.g. in Hansen and Lee (2019b)), we have that as the number of sub-cluster grows to infinity, inference with CCE clustered at the sub-cluster level also leads to correct inference.

B.4. Restricted Heterogeneity implied by Assumption 2

In this section we explain how the assumption that $S_n \to N(0,\Sigma)$ requires that $\min_j n_j \to \infty$, but does not place any restrictions on the relative rates at which each $n_j \to \infty$. We do this by way of an example. Assume that each $S_{n,j}$ have $2+\delta$ moments for $\delta > 0$ and is weakly dependent in the sense of Doukhan and Louhichi (1999). This is a general form of dependence which includes strongly mixing sequences and Bernoulli shifts as special cases. Nze and Doukhan (2004) argues for usefulness in econometrics. We show that under our assumption, $S_n \to N(0,\Sigma)$ as long as $\min_j n_j \to \infty$, even if $\frac{n_{j'}}{\min_j n_j} \to \infty$ for some $j' \in [q]$.

Consider the Cramer-Wold theorem, which gives us that $S_n \xrightarrow{d} N(0,\Sigma)$ under the null hypothesis if and only if for all $\lambda \in \mathbf{R}^q$ we have

$$E\left[\exp\left(it\lambda' S_n\right)\right] = \prod_{j=1}^{q} E\left[\exp\left(it\lambda_j S_{n,j}\right)\right] \to \prod_{j=1}^{q} E\left[\exp\left(-it\frac{\lambda_j^2\sigma_j^2 t^2}{2}\right)\right] .$$

Since characteristic functions are bounded by 1,

$$\left|\prod_{j=1}^{q} E\left[\exp\left(it\lambda_j S_{n,j}\right)\right] - \prod_{j=1}^{q} E\left[\exp\left(-it\frac{\lambda_j^2\sigma_j^2 t^2}{2}\right)\right]\right| \leq \sum_{j=1}^{q}\left|E\left[\exp\left(it\lambda_j S_{n,j}\right)\right] - E\left[\exp\left(-it\frac{\lambda_j^2\sigma_j^2 t^2}{2}\right)\right]\right| .$$

Proposition 7.1 in Dedecker et al. (2007) yields:

$$\left|E\left[\exp\left(it\lambda_j S_{n,j}\right)\right] - E\left[\exp\left(-it\frac{\lambda_j^2\sigma_j^2 t^2}{2}\right)\right]\right| \leq C n^{-c_j^*}$$

where $c_j^* > 0$ depends on the amount of dependence within sub-cluster j. Define $c = \min_j c_j$. Then we can write

$$\left| \prod_{j=1}^{q} E\left[\exp\left(it\lambda_j S_{n,j}\right)\right] - \prod_{j=1}^{q} E\left[\exp\left(-it\frac{\lambda_j^2\sigma_j^2 t^2}{2}\right)\right] \right| = O\left(\left(\min_{j\in[q]} n_j\right)^{-c}\right)$$

Hence, we have weak convergence of S_n to S as long as the slowest term converges. The relative rates at which the n_j's grow to infinity are not restricted.

For comparison, in OLS on units with cluster dependence, observations are not standardized within each cluster. As a result, the contribution of each cluster to "numerator" in the $\hat{\beta} - \beta$ is $X_j'U_j$ rather than $\frac{X_j'U_j}{\sqrt{n_j}}$, where X_j is the stacked covariates for units in cluster j and U_j are their stacked linear regression errors. Hence, large clusters have outsize influence in estimation and inference. Restricting the influence of each cluster motivates the restricted heterogeneity assumptions in Hansen and Lee (2019b) for example.

B.5. Cluster Statistics for Gneezy et al. (2019)

School	Track	Group	Size
US 1	Honors	1	325
		7	350
		11	625
		27	725
	Regular	2	300
		3	325
		4	400
		6	250
		9	225
		10	300
		12	250
		13	350
		14	375
		15	500
		17	375
		22	275
		24	300
		26	325
		28	275
	Others	5	225
		8	250
		18	400
		19	150
		23	450
		25	150
		29	25
		30	25
US 2	Honors	-	46 groups of 25
	Regular	-	60 groups of 25

School	Year	Group	Size
Shanghai 1	2016	9992	750
		9993	750
Shanghai 2	2016	9994	1000
		9995	1000
	2018	-	128 groups of 25
Shanghai 3	2016	9996	975
		9997	975
		9998	800
		9999	750
	2018	-	122 groups of 25
Shanghai 4	2018	-	126 groups of 25

Table B.1. Cluster Structure for US and Shanghai Schools.

APPENDIX C

Appendix to Chapter 4

C.1. Proofs

C.1.1. Proofs of Propositions 4.1 and 4.2

C.1.1.1. Some preliminaries and common machinery for the proofs. Our proofs in this section require some common machinery. Let $\{(Y_i^*(0), Y_i^*(1))\}_{i=1}^{\infty}$ be i.i.d. random vectors with distribution equal to that of $(Y(0) - \mu(0), Y(1) - \mu(1))$ that satisfy

$$\{(Y_i^*(0), Y_i^*(1))\}_{i=1}^{\infty} \perp\!\!\!\perp \left\{ \left\{ \widetilde{Y}_i(0), \widetilde{Y}_i(1), \widetilde{A}_i \right\}_{i=1}^{\infty}, \{Y_i(0), Y_i(1), U_i\}_{i=1}^{\infty} \right\}.$$

þC.3 provides detailed arguments showing that

$$\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) \stackrel{d}{=} \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n\overline{A}_{\widetilde{p}}} Y_i^*(1)}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{\sqrt{n}} \sum_{i=n\overline{A}_{\widetilde{p}}+1}^{n} Y_i^*(0)}{1 - \overline{A}_{\widetilde{p}}}. \tag{C.1}$$

Consider the partial-sums process

$$\mathbf{T}_n(u) = \begin{pmatrix} T_n(0,u) \\ T_n(1,u) \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor u \cdot n \rfloor} \begin{pmatrix} Y_i^*(0) \\ Y_i^*(1) \end{pmatrix} \qquad \forall u \in [0,1]. \tag{C.2}$$

By hypothesis, $\{(Y_i^*(0), Y_i^*(1))\}_{i=1}^{\infty}$ are mean zero, have finite variance and are i.i.d. across $i \in \mathbb{N}$. Denote the space of all essentially bounded functions on $[0,1]$ endowed with the essential supremum norm by $\ell^{\infty}([0,1])$. By a two-dimensional variant of Donsker's

Functional Central Limit Theorem (see for instance Whitt (2002, Theorem 4.3.5)), $\mathbf{T}_n^*(\cdot)$ converges weakly in $\ell^\infty([0,1])^2$ to a two-dimensional scaled Brownian motion $\mathbf{T}_\infty(\cdot)$

$$\mathbf{T}_\infty(u) = \begin{pmatrix} T_\infty(u,0) \\ T_\infty(u,1) \end{pmatrix} = \Sigma^{\frac{1}{2}} B(u) \qquad \forall u \in [0,1] \tag{C.3}$$

where $B(\cdot)$ is a two-dimensional standard Brownian motion and $\Sigma^{\frac{1}{2}}$ is the unique symmetric matrix satisfying $\Sigma = \left(\Sigma^{\frac{1}{2}}\right)' \Sigma^{\frac{1}{2}}$. We can write the vector comprised of the numerators in (C.1) as

$$U_n = \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n\overline{A}_{\widetilde{p}}} Y_i^*(1) \\ \frac{1}{\sqrt{n}} \sum_{i=n\overline{A}_{\widetilde{p}}+1}^{n} Y_i^*(0) \end{pmatrix} = \begin{pmatrix} T_n\left(1, \overline{A}_{\widetilde{p}}\right) \\ T_n(0,1) - T_n\left(0, \overline{A}_{\widetilde{p}}\right) \end{pmatrix} \tag{C.4}$$

so that by (C.1),

$$\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) \stackrel{d}{=} \begin{bmatrix} \frac{1}{\overline{A}_{\widetilde{p}}}, & -\frac{1}{1-\overline{A}_{\widetilde{p}}} \end{bmatrix} \cdot U_n \ . \tag{C.5}$$

C.1.1.2. Proof of Proposition 4.1.

Proof of Proposition 4.1. By Lemma C.2, $\overline{A}_{\widetilde{p}} \stackrel{\text{a.s.}}{\to} p_*$ as $m, n \to \infty$. Furthermore, van der Vaart (2000) Theorems 18.10 (v) and 18.11 imply that U_n in (C.4) satisfies

$$U_n \stackrel{d}{\to} U := \begin{pmatrix} T_\infty(1, p_*) \\ T_\infty(0,1) - T_\infty(0, p_*) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} p_* \cdot \sigma^2(1) & 0 \\ 0 & (1-p_*) \cdot \sigma^2(0) \end{pmatrix} \right). \tag{C.6}$$

where the covariance matrix is derived using the fact that a standard Brownian motion has independent increments. By (C.6), $\overline{A}_{\widetilde{p}} \xrightarrow{\text{a.s.}} p_*$ and Slutsky's Theorem, as $m, n \to \infty$,

$$\begin{aligned}
\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) &\overset{d}{=} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n\overline{A}_{\widetilde{p}}} Y_i^*(1)}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{\sqrt{n}}\sum_{i=n\overline{A}_{\widetilde{p}}+1}^{n} Y_i^*(0)}{1-\overline{A}_{\widetilde{p}}} \\
&= \begin{bmatrix} \frac{1}{\overline{A}_{\widetilde{p}}} & -\frac{1}{1-\overline{A}_{\widetilde{p}}} \end{bmatrix} \cdot U_n \\
&\xrightarrow{d} \begin{bmatrix} \frac{1}{p_*} & -\frac{1}{1-p_*} \end{bmatrix} \cdot U \\
&\sim \mathcal{N}\left(0, \frac{\sigma^2(1)}{p_*} + \frac{\sigma^2(0)}{1-p_*}\right).
\end{aligned}$$

□

C.1.1.3. Proof of Proposition 4.2.

Proof of Proposition 4.2. Note that the process $\mathbf{T}_n$ in (C.2) is independent to $\overline{A}_{\widetilde{p}}$ for every $n \in \mathbb{N}$. Additionally, $\overline{A}_{\widetilde{p}} \xrightarrow{\text{p}} \widetilde{p}$ (by Lemma C.2) and $\mathbf{T}_n \xrightarrow{\text{d}} \mathbf{T}_\infty$ as $n \to \infty$, where $\xrightarrow{\text{d}}$ denotes weak convergence in $\ell^\infty([0,1])^2$. This implies that (see for instance van der vaart and Wellner (1996, Example 1.4.6))

$$\begin{pmatrix} \overline{A}_{\widetilde{p}} \\ \mathbf{T}_n(\cdot) \end{pmatrix} \xrightarrow{\text{d}} \begin{pmatrix} \widetilde{p} \\ \mathbf{T}_\infty(\cdot) \end{pmatrix} \qquad \text{as } n \to \infty \tag{C.7}$$

where $\widetilde{p}$ and $\mathbf{T}_\infty$ are independent and $\xrightarrow{\text{d}}$ is in the sense of weak convergence in $[0,1] \times \ell^\infty([0,1])^2$. Equation (C.7) alongside Lemma C.5 and van der Vaart (2000) Theorem 18.11

imply that in this case, as $n \to \infty$, U_n in equation (C.4)

$$U_n \xrightarrow{\mathrm{d}} \widetilde{U} := \begin{pmatrix} T_\infty\left(1, \widetilde{p}\right) \\ T_\infty(0,1) - T_\infty\left(0, \widetilde{p}\right). \end{pmatrix} \tag{C.8}$$

We should note that measurability of $\widetilde{U}$ is derived in Lemma C.6. Using (C.5) in combination with (C.8), we get

$$\begin{aligned} \sqrt{n}\left(\widehat{\theta}_{\widetilde{p}} - \theta\right) &\overset{\mathrm{d}}{=} \begin{bmatrix} \frac{1}{\widehat{A}_{\widetilde{p}}} & -\frac{1}{1-\widehat{A}_{\widetilde{p}}} \end{bmatrix} \cdot U_n \\ &\xrightarrow{\mathrm{d}} \begin{bmatrix} \frac{1}{\widetilde{p}} & -\frac{1}{1-\widetilde{p}} \end{bmatrix} \cdot \widetilde{U} \\ &= \frac{1}{\widetilde{p}} T_\infty\left(1, \widetilde{p}\right) - \frac{1}{1-\widetilde{p}}\left[T_\infty(0,1) - T_\infty\left(0, \widetilde{p}\right)\right] \\ &=: \mathcal{L}_m. \end{aligned}$$

The distribution of $\mathcal{L}_m$ can be derived as a corollary of Lemma C.6, but we include the direct derivation here for completeness. To derive the distribution of $\mathcal{L}_m$ in closed form, notice that $\mathcal{L}_m | \widetilde{p} \sim \mathcal{N}\left(0, \frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right)$ so that

$$\mathbb{P}\left(\mathcal{L}_m \leq t | \widetilde{p}\right) = \Phi\left(\frac{t}{s\left(\widetilde{p}\right)}\right) \qquad \text{where } s^2\left(\widetilde{p}\right) = \frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}.$$

By the Law of Total Probability, it follows that

$$\mathbb{P}\left(\mathcal{L}_m \leq t\right) = \int_0^1 \Phi\left(\frac{t}{s\left(p\right)}\right) \, G_m(\mathrm{d}p),$$

where $G_m(\cdot)$ is the distribution of $\widetilde{p}$. □

C.1.1.4. Auxiliary lemmas.

Lemma C.1. þ Suppose Assumption 4.2 holds and additionally that, there is some $p_1 \in (0,1)$ such that $\frac{1}{m}\sum_{i=1}^{m} \widetilde{A}_i \xrightarrow{\text{p}} \mathbb{P}\left(\widetilde{A}_1 = 1\right) = p_1$ as $m \to \infty$. Then $\widetilde{p} \xrightarrow{\text{p}} p_*$ as $m \to \infty$.

Proof. Let $a \in \{0,1\}$ be given. Note that for any Borel function $h : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}[|h(Y(a))|] < \infty$, as $m \to \infty$,

$$\frac{1}{m}\sum_{i=1}^{m} h\left(\widetilde{Y}_i\right) \mathbb{I}\left\{\widetilde{A}_i = a\right\} = \frac{1}{m}\sum_{i=1}^{m} h\left(\widetilde{Y}_i(a)\right) \mathbb{I}\left\{\widetilde{A}_i = a\right\} \xrightarrow{\text{a.s.}} \mathbb{E}\left[h\left(\widetilde{Y}_1(a)\right) \mathbb{I}\left\{\widetilde{A}_1 = a\right\}\right]$$

Under Assumption 4.2, $\widetilde{Y}_1(a) \perp\!\!\!\perp \widetilde{A}_1$ and $\widetilde{Y}_1(a) \sim Y(a)$ so that,

$$\frac{1}{m}\sum_{i=1}^{m} h\left(\widetilde{Y}_i\right) \mathbb{I}\left\{\widetilde{A}_i = a\right\} \xrightarrow{\text{a.s.}} \mathbb{E}\left[h(Y(a))\right] \cdot p_1^a \cdot (1-p_1)^{1-a} \quad \text{as } m \to \infty$$

The hypothesis of the question gives us that $\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left\{\widetilde{A}_i = a\right\} \xrightarrow{\text{p}} p_1^a \cdot (1-p_1)^{1-a} \in (0,1)$. Thus, by the Continuous Mapping Theorem, as $m \to \infty$

$$\frac{\frac{1}{m}\sum_{i=1}^{m} h\left(\widetilde{Y}_i\right) \mathbb{I}\left\{\widetilde{A}_i = a\right\}}{\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left\{\widetilde{A}_i = a\right\}} \xrightarrow{\text{a.s.}} \mathbb{E}\left[h(Y(a))\right]$$

Setting h equal to the maps $y \mapsto y$ and $y \mapsto y^2$, we get respectively that

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \widetilde{Y}_i \mathbb{I}\left\{\widetilde{A}_i = a\right\}}{\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left\{\widetilde{A}_i = a\right\}} \xrightarrow{\text{p}} \mathbb{E}[Y(a)]$$

$$\frac{\frac{1}{m}\sum_{i=1}^{m} \widetilde{Y}_i^2 \mathbb{I}\left\{\widetilde{A}_i = a\right\}}{\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left\{\widetilde{A}_i = a\right\}} \xrightarrow{\text{p}} \mathbb{E}\left[Y(a)^2\right]$$

Combining these with the Continuous Mapping Theorem again, it follows that as $m \to \infty$

$$\widetilde{\sigma}_m^2(a) = \frac{1}{\sum_{i=1}^m \mathbb{I}\left\{\widetilde{A}_i = a\right\} - 1} \sum_{i=1}^m \left(\widetilde{Y}_i \mathbb{I}\left\{\widetilde{A}_i = a\right\} - \frac{1}{\sum_{i=1}^m \mathbb{I}\left\{\widetilde{A}_i = a\right\}} \sum_{i=1}^m \widetilde{Y}_i \mathbb{I}\left\{\widetilde{A}_i = a\right\} \right)^2$$
$$\xrightarrow{\text{p}} \sigma^2(a)$$

By another application of the Continuous Mapping Theorem,

$$\widetilde{p} = \frac{\widetilde{\sigma}_m(1)}{\widetilde{\sigma}_m(1) + \widetilde{\sigma}_m(0)} \xrightarrow{\text{p}} \frac{\sigma(1)}{\sigma(1) + \sigma(0)} = p_*.$$

□

Lemma C.2. Let $\overline{A}_{\widetilde{p}} = \frac{1}{n}\sum_{i=1}^n A_{\widetilde{p},i}$. Under Assumptions 4.2, 4.3, equation (4.3) and fixed $m \in \mathbb{N}$, $\overline{A}_{\widetilde{p}} \xrightarrow{\text{p}} \widetilde{p}$ as $n \to \infty$. Additionally, $\overline{A}_{\widetilde{p}} \xrightarrow{\text{p}} p_*$ as $m, n \to \infty$.

Proof. Note that given any fixed $m \in \mathbb{N}$, $A_{\widetilde{p},i}|\widetilde{p} \sim$ Bernoulli $(\widetilde{p})$ independently across $i \in \mathbb{N}$. By De Finetti's Theorem, $\{A_{\widetilde{p},i}\}_{i=1}^\infty$ forms an exchangeable sequence of random variables. By the Strong Law of Large Numbers for exchangeable sequences (see for instance Schervish (1995, Theorem 1.62) or Kingman (1978)), it follows that $\overline{A}_{\widetilde{p}} \xrightarrow{\text{a.s.}} \widetilde{p}$ as $n \to \infty$. Next, we consider the case with both $m, n \to \infty$. Using the triangle inequality,

$$\left|\overline{A}_{\widetilde{p}} - p_*\right| \leq \left|\overline{A}_{\widetilde{p}} - \widetilde{p}\right| + \left|\widetilde{p} - p_*\right| \leq \sup_{p\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{U_i \leq p\right\} - p \right| + \left|\widetilde{p} - p_*\right|.$$

By Lemma C.1, $|\widetilde{p} - p_*| \xrightarrow{\text{p}} 0$ as $m \to \infty$. An immediate consequence of the Glivenko-Cantelli Theorem is that $\sup_{p\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{U_i \leq p\right\} - p\right| \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. □

Lemma C.3. Under Assumptions 4.2, 4.3 and equation (4.3), we can write

$$\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}}-\theta\right) \stackrel{\mathrm{d}}{=} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n\overline{A}_{\widetilde{p}}} Y_i^*(1)}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{\sqrt{n}}\sum_{i=n\overline{A}_{\widetilde{p}}+1}^{n} Y_i^*(0)}{1-\overline{A}_{\widetilde{p}}}$$

where $\{(Y_i^*(0), Y_i^*(1))\}_{i=1}^{\infty}$ are i.i.d. random vectors with distribution equal to that of $(Y(0)-\mu(0), Y(1)-\mu(1))$ and

$$\{(Y_i^*(0), Y_i^*(1))\}_{i=1}^{\infty} \perp\!\!\!\perp \left\{\left\{\widetilde{Y}_i(0), \widetilde{Y}_i(1), \widetilde{A}_i\right\}_{i=1}^{\infty}, \{Y_i(0), Y_i(1), U_i\}_{i=1}^{\infty}\right\}. \tag{C.9}$$

Proof. We can rewrite $\widehat{\theta}_{\widetilde{p}}$ using the potential outcomes as is standard:

$$\widehat{\theta}_{\widetilde{p}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(1)A_{\widetilde{p},i}}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(0)\left(1-A_{\widetilde{p},i}\right)}{1-\overline{A}_{\widetilde{p}}}.$$

where $\overline{A}_{\widetilde{p}} = \frac{1}{n}\sum_{i=1}^{n} A_{\widetilde{p},i}$. Furthermore, since $\theta = \mu(1)-\mu(0)$,

$$\begin{aligned}\widehat{\theta}_{\widetilde{p}}-\theta &= \left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(1)A_{\widetilde{p},i}}{\overline{A}_{\widetilde{p}}} - \mu(1)\right) - \left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i(0)\left(1-A_{\widetilde{p},i}\right)}{1-\overline{A}_{\widetilde{p}}} - \mu(0)\right) \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i(1)-\mu(1)\right]A_{\widetilde{p},i}}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i(0)-\mu(0)\right]\left(1-A_{\widetilde{p},i}\right)}{1-\overline{A}_{\widetilde{p}}}.\end{aligned}$$

Thus,

$$\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}}-\theta\right) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[Y_i(1)-\mu(1)\right]A_{\widetilde{p},i}}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[Y_i(0)-\mu(0)\right]\left(1-A_{\widetilde{p},i}\right)}{1-\overline{A}_{\widetilde{p}}}.$$

Note that the distribution $\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}}-\theta\right)$ is invariant to permutation of the sample indices. Hence the distribution of $\sqrt{n}\left(\widehat{\theta}_{\widetilde{p}}-\theta\right)$ does not change if we reorder the sample indices to have $i = 1, \ldots, n\overline{A}_{\widetilde{p}}$ correspond to the observations in the treatment group and $i = n\overline{A}_{\widetilde{p}}+1, \ldots, n$ correspond to the observations in the control group. Notice also, that under

Assumptions 4.2, 4.3, and (C.9), the resulting permuted sum has the same distribution as

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n\overline{A}_{\widetilde{p}}} Y_i^*(1)}{\overline{A}_{\widetilde{p}}} - \frac{\frac{1}{\sqrt{n}}\sum_{i=n\overline{A}_{\widetilde{p}}+1}^{n} Y_i^*(0)}{1-\overline{A}_{\widetilde{p}}}.$$

□

Lemma C.4. Under Assumptions 4.1, 4.2 and 4.3, if m stays fixed as $n \to \infty$, $\widehat{\sigma}^2_{\widetilde{p}}(a) \xrightarrow{\text{p}} \sigma^2(a)$.

Proof. Recall that we define

$$\widehat{\sigma}^2_{\widetilde{p}}(a) = \frac{1}{n_{\widetilde{p},a}-1}\sum_{i=1}^{n}\left(Y_{\widetilde{p},i}\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\} - \frac{1}{n_{\widetilde{p},a}}\sum_{i=1}^{n} Y_{\widetilde{p},i}\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\}\right)^2$$
$$n_{\widetilde{p},a} = \sum_{i=1}^{n}\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\}$$

By Lemma C.2, it follows that

$$\frac{n_{\widetilde{p},a}}{n} \xrightarrow{\text{p}} \widetilde{p}^a\left(1-\widetilde{p}\right)^{1-a}. \tag{C.10}$$

By (4.3), we have that for each $i \in \mathbb{N}$,

$$Y_{\widetilde{p},i}\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\} = Y_i(a)\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\} \qquad Y^2_{\widetilde{p},i}\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\} = Y_i^2(a)\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\}$$

By the Continuous mapping theorem, the proof is completed if we show both

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Y_i(a)\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\} &\xrightarrow{\text{p}} \mu(a)\cdot\widetilde{p}^a\left(1-\widetilde{p}\right)^{1-a}\\ \frac{1}{n}\sum_{i=1}^{n} Y_i(a)^2\mathbb{I}\left\{A_{\widetilde{p},i}=a\right\} &\xrightarrow{\text{p}} \left(\sigma^2(a)+\mu(a)^2\right)\cdot\widetilde{p}^a\left(1-\widetilde{p}\right)^{1-a}\end{aligned} \tag{C.11}$$

Similarly to the arguments in the proof of Lemma C.2, conditional on $\widetilde{p}$,

$$\{Y_i(a)\mathbb{I}\{A_{\widetilde{p},i}=a\}\}_{i=1}^{\infty} \text{ and } \left\{Y_i(a)^2\mathbb{I}\{A_{\widetilde{p},i}=a\}\right\}_{i=1}^{\infty}$$

are i.i.d. sequences. Hence both are also exchangeable sequences (see Schervish (1995, Problem 4, page 73)). Thus, (C.11) follows from þ4.1 and the Strong Law of Large Numbers for exchangeable sequences (see for instance Schervish (1995, Theorem 1.62) or Kingman (1978)). Additionally, convergence almost surely implies convergence in probability. The conclusion of the lemma then follows by (C.10), (C.11) and the Continuous Mapping Theorem. □

For the next lemma, for $k \in \mathbb{N}$ and let $C([0,1])$ denote the space of all continuous real-valued functions on $[0,1]$ endowed with the supremum norm. We endow the Cartesian product $[0,1] \times C([0,1])^k$ with the metric ρ defined by

$$\rho((x,f),(y,h)) = |x-y| + \sup_{z\in[0,1]} \sqrt{\sum_{j=1}^{k} (f_j(z)-h_j(z))^2} \tag{C.12}$$

Lemma C.5. Define the evaluation functional, $g : [0,1]\times C([0,1])^k \to \mathbb{R}^k$ by $g(x,f) = f(x)$. Then, in the metric space $\left([0,1] \times C([0,1])^k, \rho\right)$ with ρ as defined in (C.12), g is continuous at every pair (x,f) such that f is a continuous function mapping $[0,1]$ into $\mathbb{R}^k$. It follows from this that g is a measurable function against the Borel sets of $\left([0,1] \times C([0,1])^k, \rho\right)$.

Proof. We prove this for $k = 1$, since the extension to $k \in \mathbb{N}$ follows in similar fashion with more complicated notation. Let $\varepsilon > 0$ be given. Note that

$$\begin{aligned} |g(y,h) - g(x,f)| &= |h(y) - f(x)| = |h(y) - f(y) + f(y) - f(x)| \\ &\leq |h(y) - f(y)| + |f(y) - f(x)| \\ &\leq \left\{ \sup_{z \in [0,1]} |h(z) - f(z)| \right\} + |f(y) - f(x)|. \end{aligned}$$

Since f is continuous, and $[0,1]$ is a compact set, f is uniformly continuous. Hence, there is $\delta_{f,\varepsilon} > 0$ such that if $|x - y| < \delta_{f,\varepsilon}$, then $|f(y) - f(x)| < \varepsilon$. Let $\delta_\varepsilon = \max\{\delta_{f,\varepsilon}, \varepsilon\}$. Then if $\rho((y,h),(x,f)) < \delta_\varepsilon$, it follows that $\sup_{z \in [0,1]} |h(z) - f(z)| < \varepsilon$ and $|x - y| < \delta_{f,\varepsilon}$ so that $|g(y,h) - g(x,f)| < 2\varepsilon$. □

Lemma C.6. þ Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $U : \Omega \to \mathbb{R}$ be a random variable supported in $[0,1]$. Furthermore, let $B : [0,1] \times \Omega \to \mathbb{R}^k$ be a $\mathbb{R}^k$-valued stochastic process such that $\omega \mapsto B(\cdot, \omega)$ is measurable against the Borel σ-algebra over $C([0,1])^k$. Then $H : \Omega \to \mathbb{R}$ defined by $H(\omega) = B(U(\omega), \omega)$ is a random variable (it is $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable). Additionally, if U and B are independent, then the distribution of H is a mixture defined by

$$\mathbb{P}(H \leq t) = \int_0^1 \mathbb{P}\left(B(u) \leq t\right) \mathbb{P}_U(\mathrm{d}u) \quad \forall t \in \mathbb{R}$$

where $\mathbb{P}_U$ is the pushforward measure of U against $\mathbb{P}$.

Proof. Notice that H is defined by the evaluation functional g as defined in Lemma C.5, since $H(\omega) = B(U(\omega), \omega) = g(U(\omega), B(\cdot, \omega))$. Since g is a measurable function

against the product Borel σ-algebra over $[0,1] \times C([0,1])^k$ (Lemma C.5), it follows that H is also measurable since the composition of measurable functions is itself measurable. Now, if U and B are independent, notice that

$$\mathbb{P}(H \leq t | U = u) = \mathbb{P}(B(U) \leq t | U = u) = \mathbb{P}(B(u) \leq t).$$

The final claim then follows from the Law of Total Probability. □

C.1.2. Proof of Corollary 4.1

Let $\mathcal{L}_m$ denote a random variable whose distribution is given by

$$\mathbb{P}\left(\mathcal{L}_m \leq t\right) = \int_0^1 \Phi\left(\frac{t}{s\left(p\right)}\right) \, G_m(\mathrm{d}p)$$

Using the Law of Iterated Expectations, it can be shown that $\mathcal{L}_m$ has mean zero and variance given by

$$\mathbb{E}\left[\mathcal{L}_m^2\right] = \mathbb{E}\left[s^2\left(\widetilde{p}\right)\right] = \mathbb{E}\left[\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right]$$

Next, note that for all $\widetilde{p} \in [0,1], \widetilde{p} \neq p_*$,

$$\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}} > \frac{\sigma^2(1)}{p_*} + \frac{\sigma^2(0)}{1-p_*} = \Sigma_* \ .$$

Since $\sigma^2(0) > 0$ and $\sigma^2(1) > 0$, $\mathbb{P}\left(\widetilde{p} \neq p_*\right) > 0$. That is,

$$\mathbb{P}\left(\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}} > \Sigma_*\right) > 0 \ .$$

By the strict monotonicity of expectation, we are done. □

C.1.3. Proof of Theorem 4.3

By our earlier derivation, $\widehat{\theta}_p$ has lower asymptotic variance than $\widehat{\theta}_{\widetilde{p}}$ if and only if

$$\mathbb{E}\left[\frac{\sigma^2(1)}{\widetilde{p}} + \frac{\sigma^2(0)}{1-\widetilde{p}}\right] \geq \mathbb{E}\left[\frac{\sigma^2(1)}{1/2} + \frac{\sigma^2(0)}{1-1/2}\right]$$

Define the variable Z_m:

$$Z_m = \frac{\widehat{\sigma}(1)}{\sigma(1)} \Big/ \frac{\widehat{\sigma}(0)}{\sigma(0)} \ .$$

We can then rewrite the above condition as

$$\begin{aligned}
&\mathbb{E}\left[\left(1 + \frac{1}{Z_m}\frac{\sigma(0)}{\sigma(1)}\right)\sigma^2(1) + \left(1 + Z_m\frac{\sigma(1)}{\sigma(0)}\right)\sigma^2(0)\right] \geq 2\sigma^2(1) + 2\sigma^2(0) \\
\Leftrightarrow &\mathbb{E}\left[\frac{1}{Z_m} + Z_m\right]\sigma(1)\sigma(0) \geq \sigma^2(1) + \sigma^2(0) \\
\Leftrightarrow &\frac{\sigma^2(1)}{\sigma^2(0)} - \mathbb{E}\left[\frac{1}{Z_m} + Z_m\right]\frac{\sigma(1)}{\sigma(0)} + 1 \leq 0
\end{aligned}$$

By the quadratic formula, the above inequality is satisfied whenever

$$\frac{\sigma(1)}{\sigma(0)} \in \left[B_m - \sqrt{B_m^2 - 1}, B_m + \sqrt{B_m^2 - 1}\right] = C_m \ ,$$

where $B_m := \frac{1}{2}\mathbb{E}\left[\frac{1}{Z_m} + Z_m\right]$. Note that

$$B_m^2 - (B_m^2 - 1) = 1 \Rightarrow \left(B_m - \sqrt{B_m^2 - 1}\right)\left(B_m + \sqrt{B_m^2 - 1}\right) = 1 \ ,$$

so that

$$B_m - \sqrt{B_m^2 - 1} = \frac{1}{B_m + \sqrt{B_m^2 - 1}} = \frac{1}{x} \ .$$

Next note that on R_+, $f(x) = \frac{1}{x} + x$ is strictly convex and attains its strict minimum at $x = 1$. Since Z_m is non-degenerate, Jensen's inequality yields

$$2B_m > \frac{1}{\mathbb{E}[Z_m]} + \mathbb{E}[Z_m] \geq 2$$

Hence, $|C_m| > 0$. We can then write:

$$|C_m| = 2\sqrt{B_m^2 - 1} = 2\sqrt{(B_m - 1)^2 + 2(B_m - 1)}$$

where

$$B_m - 1 = \frac{1}{2}\left[\frac{\sigma(1)}{\sigma(0)}\,\text{Bias}\,(\widetilde{p}) + \frac{\sigma(0)}{\sigma(1)}\,\text{Bias}\left(\frac{1}{\widetilde{p}}\right)\right] = W_m\ . \qquad \square$$

C.1.4. Proof of Theorem 4.4

We start by evaluating the asymptotic distributions of Z_m. First,

$$\begin{aligned}\sqrt{m}\left(\widehat{\sigma}^2(1) - \sigma^2(1)\right) &= \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left((Y_i(1) - \mu(1))^2 - \sigma^2(1)\right) - \sqrt{n}\left(\bar{Y}(1) - \mu(1)\right)^2 + o_p(1) \\ &\xrightarrow{d} \mathcal{N}\left(0\,, \mathbb{E}\left[(Y_i(1) - \mu(1))^4\right] - \sigma^4(1)\right)\ .\end{aligned}$$

By the Delta Method,

$$\sqrt{m}\left(\frac{\widehat{\sigma}^2(1)}{\sigma^2(1)} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[(Y_i(1) - \mu(1))^4\right] - \sigma^4(1)}{4\sigma^4(1)}\right)\ .$$

Similarly,

$$\sqrt{m}\left(\frac{\widehat{\sigma}^2(0)}{\sigma^2(0)} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[(Y_i(0) - \mu(0))^4\right] - \sigma^4(0)}{4\sigma^4(0)}\right)\ .$$

Since the above two displays contain independent random variables, another application of the Delta Method yields that

$$\sqrt{m}\,(Z_m - 1) \xrightarrow{d} \mathcal{N}(0, V)\ .$$

Next, let $f(x) = x + 1/x$. Note that

$$\begin{aligned} f'(x) = 1 - \frac{1}{x^2} \quad &, \quad f'(1) = 0 \\ f''(x) = \frac{2}{x^3} \quad &, \quad f''(1) = 2 \end{aligned}$$

By the second order Delta Method,

$$m\left(\frac{1}{2}\left(\frac{1}{Z_m} + Z_m\right) - 1\right) \xrightarrow{d} \frac{1}{2} V \cdot \chi_1^2$$

Since the left hand side is an analytic function of sub-Gaussian random variables, all moments can be bounded uniformly in m. Conclude that:

$$m\,(B_m - 1) \to \frac{V}{2}$$

Hence,

$$\sqrt{m}\left(\left(B_m + \sqrt{B_m^2 - 1}\right) - 1\right) = \frac{m(B_m - 1)}{\sqrt{m}} + \sqrt{m(B_m - 1)\left(\frac{m(B_m - 1)}{m} + 2\right)} \to \sqrt{V}\ .$$

□

C.2. Additional Empirical Examples

We revisit the first 10 completed RCTs in the AER RCT Registry. This section contains the empirical examples omitted from the main text. They are:

- Section C.2.1: Dillon et al. (2017)
- Section C.2.2: Finkelstein et al. (2012)
- Section C.2.3: McKenzie (2017)
- Section C.2.4: Chong et al. (2015)
- Section C.2.5: Bloom et al. (2014)
- Section C.2.6: Deming et al. (2016)
- Section C.2.7: Bryan et al. (2015)
- Section C.2.8: Galiani and McEwan (2013)

C.2.1. Dillon et al. (2017)

Dillon et al. (2017) conduct an RCT to test the hypothesis that math game play in pre-school prepares poor children for formal math in primary school. Their study, conducted in Delhi, India, with the organization Pratham, involved 214 pre-schools with 1540 children, with treatment assigned at the school level. In the Math treatment arm, children were led by facilitators to play math games over the course of four months, while the control group received lessons according to Pratham's usual curriculum. To distinguish the effect of the math games from the effect of engagement with adults, the experiment further involved a Social treatment arm, where social games were played. The outcomes of interest are Math Skills – subdivided into Symbolic Math Skills and Non-Symbolic Math Skills – as well as Social Skills, as measured by Pratham's standardized tests. Since the authors

are interested in persistence of treatment effects, they measure these outcomes at the following times after intervention: 0-3 months (Endline 1), 6-9 months (Endline 2) and 12-15 months (Endline 3). They find that the Math intervention has positive effects on Non-Symbolic Mathematical skills across all three Endlines, while Symbolic Mathematical skills only improves in Endline 1.

Table C.1 displays the standard deviation of each outcome variable by treatment arm, computed at the individual level. In the absence of correlation among students in the same pre-school, and assuming that treatment is assigned at the individual level, the numbers shown are the relevant empirical counterparts to $\sigma(1)$ and $\sigma(0)$. We see that the outcomes are relatively homoskedastic across the outcome variables and across time. The ratio $\sigma(1)\sigma(0)$ falls between 0.94 and 1.31, suggesting that naive experiment will do well when pilots are small.

Table C.1. Individual Level Heteroskedasticity in Dillon et al. (2017).

Endline	Outcome	Math	Social	Control	Math/Control	Social/Control
1	Math	0.73	0.68	0.69	1.06	0.99
	Symbolic Math	0.74	0.78	0.77	0.96	1.01
	Non-Symbolic Math	0.94	0.81	0.77	1.21	1.04
	Social	1.18	1.41	1.07	1.10	1.31
2	Math	0.71	0.73	0.69	1.04	1.06
	Symbolic Math	0.72	0.74	0.71	1.02	1.05
	Non-Symbolic Math	0.98	0.99	0.92	1.06	1.07
	Social	0.92	0.96	0.99	0.94	0.97
3	Math	0.78	0.70	0.75	1.04	0.93
	Symbolic Math	0.72	0.68	0.74	0.98	0.92
	Non-Symbolic Math	1.17	1.09	1.06	1.11	1.03
	Social	1.01	1.05	1.07	0.94	0.98

Suppose we are concerned about correlation across students in the same pre-school. We can redefine our unit of observation to be the school by taking averages across students in the same school. The Neyman Allocation then tells us how many schools to allocate to treatment. In this case, the standard deviation in the mean across schools is the relevant counterpart to $\sigma(1)$ and $\sigma(0)$. They are presented in Table C.2. As before, we see that outcomes are relatively homoskedastic across schools, though less so than in the individual level case. Nonetheless, the ratio falls between 0.84 and 1.55. Once we consider schools to be the unit of treatment, however, the effective pilot size also shrinks, such that the drawbacks of the estimated Neyman Allocation may be even more pronounced. All in all, the Dillon et al. (2017) example supports our case of relative homoskedasticity in empirical applications.

Table C.2. School Level Heteroskedasticity in Dillon et al. (2017).

Endline	Outcome	Math	Social	Control	Math/Control	Social/Control
1	Math	0.41	0.38	0.30	1.34	1.24
	Symbolic Math	0.39	0.42	0.34	1.14	1.22
	Non-Symbolic Math	0.51	0.40	0.33	1.55	1.22
	All Social	0.51	0.71	0.46	1.10	1.55
2	All Math	0.39	0.35	0.36	1.10	0.97
	Symbolic Math	0.40	0.35	0.36	1.11	0.96
	Non-Symbolic Math	0.46	0.42	0.43	1.06	0.98
	Social	0.41	0.48	0.49	0.84	0.97
3	All Math	0.44	0.40	0.39	1.15	1.04
	Symbolic Math	0.40	0.35	0.37	1.06	0.94
	Non-Symbolic Math	0.65	0.63	0.53	1.23	1.20
	Social	0.55	0.59	0.50	1.10	1.20

C.2.2. Finkelstein et al. (2012)

Finkelstein et al. (2012) study the Oregon Health Insurance Experiment, in which uninsured, low-income adults were randomly given the opportunity to apply for Medicaid. Over the course of a month in February 2008, Oregon conducted extensive public awareness campaign to encourage participation in the lottery. From a total of 89,824 sign-ups, 35,169 individuals (from 29,664 households) were selected. They, and any members of their households were then given the opportunity to apply for Medicaid. Hence, treatment occurred at the household level.

The authors used the data to study a variety of outcomes. In this section, we focus on their first set of results, which concern healthcare utilization. In particular, we revisit the outcome variables used in Tables V and VI of Finkelstein et al. (2012), which are obtained from survey data (as opposed to administrative data), and are hence publicly available. Inline with the authors' results on the Intent-to-Treat effect, we define the treated group as those selected by the lottery. We note that the authors apply sampling weights to correct for differential response rates to the survey. We follow their weighting scheme in computing our results.

The standard deviation of individual-level outcome are presented in Table C.3. Results taking household to be the unit of observation are presented in Table C.4. Across both tables, we see that that the standard deviations in outcomes are remarkably similar across treatment and control groups. They are also very similar across individual and household level groups, since households with more than one person represented less than 5% of the survey sample.

Table C.3. Individual Level Heteroskedasticity in Finkelstein et al. (2012).

	Outcome	Treatment	Control	Treat./Cont.
Extensive Margin	Prescription drugs currently	0.48	0.48	0.99
	Outpatient visits last six months	0.48	0.49	0.98
	ER visits last six months	0.44	0.44	1.00
	Inpatient hospital admissions last six months	0.26	0.26	1.00
Total Utilization	Prescription drugs currently	2.90	2.88	1.01
	Outpatient visits last six months	3.29	3.09	1.07
	ER visits last six months	1.01	1.04	0.97
	Inpatient hospital admissions last six months	0.42	0.40	1.04
Preventative Care	Blood cholesterol checked (ever)	0.48	0.48	0.98
	Blood tested for high blood sugar (ever)	0.48	0.49	0.99
	Mammogram within last 12 months (women $\geq$ 40)	0.48	0.46	1.04
	Pap test within last 12 months (women)	0.50	0.49	1.01

Table C.4. Household Level Heteroskedasticity in Finkelstein et al. (2012).

	Outcome	Treatment	Control	Treat./Cont.
Extensive Margin	Prescription drugs currently	0.46	0.47	0.99
	Outpatient visits last six months	0.47	0.48	0.97
	ER visits last six months	0.43	0.44	0.99
	Inpatient hospital admissions last six months	0.26	0.26	0.99
Total Utilization	Prescription drugs currently	2.88	2.86	1.01
	Outpatient visits last six months	3.30	3.09	1.07
	ER visits last six months	1.01	1.04	0.98
	Inpatient hospital admissions last six months	0.41	0.40	1.02
Preventative Care	Blood cholesterol checked (ever)	0.46	0.48	0.97
	Blood tested for high blood sugar (ever)	0.47	0.48	0.98
	Mammogram within last 12 months (women $\geq$ 40)	0.48	0.46	1.04
	Pap test within last 12 months (women)	0.50	0.49	1.01

C.2.3. McKenzie (2017)

Business plan competitions are growing in popularity as a way of fostering high growth entrepreneurship in developing countries. McKenzie (2017) studies the Youth Enterprise With Innovation in Nigeria (YouWiN!) program, which distributed up to US$64,000 to winners. A portion of the awards were reserved for business plans that were clearly

superior to the rest. 1,841 entrepreneurs, determined to be of medium quality, were entered into a lottery, from which 729 were selected for the award. Three rounds of surveys were then conducted at 1,2 and 3 years after the application respectively.

We focus on the first set of results in McKenzie (2017) – presented in Table 2 – concerning the effect of the award on start-up and survival. Here, they find that the grant persistently increased the probability that the entrepreneur was operating a firm, as well as the number of hours they spent in self-employment. We present the standard deviations of these outcome variables in Table C.5. Here we see that hours in self employment is roughly homoskedastic across all three periods. However, the outcome on whether the entrepreneur is operating a firm is arguably highly heteroskedastic, with standard deviations that is as small as 0.44 that of the control group. As in our earlier example, we find high kurtosis in these outcome variables, displayed in Table C.6.

We do not estimate C_m in this example because almost all entrepreneurs in the treated group operate their own firms in the sample. As such, a small random sub-sample (e.g. of size below 200) from this group has variance 0 with high probability, impeding the estimation of C_m. These pathological cases are revealing. In a pilot, if the treated group has variance 0 in the outcome, the estimated Neyman Allocation assigns 0 units to treatment in the full experiment. The high probability of such an “extreme" outcome with small pilots is precisely the danger which we are warning against. We conclude that the estimated Neyman from a small pilot will likely lead to adverse results given the DGP in McKenzie (2017).

Table C.5. Heteroskedasticity in McKenzie (2017).

Outcome	New Firms			Existing Firms		
	Treat.	Cont.	Ratio	Treat.	Cont.	Ratio
Operates a Firm at Round 1	0.43	0.50	0.86	0.21	0.34	0.63
Operates a Firm at Round 2	0.27	0.50	0.55	0.16	0.36	0.44
Operates a Firm at Round 3	0.28	0.50	0.56	0.20	0.43	0.48
Weekly Hours of Self Emp. at Round 1	29.40	29.75	0.99	25.74	27.81	0.93
Weekly Hours of Self Emp. at Round 2	24.98	28.62	0.87	24.51	29.71	0.82
Weekly Hours of Self Emp. at Round 3	24.77	25.85	0.96	25.09	26.10	0.96

Table C.6. Kurtosis in McKenzie (2017).

Outcome	New Firms		Existing Firms	
	Treat.	Cont.	Treat.	Cont.
Operates a Firm at Round 1	2.52	1.04	19.68	5.91
Operates a Firm at Round 2	10.69	1.08	35.46	4.58
Operates a Firm at Round 3	9.62	1.03	21.05	2.47
Weekly Hours of Self Emp. at Round 1	1.99	3.08	3.93	3.08
Weekly Hours of Self Emp. at Round 2	2.77	2.92	3.03	2.42
Weekly Hours of Self Emp. at Round 3	2.09	3.28	3.47	2.15

C.2.4. Chong et al. (2015)

Partnering with the Peruvian nongovernmental organization PRISMA, Chong et al. (2015) conducted two RCTs to investigate the efficacy of various interventions in encouraging recycling. First, the Participation Study considers the following 9 different messaging strategies and their relative success in enrolling members into recycling programs:

(1) Norms: Rich and Poor. Norm messaging focus on communicating high recycling rate of either a rich or poor reference neighborhoods, encouraging conformity.

(2) Signal: Rich, Poor and Local. Signal messaging informs the targets that their recycling behavior will be known to either a nearby neighborhood (Local), a distal neighborhood of varying wealth (Rich or Poor), affecting the targets reputation.

(3) Authority: Religious or Municipal. Authority messaging communicates that a higher authority, either religious or local governmental, advocates recycling.

(4) Information: Environmental or Social. Informational messaging communicated the benefits of recycling, either to the environment or to the local society (e.g. by creating jobs).

Out of a total of 6,718 households, approximately 600 were assigned to each treatment arm, with the exception of Signal: Local, which were assigned 932 participants. 1,157 households were assigned to the control group. Three measures of participation were considered:

(1) “Participates any time" is an indicator that takes the value 1 if a household turned in residuals over the course of the study.

(2) “Participation Ratio" is the number of times a household turns in residual over the total number of opportunities they had to turn in residuals.

(3) “Participates during either of last two visits" is an indicator that takes value 1 if the household turned in residual during one of the last two canvassing weeks.

The results in Table 3 of Chong et al. (2015) shows that messaging had no effect in increasing participation in the program. Table C.7 displays the standard deviation of the various outcomes by treatment type. Table C.8, shows the ratio of the standard deviation in the outcome variable of each treatment group, with respect to that of the control group. Here, we see that the outcomes are highly homoskedastic, suggesting little scope for improvement over the naive experiment.

The second experiment is the Participation Intensity Study. The outcomes of interest are the following measures recycling intensity:

Table C.7. S.D. of outcome by treatment type in the Participation Study. See text for definitions of outcome and treatment.

Outcome	Control	Norms		Signal			Authority		Info.	
		Rich	Poor	Rich	Poor	Local	Reli.	Muni.	Env.	Social
1	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
2	0.389	0.389	0.392	0.403	0.385	0.393	0.397	0.388	0.396	0.392
3	0.490	0.486	0.494	0.493	0.486	0.489	0.491	0.491	0.494	0.493

Table C.8. Ratio of the S.D w.r.t. the control group in the Participation Study.

Outcome	Norms		Signal			Authority		Info.	
	Rich	Poor	Rich	Poor	Local	Reli.	Muni.	Env.	Social
1	1.000	0.999	1.000	1.000	1.000	1.000	0.999	1.000	1.000
2	1.000	1.008	1.036	0.989	1.012	1.022	0.998	1.018	1.008
3	0.992	1.007	1.005	0.990	0.997	1.000	1.001	1.007	1.005

(1) Percentage of visits in which household turned in residuals

(2) Average number of bins turned in per week

(3) Average weight (in kg) of recyclables turned in per week

(4) Average market value of recyclables given per week

(5) Average percentage of contamination (non-recyclables mixed into recycling) per week.

The treatments of interest are (1) providing recycling bins to households and (2) sending SMS reminders for recycling.[1] Of the 1,781 households in this study, 182 were received Bin and SMS. 417 received the Bin only treatment. 369 received the SMS only treatment, leaving 817 in the control group. The authors find, in Table 4A that bin provision was highly effective in increasing recycling, though SMS reminders had no effect. We compute

[1]The authors also consider providing bins with and without instructions as well as generic vs personalized SMSes. However, these finer definitions leads to treatment arms with fewer 50 households. Hence we focus on the coarser definition of the treatments, as employed in panel 4A.

the standard deviation in each of their outcome variables by treatment type in Table C.9. Also displayed is the ratio of these standard deviation with respect to the control group. Again, we see strong evidence of homoskedasticity, suggesting that the naive experiment will perform well in this scenario.

Table C.9. S.D. of outcome by treatment type in the Participation Intensity Study. See text for definitions of outcome and treatment.

Outcome	Standard Deviation				Ratio of S.D. w.r.t. Control		
	Control	SMS only	Bin only	SMS & Bin	SMS only	Bin only	SMS & Bin
1	0.262	0.286	0.233	0.227	1.092	0.891	0.867
2	0.404	0.371	0.441	0.375	0.919	1.092	0.927
3	0.744	0.646	0.756	0.727	0.869	1.017	0.978
4	0.418	0.371	0.416	0.399	0.889	0.996	0.955
5	0.156	0.145	0.136	0.128	0.928	0.870	0.822

C.2.5. Bloom et al. (2014)

Bloom et al. (2014) study the effect of working from home on employees' productivity via a randomized experiment at Ctrip, a NASDAQ-listed Chinese travel agency with 16000 employees. The main concern is whether or not working from leads to shirking. Ctrip decided to run a nine-month experiment on working from home. They asked the 996 employees in the airfare and hotel departments of the Shanghai call center if they were interested in working from home four days a week and one day in the office. 503 of these employees were interested and of these, 249 were qualified to take part on the basis of tenure, broadband access and access to private work space at home. Qualified employees were assigned to working from home if they had even-numbered birthdays so that those with odd-numbered birthdays formed the control group. The treatment and control groups were comprised of 131 and 118 employees respectively. The only difference

between the two groups was location of work – both groups used the same equipment, faced the same workload and were compensated under the same pay system. The authors find a 13% increase in productivity of which the main source of improvement was a 9% increase in the number of minutes worked during a shift. The remaining 4% came from an increase in the number of calls per minute worked.

Table C.10 reports standard deviations for treatment and control groups in the experiment as well as the standard deviation ratios for the main outcomes of interest in Bloom et al. (2014). Strong evidence of relative homoskedasticity with respect to treatment status presents in this study as well – suggesting that the balanced allocation would outperform the FNA.

Table C.10. Performance impact outcomes: standard deviations.

Outcome Variables	Control S.D.	Treated S.D.	Ratio
Overall Performance	1.0049	1.0035	0.9986
Phone calls	0.9775	0.7502	0.7675
Log phone calls	0.2476	0.1764	0.7123
Log call per sec	0.0217	0.0299	1.3786
Log call length	0.2701	0.2729	1.0105

C.2.6. Deming et al. (2016)

Deming et al. (2016) study employers' perceptions of the value of post-secondary degrees using a field experiment. The experimental units are fictitious resumes to be used in applications to vacancies posted on a large online job board. Their focus is on degrees and certificates awarded in the two largest occupational categories in the United States: business and health. Resumes are randomly assigned sector and selectivity of (degree-awarding) institutions. Fictitious resumes are created using a vast online database of

actual resumes of job seekers, with applicant characteristics varying randomly (i.e. characteristics are randomly assigned). Outcomes are callback rates. There are three main comparisons in the paper:

- for-profit vs. public institutions,
- for-profits that are online vs. brick-and-mortar (with a local presence),
- more selective vs. less selective public institutions.

Deming et al. (2016) find that BA degrees in business from large online for-profit institutions are more 22% less likely to receive a callback than applicants with similar degrees from non-selective public schools when the job vacancy requires a BA. When a business job opening does not list a BA requirement, they find no significant overall advantage to having a post-secondary degree. For health jobs, resumes with certificates from for-profit institutions are 57% less likely to receive a callback than those with similar certificates from public institutions when the job listing does not require a post-secondary certificate. No significant difference in callback rates are found when the health job listing requires a certificate.

Table C.11 reports standard deviations across treatment arms across the various subpopulations of interest in Deming et al. (2016). Since in most sub-populations, there are more than two treatment arms, we do not report ratios. However, pairwise comparisons between treatment arms within any chosen subpopulation shows strong evidence of relative homoskedasticity across the board.

Table C.11. Performance impact outcomes: standard deviations.

Experimental Population	Treatment Arm	S.D.
Business jobs without BA requirements	No degree	0.3054
	AA (for profit)	0.3026
	AA (public)	0.3053
	BA (for profit)	0.3071
Business jobs with BA requirements	BA (for profit, online)	0.2522
	BA (for profit, not online)	0.2209
	BA (public, selective)	0.2879
	BA (public, not selective)	0.2595
Health job without cert. requirement	No certificate	0.2900
	Certificate (for profit)	0.2922
	Certificate (public)	0.3014
Health job with cert. requirement	Certificate (for profit)	0.2400
	Certificate (public)	0.2681

C.2.7. Bryan et al. (2015)

Bryan et al. (2015) conduct a field experiment to study efficacy of peer intermediation in mitigating adverse selection and moral hazard in credit markets. To identify the effects of peer screening and enforcement, they use a two-stage referral incentive field experiment. The experiment was conducted through Opportunity Finance South Africa (Opportunity), a for-profit lender in the consumer micro-loan market. Over the period of February 2008 through July 2009, Opportunity offered individuals approved for a loan the option to participate in its "Refer-A-Friend" program. Referred individuals earned R40 if they brought in a referral card and were approved for a loan. The referrer could earn R100 for referring someone who was subsequently approved for and/or repaid the loan, depending on the referrer's incentive contract. Referrers were randomly assigned to one of two ex-ante incentive contracts:

- Approval incentives: the referrer would be paid only if the referred was approved for a loan.
- Repayment incentive: the referrer would be paid only if the referred successfully repaid the loan.

Among referrers whose referred friends were approved for a loan, Opportunity randomly selected half to be surprised with an ex-post incentive change:

- Half among the ex-ante approval group were phoned and told that in addition to the R100 approval bonus, they would receive an additional R100 if the loan was successfully repaid by the referrer.
- Half among the ex-ante repayment group were phoned and told that they would receive the R100 now, and that receipt of the bonus would no longer be conditional on repayment of the loan by the referrer.

The overall incentive structure is as follows

- Ex-ante and ex-post approval (EA = A): no enforcement or screening incentive.
- Ex-ante repayment and ex-post approval (EA = R): screening incentive.
- Ex-ante approval and ex-post repayment (EA = A, EP = R): Enforcement incentive.
- Ex-ante repayment and ex-post repayment (EA = R, EP = R): Enforcement and screening incentive.

The authors find no evidence of screening but do find large enforcement effects.

Table C.12 reports standard deviations in the main outcomes of interest in Bryan et al. (2015) across the four treatment arms. For a given outcome, pairwise comparisons across

arms yields evidence of relative homoskedasticity with respect to treatment status. We conclude that in this case, the balanced allocation would outperform the FNA.

Table C.12. S.D. across treatment groups G1, G2, G3 and G4.

Outcome	EA = A	EA = A, EP = R	EA = R	EA = R, EP = R
Penalty interest	0.4919	0.4407	0.4488	0.5043
Positive balance owing at maturity	0.4086	0.2959	0.3613	0.4225
Proportion of value owing at maturity	0.4088	0.3106	0.3153	0.5526
Loan charged off	0.3652	0.2147	0.2917	0.3950

C.2.8. Galiani and McEwan (2013)

Galiani and McEwan (2013) use the Honduran PRAF experiment to study the impact of conditional cash transfers (CCT) on the likelihood of children to work versus enrolling in school. The PRAF experiment randomly allocated CCT's among 70 municipalities. These 70 were chosen out of a total of 298 on the basis of mean heights-for-age z-scores of first graders. The 70 municipalities were further assigned to four treatment arms termed G1, G2, G3, G4. G1 received CCT's in education and health. G2 received CCT's in addition to direct investment in education and health centers. G3 received only direct investments and finally, G4 served as the control group and received no interventions. The 70 municipalities were further divided into 5 strata each consisting of 14 municipalities on the basis of quintiles of mean height-for-age. Random assignment was performed within these strata (stratified randomization). The final sample consisted of 20 municipalities in G1, 20 in G2, 10 in G3, and 20 in G4. The authors match the experimental data with census data and use the latter to construct the outcomes of interest which are three dummy variables. The first is an indicator for whether a child is enrolled in and attending

school during the time of the census. The second indicates whether the child worked during the week prior to the census or conditional on a negative response to the former, whether they reported non-wage employment in a family farm or business. The third indicates whether the child worked exclusively on household chores. The authors find that overall, children eligible for CCTs were 8% more likely to enroll in school and 3% less likely to work.

Table C.13 reports standard deviations for the outcomes of interest across treatment arms. Again, for a given outcome, pairwise comparisons across arms yields evidence of relative homoskedasticity with respect to treatment status. We conclude that in this case, the balanced allocation would outperform the FNA.

Table C.13. S.D. across treatment groups G1, G2, G3 and G4.

Outcome	G1	G2	G3	G4
Enrolled in school	0.4393	0.4474	0.4812	0.4769
Works outside home	0.2637	0.2267	0.2893	0.2986
Works only in home	0.3015	0.2841	0.3478	0.3409