Dependence-Robust Inference Using Randomized Subsampling

Michael P. Leung*

December 14, 2017

Abstract. We study hypothesis tests and confidence intervals constructed using a randomized subsampling procedure first proposed by Song (2016). We show that these procedures are robust to general forms of weak dependence in the sense that they are asymptotically valid under the weak requirement that the target parameter can be consistently estimated at the $\sqrt{n}$ rate by an asymptotically linear estimator. The implementation of randomized subsampling does not depend on the correlation structure of the data and is computationally simple. We consider applications to constructing clustered standard errors when the level of clustering is unknown and the number of clusters is potentially small, inference on network data when the network is partially observed or the network-formation model is unknown, and testing for power laws with dependent data. We also develop randomized subsampling tests of moment inequalities.

JEL Codes: C12, C31
Keywords: subsampling, dependent data, social networks, clustered standard errors, power laws

*Department of Economics, University of Southern California. E-mail: leungm@usc.edu. I thank Vittorio Bassi, Jinyong Hahn, Roger Moon, and Hashem Pesaran for useful comments. I also thank seminar audiences at UC Davis and USC INET.

1 Introduction

Suppose the econometrician believes her data to be correlated but lacks information regarding its dependence structure. For example, suppose she has data on individuals sampled from a geographic region with only a small number of zip codes, but all she observes about locations are individuals' zip codes. If she believes her data to be spatially dependent, it is difficult to account for this, since locations are imperfectly observed. For instance, it is impossible to compute a spatial HAC variance estimator, since distances between pairs of observations are unknown. Another example is clustered data with imperfectly observed cluster memberships, for instance if there are multiple potential levels at which one can cluster, but the right level of clustering is unknown. Alternatively, it is possible that the econometrician has the option of computing a variety of variance estimators that control for different types of dependence, but it is unclear which form of dependence is of first-order importance. Should she cluster geographically? At which level? Should she use a spatial HAC estimator to control for a more heterogeneous form of spatial dependence? Or is the dependence structure more complicated, for example network dependence?

This paper studies inference procedures robust to general forms of weak dependence. These procedures are constructed using "randomized subsampling," a resampling procedure first proposed in an insightful paper by Song (2016), the implementation of which is independent of the correlation structure of the data. We prove that randomized subsampling procedures are asymptotically valid under weak conditions on the correlation structure that are satisfied by most forms of weakly dependent data. In this sense, randomized subsampling is robust to general forms of weak dependence.

To be more concrete, consider inference on the population mean. The randomized subsampling procedure consists computing a resampled test statistic that averages over random subsamples of the original data. We show that this statistic is asymptotically normal under the weak requirement that the sample mean is $\sqrt{n}$-consistent. One can then easily construct confidence intervals using normal quantiles. Song (2016) proves that the randomized subsampling statistic has a normal limit under a weak-dependence condition he refers to as local dependence. He verifies this condition holds under M-dependence or strong mixing, under certain restrictions on the mixing coefficients. It is less clear how to verify this condition for other forms of dependence such as near-epoch dependence or forms of network dependence that occur in strategic models of network formation. In contrast, $\sqrt{n}$-consistency condition is a very minimal requirement that follows directly from most conventional forms of weak dependence, which makes randomized subsampling applicable to a wider variety of contexts. Also, Song (2016) focuses on equality tests, while we also develop a method for testing moment inequalities.

We consider four applications. The first is inference under cluster dependence when the level of clustering is unknown and the number of clusters is potentially small, which are well-known problems with implementing existing clustered standard

errors (see e.g. Cameron and Miller, 2015, §IV-VI). The second is inference on network statistics, such as the average clustering coefficient or degree distribution. The motivation is that there are many possible models of network formation, and the same network statistics under different models may have different asymptotic variances. However, for many such models, these statistics are $\sqrt{n}$-consistent, so randomized subsampling can be used for inference robust to the underlying network formation model. The third application is treatment effects with network spillovers when the network is imperfectly observed. The fourth application is testing for a power law distribution, a problem that has received a great deal of attention in economics, network science, biology, and physics (Barabási and Albert, 1999; Gabaix, 2009; Newman, 2005). Widely used methods in practice assume that the underlying data is i.i.d. (Clauset et al., 2009; Klaus et al., 2011), which is generally implausible in economic applications involving spatial or financial data and network applications.

Many resampling methods are available for inference on spatial, temporal, and clustered data when the dependence structure is known (see e.g. Cameron et al., 2008; Lahiri, 2013; Politis et al., 1999). Knowledge of the dependence structure is commonly exploited by resampling blocks of neighboring observations, but this requires data on which observations are neighbors, which is unavailable if, for example, locations are imperfectly observed. It is also an open question how to devise valid resampling procedures for dependent data that lack a temporal or spatial structure, such as network data.

Conventional resampling procedures are used to construct critical values for a test statistic computed on the original dataset. For the critical values to be asymptotically valid, resampling has to be implemented in a way that mimics the dependence structure of the data, which requires information about the dependence structure. In contrast, with randomized subsampling, one computes a resampled test statistic and critical values for this new statistic based on its limiting distribution. Thus, the objective is not to mimic the actual dependence structure, which is intuitively why randomized subsampling is dependence-robust.

Asymptotic normality of randomized subsampling statistics follows from the fact that we draw random subsamples in an i.i.d. fashion, so that conditional on the data, the statistic has a normal limit when centered at the conditional mean. Under appropriate conditions on the number of subsamples and subsample size, the difference between the conditional mean and target unconditional mean is negligible in large samples, which establishes asymptotical validity of the procedure.

Of course, the broad applicability of the randomized subsampling procedure comes at some cost. First, implementation requires choosing two tuning parameters, the subsample size and the number of random subsamples to average over. However, note that conventional subsampling for dependent data also requires two tuning parameters: the subsample size and the block size. Most inference methods for dependent data depend on some tuning parameters, including the bootstrap and autocovariance-consistent variance estimators, and it is generally a difficult problem to devise data-

dependent choices of these parameters. In this paper, we provide simple rules of thumb for tuning parameter choice that work well across a variety of data-generating processes in our simulation study.

A second problem with randomized subsampling is that, in settings where existing inference procedures exist, randomized subsampling will typically yield wider confidence intervals and tests with lower power, a problem shared with conventional subsampling. This is the price to be paid for having a widely applicable procedure. Our objective is not to propose a procedure that is competitive with existing procedures but rather to provide a broadly applicable and robust inference procedure that is advantageous when little is known about the dependence structure and useful for complex forms of dependency for which no procedure is presently available.

The outline of the paper is as follows. The next section introduces the randomized subsampling statistic and shows how to construct CIs for a population mean. We then discuss the four applications and related literature. We prove the asymptotic validity of the inference procedures for the population mean problem in §3 and provide a test for moment inequality models in §4. We discuss results from an empirical application to testing for power law degree distributions in §5. Then §6 presents simulation results for four different data-generating processes, which motivate our recommendations for tuning parameter choice. Finally, §7 concludes.

2 Overview and Applications

Let $\mathbf{X} = \{X_i\}_{i=1}^n \subset \mathbb{R}^m$, a set of identically distributed random vectors with possibly dependent row elements. Denote the mean of $\mathbf{X}$ by $\bar{X}$. Our main assumption requires $\mathbf{X}$ to be weakly dependent in the sense that $\bar{X}$ is $\sqrt{n}$-consistent for a parameter $\mu_0 \in \mathbb{R}^m$. For now, consider the case $\mu_0 = \mathbf{E}[X_1]$; we will discuss other examples below. In this section, we consider the problems of testing that μ_0 equals a hypothesized value and constructing a confidence region for μ_0. We consider testing moment inequalities in §4.

Let b_n be a natural number less than n, denoting the subsample size, and R_n a natural number denoting the number of subsamples. Let Π be the set of all bijections (permutation functions) on $\{1, \ldots, n\}$. Draw $\{\pi_r\}_{r=1}^{R_n}$ i.i.d. and uniformly from Π, and let $\pi = (\pi_1, \ldots, \pi_{R_n})$. Define the sample variance matrix $\hat{\Sigma} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'/n$. Note that $\hat{\Sigma}$ is generally not consistent for $n\text{Var}(\bar{X})$ because we allow for dependent data.

We define two statistics, following Song (2016). The first is the randomized subsampling mean-type statistic, given by

$$T_n(\mu_0; \pi) = \frac{1}{\sqrt{R_n b_n}} \sum_{r=1}^{R_n} \sum_{i=1}^{b_n} \hat{\Sigma}^{-1/2} \left(X_{\pi_r(i)} - \mu_0 \right).$$

This is computed by drawing R_n subsamples of size b_n and averaging. The second is

a variant of Song's randomized subsampling U-type statistic,[1] given by

$$S_n(\mu_0;\pi) = \frac{1}{\sqrt{2R_n}b_n}\sum_{r=1}^{R_n}\sum_{i=1}^{b_n}\sum_{j=1,j\neq i}^{b_n}(X_{\pi_r(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi_r(j)}-\mu_0).$$

Inference. In the next section, we show that if $\bar{X}$ is $\sqrt{n}$-consistent for μ_0, then (under regularity conditions)

$$\begin{aligned} T_n(\mu_0;\pi) &\xrightarrow{d} \mathcal{N}(0,I_m) \quad \text{if} \quad \frac{R_n b_n}{n} \to 0, \\ S_n(\mu_0;\pi) &\xrightarrow{d} \mathcal{N}(0,1) \quad \text{if} \quad \frac{\sqrt{R_n}b_n}{n} \to 0, \end{aligned} \tag{1}$$

where I_m is the $m \times m$ identity matrix. Then to test the null that $\mu_0 = \mu$ against two-sided alternatives, we can use

$$\mathbf{1}\{T_n(\mu;\pi)'T_n(\mu;\pi) > c\} \quad \text{or} \quad \mathbf{1}\{S_n(\mu;\pi) > z\}, \tag{2}$$

where c is the appropriate chi-square quantile (with m degrees of freedom) and z the appropriate normal quantile. To construct a CI for a component of μ_0 for the case $m = 1$, (1) suggests the following simple CI:

$$\frac{1}{R_n b_n}\sum_{r=1}^{R_n}\sum_{i=1}^{b_n} X_{\pi_r(i)} \pm z\frac{\hat{\Sigma}^{1/2}}{\sqrt{R_n b_n}},$$

where z is the appropriate normal quantile. Alternatively, we can use the U-type statistic to obtain a CI by test inversion:

$$\{\mu_0 \in \mathbb{R}^m : S_n(\mu_0;\pi) < z\}.$$

Tuning Parameters. Based on our simulation study in §6, in practice we suggest choosing

$$\begin{aligned} (R_n, b_n) &= (n^{1/4}, n^{1/4}) \quad \text{for} \quad T_n(\mu_0;\pi), \\ (R_n, b_n) &= (n^{2/3}, n^{1/3}) \quad \text{for} \quad S_n(\mu_0;\pi), \end{aligned}$$

which work well for both tests across a variety of testing problems and data-generating processes.

In general, we face the following trade-off in tuning parameter choice: larger values of R_n and b_n mean higher power and narrower confidence intervals, but this also inflates a bias term, leading to a poorer limiting approximation. Consider the case $\mu_0 = 0$. For the mean-type statistic, the test has power against local alternatives

[1] Our U-type statistic is scaled differently, and we do not partially bias-correct, since this appears to worsen coverage in our simulations.

that vanish no faster than $(R_n b_n)^{-1/2}$ (call this rate the test's rate of convergence for short), but the bias term vanishes at the rate $(R_n b_n/n)^{1/2}$. For the U-type statistic, the test's rate of convergence is $R_n^{-1/4} b_n^{-1/2}$, but the bias term is order $R_n^{1/2} b_n/n$.[2] Thus for the first choice of tuning parameters above, for the mean-type statistic, the bias and rate of convergence are both $n^{-1/4}$, whereas for the U-type statistic, they are respectively $n^{-5/8}$ and $n^{3/16}$. In contrast, for the second choice of tuning parameters above, the bias and rate of convergence are respectively 1 and $n^{-1/2}$ for the mean-type statistic, and for the U-type statistic, both equal $n^{-1/3}$.

Choice of Statistic. The mean-type statistic leads to confidence intervals that are easy to compute because they have the simple form of being a mean of randomized subsamples plus or minus a "standard error," which does not require test inversion. Also, the U-type statistic can only be used to test against two-sided alternatives, whereas the mean-type statistic can be adopted for one-sided alternatives. However, the U-type statistic has a better rate of convergence under our recommended tuning parameters and should therefore be preferred.

Asymptotically Linear Estimators. Our setup generalizes beyond the sample-mean case. Suppose we observe data $\mathbf{Z} = \{Z_i\}_{i=1}^n$, and we are interested in a parameter $\beta_0 \in \mathbb{R}^d$. Let $\hat{\theta}$ be an asymptotically linear estimator in the sense that

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Z_i; \beta_0, \hat{\gamma}) + o_p(1) \tag{3}$$

for some function ψ and first-stage estimator $\hat{\gamma}$ that is a function of $\mathbf{Z}$. For example, in the case of maximum likelihood, $\hat{\gamma}$ is the sample Hessian, and ψ is the score function times the Hessian. We can then apply randomized subsampling to the asymptotically linear form of $\hat{\beta}$ to conduct inference on β_0 by defining

$$X_i = \psi(Z_i; \beta_0, \hat{\gamma}) \tag{4}$$

As discussed in Remark 2, under regularity conditions, the resulting procedure is asymptotically valid if $\hat{\beta}$ is $\sqrt{n}$-consistent for β_0 and $\hat{\gamma}$ is $\sqrt{n}$-consistent for its population analog γ.

In the remainder of this section, we discuss four applications of randomized subsampling.

2.1 Cluster Dependence

Let Y be an n-dimensional outcome vector, W an $n \times k$ matrix of covariates, and W_i the ith row of X. Consider the standard linear model

$$Y_i = W_i' \beta_0 + \varepsilon_i,$$

[2] For the bias terms, see the proof of Theorem 1 in the appendix. For the rates of convergence, see Remark 3 in §3.

where $\{\varepsilon_i\}_{i=1}^n$ is identically distributed with mean zero but possibly dependent. We are interested in the jth component of β, denoted by β_j.

It is a common concern that $\{(W_i, \varepsilon_i)\}_{i=1}^n$ are dependent, perhaps because the data is clustered (Bertrand et al., 2004), there is spatial dependence (Barrios et al., 2012; Bester et al., 2011), or ε_i is some function of an underlying social network and therefore exhibits network autocorrelation (Acemoglu et al., 2015). We may often not know the precise form of dependence or not have enough data to use conventional standard error formulas, for example if we do not fully observe the clusters, the spatial locations of the observations, or the network. Nonetheless, we can still construct a valid CI for β_j using randomized subsampling.

To apply our procedure, we write the estimator in the form of a sample mean. Let $A = (W'W/n)^{-1}W'$ and A_{ji} its jith component. Then the least-squares estimator for β_j can be written as

$$\frac{1}{n}\sum_{i=1}^{n} A_{ji}Y_i.$$

This fits into our setup, specifically (4), if we define $Z_i = (Y_i, X_i)$, $\hat{\gamma} = W'W/n$, and $\psi(Z_i; \beta_0, \hat{\gamma}) = A_{ji}Y_i$. That is, we apply randomized subsampling to the summands $X_i = A_{ji}Y_i$.

The main assumptions required for these CIs to have correct asymptotic coverage are $\sqrt{n}$-consistency of the least-squares estimator and $W'W/n$. For cluster dependence, this holds under conventional many-cluster asymptotics, where the number of observations in each clusters is small, but the number of clusters is large. Importantly, we need not know the right level of clustering or even observe cluster memberships. For spatial dependence, this holds under standard mixing or near-epoch dependence conditions (Jenish and Prucha, 2009, 2012). For forms of dependence captured by dependency graphs, this holds under restrictions on the degree distribution (Aronow and Samii, forthcoming; Leung, 2017a).

We can also allow the number of clusters to be small, perhaps even equal to one, so long as the data is weakly dependent within clusters in the sense of $\sqrt{n}$-consistency of the estimator. Bakirov and Székely (2006), Canay et al. (2017), Ibragimov and Müller (2010), and Ibragimov and Müller (2016) propose novel inference procedures for cluster dependence when the number of clusters is small, also assuming weak dependence within clusters. An advantage of randomized subsampling is that we can allow for only a single cluster and do not require knowledge of cluster memberships.

2.2 Network Statistics

Much of the literature in network science is motivated by a handful of stylized facts about real world social networks (Barabási, 2015; Jackson, 2008). These facts are obtained by computing various network statistics from networks across a wide variety of social and economic domains. However, we have no measure of sampling variation for these point estimates. In part, this is due to the wide variety of network

formation models, many of which induce different dependence structures among the set of potential links. This motivates the use of randomized subsampling, which can be used to construct CIs for network statistics without taking a stance on the particular data-generating process. Network statistics are also important for inference on strategic models of network formation (Sheng, 2016; de Paula et al., forthcoming; Leung, 2017b). Leung and Moon (2017) develop a central limit theorem applicable to such models, but the asymptotic variance has a complicated form for which it is difficult to construct a consistent estimator. For this reason, the authors resort to randomized subsampling for inference.

We consider two stylized facts that have arguably received the most attention in the literature: clustering and power law degree distributions.[3] This subsection focuses on the former, while the latter is discussed in the more general context of testing for power law distributions in §2.4. For a set of n nodes, let G be a symmetric, binary adjacency matrix that represents a network. Define the *individual clustering* for a node i under network G as

$$Cl_i(G) = \frac{\sum_{j \neq i; k \neq j; k \neq i} G_{ij} G_{ik} G_{jk}}{\sum_{j \neq i; k \neq j; k \neq i} G_{ij} G_{ik}},$$

with $Cl_i(G) \equiv 0$ if i has at most one link. The numerator counts the number of pairs (j, k) linked to i that are themselves linked, while the denominator counts the number of pairs linked to i. The *average clustering coefficient* of G is defined as $\sum_{i=1}^{n} Cl_i(G)/n$.

This statistic is a well-known measure of *transitivity* or *clustering*, the tendency for individuals with partners in common to associate. A well-known stylized fact in the network literature is that most social networks exhibit nontrivial clustering, where "nontrivial" is defined relative to the null model in which links are i.i.d. (Jackson, 2008). Under the null model, the average clustering coefficient in expectation equals the probability of forming a link, which is typically order n^{-1}, since most networks are sparse. Yet, the average clustering coefficient typically appears to be quite larger than zero in practice (Barabási, 2015, Ch. 3), hence the stylized fact.

In order to assess formally whether average clustering is significantly different from the probability of link formation, we can use the tests given by (2) with

$$X_i = Cl_i(G) - \frac{2}{n-1} \sum_{j=1}^{n} G_{ij}.$$

Then $\bar{X}$ is the difference between the average clustering coefficient and the empirical linking probability. To verify the main regularity condition that the sample clustering coefficient is $\sqrt{n}$-consistent, we can apply results in Bickel et al. (2011), Leung and Moon (2017), and Resnick and Samorodnitsky (2016), which prove CLTs for various classes of network statistics and network formation models.

[3]The degree of a node is the total number of links it forms.

2.3 Treatment Effects with Spillovers

Suppose we observe data from a randomized experiment on a single network, where for each node i, we observe an outcome Y_i, a binary treatment assignment D_i, the number of treated network neighbors T_i, and the number of such neighbors γ_i. Consider the following outcome model studied in Leung (2017a):

$$Y_i = r\left(D_i, T_i, \gamma_i, \varepsilon_i\right)$$

(also see Aronow and Samii, forthcoming). This departs from the conventional potential outcomes model by allowing $r(\cdot)$ to depend on T_i and γ_i, which violates the conventional stable unit treatment value assumption. The object of interest is the following measure of treatment/spillover effects:

$$\mathbf{E}\left[r(d, t, \gamma, \varepsilon_i(\gamma))\right] - \mathbf{E}\left[r(d', t', \gamma, \varepsilon_i(\gamma))\right], \tag{5}$$

where $t, t' \leqslant \gamma$, and $\varepsilon_i(\gamma)$ is the conditional distribution of ε_i given $\gamma_i = \gamma$. The latter allows for some dependence between the network and unobserved heterogeneity. Treatment effects are obtained from differences between d, d' and spillover effects from differences between t, t'. Leung (2017a) provides conditions on the network and dependence structure of $\{\varepsilon_i\}_{i=1}^n$ under which the sample analog of (5) is $\sqrt{n}$-consistent. In particular, we can allow ε_i and ε_j to be correlated if the network distance between i and j is at most M.

Suppose the econometrician obtains data $\{W_i\}_{i=1}^n$ for $W_i = (Y_i, D_i, T_i, \gamma_i)$ by snowball-sampling 1-neighborhoods. That is, she first obtains a random sample of units, from which she gathers (Y_i, D_i), and then she obtains the network neighbors of those units and their treatment assignment, from which she gathers (T_i, γ_i). This is a very common method of network sampling. However, standard error formulas provided by Aronow and Samii (forthcoming) and Leung (2017a) rely on knowing for each unit which alters are of path distance at most M, due to the dependence structure on the unobservables. For this, we would have to instead snowball-sample M-neighborhoods, which requires knowledge of M and can be costly in practice.

An alternative is to use randomized subsampling. Let $\mathbf{1}_i(d, t, \gamma) = \mathbf{1}\{D_i = d, T_i = t, \gamma_i = \gamma\}$. The frequency estimator for the average treatment/spillover effect is given by

$$\frac{\sum_{i=1}^n Y_i \mathbf{1}_i(d, t, \gamma)}{\sum_{i=1}^n \mathbf{1}_i(d, t, \gamma)} - \frac{\sum_{i=1}^n Y_i \mathbf{1}_i(d', t', \gamma)}{\sum_{i=1}^n \mathbf{1}_i(d', t', \gamma)}.$$

This fits into our setup, specifically (4), if we define $Z_i = W_i$,

$$\hat{\gamma} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_i(d, t, \gamma), \frac{1}{n}\sum_{i=1}^n \mathbf{1}_i(d', t', \gamma)\right),$$

and

$$\psi(Z_i; \beta_0, \hat{\gamma}) \equiv \frac{Y_i \mathbf{1}_i(d, t, \gamma)}{\frac{1}{n}\sum_{i=1}^n \mathbf{1}_i(d, t, \gamma)} - \frac{Y_i \mathbf{1}_i(d', t', \gamma)}{\frac{1}{n}\sum_{i=1}^n \mathbf{1}_i(d', t', \gamma)}.$$

Thus, we can construct confidence intervals for (5) using randomized subsampling with X_i given by (4).

2.4 Testing for Power Laws

Testing for whether the data follows a power law distribution is of wide empirical interest in economics, finance, network science, neuroscience, biology, and physics (Barabási, 2015; Gabaix, 2009; Klaus et al., 2011; Newman, 2005). Existing methods rely on the assumption of i.i.d. data, which is unrealistic for spatial and financial data. It is also unrealistic in network applications, which study the degree distribution, since even if links are formed i.i.d., the degrees of different nodes are dependent.

By "power law" we mean that the probability density or mass function of the data $f(x)$ is proportional to $x^{-\alpha}$ for some positive exponent α. Many methods are available for estimating α, for example maximum likelihood.[4] When the data is dependent, this becomes a pseudo-maximum likelihood, but the estimator is still consistent under weak dependence.

With an estimate of the power law exponent in hand, it is of interest to test how well the data accords with or deviates from a power law. Standard methods assume that the underlying data is independent, which motivates the use of randomized subsampling. The hypothesis we test is motivated by Klaus et al. (2011): the power law fits no better than some chosen null distribution, for example exponential or log-normal.[5] We operationalize this by testing the null that the log-likelihood ratio is less than or equal to zero, where the numerator of the likelihood is the power law distribution with exponent estimated using pseudo-maximum likelihood and the denominator is the estimated null distribution. This is a non-nested model selection test. Under general misspecification, the log-likelihood ratio is zero if both models are misspecified and poor fits and less than (greater than) zero if the null distribution fits better (worse) (Pesaran, 1987; Vuong, 1989). The reason we test the inequality $\leqslant 0$, rather than the equality $= 0$ as in Klaus et al. (2011), is that this literature is solely interested in making a decision about the power law hypothesis.[6]

Formally, for identically distributed data $\{Z_i\}_{i=1}^n$, let $\ell_{PL}(Z_i, \alpha)$ be the likelihood of observation i under a power law and $\ell_0(Z_i, \theta)$ the likelihood under the null distri-

[4]Regression estimators are also popular. See e.g. Ibragimov et al. (2015); Nicolau and Rodrigues (2015).

[5]It is common in practice to test for a power law using a Kolmogorov-Smirnov test (Clauset et al., 2009). However, as pointed out by Klaus et al. (2011), nonrejection does not constitute evidence for a power law, and with large enough samples, the test will eventually reject, since no distribution perfectly follows a power law in practice.

[6]It is certainly possible to use, for example, a randomized subsampling version of the Vuong test (Vuong, 1989) to test the equality, that the log-likelihood ratio is equal to zero against one-sided alternatives. However, this requires using the mean-type statistic, which has worse power compared to the moment inequalities test.

bution, which we assume to be parameterized by θ. Then the null is

$$H_0 : \mathbf{E}\left[\log \ell_{PL}(Z_i, \alpha) - \log \ell_0(Z_i, \theta)\right] \leqslant 0.$$

This fits into our setup (4) by defining $\hat{\gamma} = (\hat{\alpha}, \hat{\theta})$, the pseudo-maximum likelihood estimates of (α, θ), and

$$\psi(Z_i; \beta_0, \hat{\gamma}) = \log \ell_{PL}(Z_i, \hat{\alpha}) - \log \ell_0(Z_i, \hat{\theta}).$$

A key difference, however, is that this is a test of the moment inequality $\mu_0 \equiv \mathbf{E}[\psi(Z_i; \beta_0, \gamma)] \leqslant 0$. Because the methods in §2 only apply to tests of moment equalities, in §4, we develop a randomized subsampling procedure applicable to testing moment inequalities.

3 Large-Sample Theory

Consider a generalization of the setup in the previous section where $\mathbf{X}$ is a triangular array. Hence, X_i and μ_0 may implicitly depend on n, but we suppress this in the notation. This is important to accommodate the previous network applications. The next theorem provides conditions for asymptotic validity of the randomized subsampling procedures previously introduced.

Theorem 1. *Suppose $R_n \to \infty$ and $b_n^{-1} = O(1)$ as $n \to \infty$, and the following conditions hold.*

(a) $\sum_{i=1}^n (X_i - \mu_0)/\sqrt{n} = O_p(1)$.

(b) $\Sigma = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{E}[(X_i - \mu_0)(X_i - \mu_0)']/n$ *is finite and positive definite, and there exists an estimator $\hat{\Sigma}$ consistent for Σ.*

(c) $\frac{1}{n}\sum_{i=1}^n ||X_i||^3 = O_p(1)$.

If $R_n b_n/n = o(1)$, then

$$T_n(\mu_0; \pi) \xrightarrow{d} \mathcal{N}(0, I_m) \quad \textit{and} \quad T_n(\bar{X}; \pi) \xrightarrow{d} \mathcal{N}(0, I_m).$$

If $\frac{\sqrt{R_n b_n}}{n} \to 0$, then

$$S_n(\mu_0; \pi) \xrightarrow{d} \mathcal{N}(0, 1) \quad \textit{and} \quad S_n(\bar{X}; \pi) \xrightarrow{d} \mathcal{N}(0, 1).$$

Remark 1. Assumption (b) requires an asymptotically nondegenerate sample variance. Note that if a CLT exists, unless the data is independent, $\hat{\Sigma}$ will typically not be consistent for the asymptotic variance of $\bar{X}$. Song (2016) assumes locally dependent data and that 8th moments exist. Assumptions (a) and (c) substantially weaken these requirements, the latter of which is important for the power law application.

Remark 2. Assumption (a) allows X_i to depend on a "first-stage" estimator, which is important for many of the applications in §2.4 and asymptotically linear estimators. Suppose we have data $\mathbf{Z} = \{Z_i\}_{i=1}^n$, and $\hat{\theta}$ is an estimator of a vector $\theta_0 \in \Theta$ that is a function of $\mathbf{Z}$. Define $X_i = h(Z_i, \hat{\theta})$ and $\mu_0 = \mathbf{E}[h(Z_i, \theta_0)]$. The following are primitive conditions for the assumptions in Theorem 1:

(i) $\sum_{i=1}^n (h(Z_i, \theta_0) - \mu_0)/\sqrt{n}$ and $\sqrt{n}(\hat{\theta} - \theta_0)$ are $O_p(1)$.

(ii) Suppose $\mathbf{S} = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{E}[(h(Z_i, \theta_0) - \mu)(h(Z_i, \theta_0) - \mu)']/n$ is finite and positive definite. There exists an estimator $\hat{\mathbf{S}}$ consistent for $\mathbf{S}$.

(iii) $\sum_{i=1}^n ||h(Z_i, \theta_0)||^3/n = O_p(1)$.

(iv) $\sup_{\theta\in\Theta} \left| \sum_{i=1}^n \nabla_\theta h(X_i, \theta) - \mathbf{E}\left[\sum_{i=1}^n \nabla_\theta h(X_i, \theta)\right] \right| /n = o_p(1)$.

Note that for asymptotically linear estimators of the form (3), $h(\cdot) = \psi(\cdot, \beta_0, \cdot)$, and $\hat{\theta} = \hat{\gamma}$. Hence, the second requirement of (i) is $\sqrt{n}$-consistency of $\hat{\gamma}$, and the first requirement is

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(Z_i; \beta_0, \gamma) = O_p(1),$$

which follows from standard weak-dependence conditions.

Remark 3. As with all other subsampling procedures, the generality of randomized subsampling comes at the cost of having power against fewer sequences of alternatives. If the econometrician could consistently estimate the asymptotic variance of $\bar{X}$, then the usual trinity of tests would have power against local alternatives $\mu_n = \mu_0 + h/\sqrt{n}$. In contrast, it is easy to see that the test in (2) using the mean-type statistic only has nontrivial asymptotic power against alternatives $\mu_n = h/\alpha_n$, where $\alpha_n \to \infty$ but $\alpha_n(R_n b_n)^{-1/2} \to c \in [0, \infty)$. For the test using the U-type statistic, we have instead $\alpha_n R_n^{-1/4} b_n^{-1/2} \to c \in [0, \infty)$ (Song, 2016, Theorem 3.3). We interpret the lower power of randomized subsampling tests as the cost of having a dependence-robust procedure.

Remark 4. It should be straightforward to modify the rate conditions on R_n, b_n to allow for estimators that converge at rates slower than $\sqrt{n}$, for example kernel estimators. We do not pursue these extensions here because, in light of the previous remark, the power properties of the resulting tests will be rather poor.

Remark 5. The theorems prove convergence of $T_n(\bar{X}; \pi)$ and $S_n(\bar{X}; \pi)$. These results can be used to construct a permutation critical values, which may have better finite-sample properties than the asymptotic critical values suggested in §2. Let L be a natural number. For each $l = 1, \ldots, L$, let $\tilde{\pi}_l = (\tilde{\pi}_{1l}, \ldots, \tilde{\pi}_{Rl})$, where the components of $\tilde{\pi}$ are drawn independently and uniformly from Π. Following §3.2 of Song (2016),

the permutation critical value for the test in (2) using the mean-type statistic is

$$c_n(\alpha, \tilde{\pi}) = \inf\left\{c' > 0 : \frac{1}{L}\sum_{l=1}^{L} \mathbf{1}\left\{T_n(\bar{X}, \tilde{\pi}_l)' T_n(\bar{X}; \tilde{\pi}_l) > c'\right\} \leqslant \alpha\right\}.$$

For the U-type statistic we replace $T_n(\bar{X}, \tilde{\pi}_l)' T_n(\bar{X}; \tilde{\pi}_l)$ with $S_n(\bar{X}, \tilde{\pi}_l)$.

We give a quick sketch of the proof of Theorem 1 for mean-type statistics. Define

$$\tilde{X}_{T,r} = \frac{1}{\sqrt{b_n}} \sum_{i=1}^{b_n} \hat{\Sigma}^{-1/2}\left(X_{\pi_r(i)} - \mu_0\right). \tag{6}$$

Consider the decomposition

$$T_n(\mu_0; \pi) = \underbrace{\frac{1}{\sqrt{R_n}} \sum_{r=1}^{R_n} \left(\tilde{X}_{T,r} - \mathbf{E}[\tilde{X}_{T,r} \,|\, \mathbf{X}]\right)}_{[I]} + \underbrace{\frac{1}{\sqrt{R_n}} \sum_{r=1}^{R_n} \mathbf{E}[\tilde{X}_{T,r} \,|\, \mathbf{X}]}_{[II]}.$$

It is not too hard to calculate that

$$[II] = \sqrt{\frac{R_n b_n}{n}} \hat{\Sigma}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_0),$$

which is immediately $o_p(1)$ by the assumptions of the theorem. Since the random permutations are i.i.d. conditional on $\mathbf{X}$, we can then show that $[I] \xrightarrow{d} \mathcal{N}(0, I_m)$ using a martingale difference CLT. The proof for $S_n(\mu_0; \pi)$ follows a similar logic.

4 Testing Moment Inequalities

The results in the previous section are relevant testing the hypothesis that $\mu_0 = \mu$. In this section, we instead consider the null hypothesis

$$H_0 : \mu_0 \leqslant 0, \tag{7}$$

where "$\leqslant$" denotes a component-wise inequality. Let $T_{nk}(\mu_k; \pi)$ be the mean-type statistic applied to data $\{X_{ik}\}_{i=1}^n$, where X_{ik} is the kth component of X_i. We focus on the test statistic

$$Q_n(\pi) = \max_{1 \leqslant k \leqslant m} T_{nk}(0; \pi)$$

studied in, e.g., Romano et al. (2014). From the proof of the theorem below, it is evident that we could also use test statistics other than the max statistic above satisfying the regularity conditons of Andrews and Soares (2010).

Let $\hat{\lambda}_k = \sqrt{R_n b_n}\hat{\Sigma}_{kk}^{-1/2}\bar{X}_k$, where $\bar{X}_k$ is the kth component of $\bar{X}$ and $\hat{\Sigma}_{kk}$ is the kkth element of $\hat{\Sigma}$. Then

$$Q_n(\pi) = \max_{1 \leqslant k \leqslant m} \{T_{nk}(\bar{X}_k; \pi) + \hat{\lambda}_k\}. \tag{8}$$

We construct a critical value for $Q_n(\pi)$ by replacing $\hat{\lambda}_k$ with $\min\{\hat{\lambda}_k, 0\}$ (which is asymptotically equivalent only under the null) and resampling π L times and taking the appropriate quantile of the resulting permutation distribution of statistics, as in Remark 5. Formally, the critical value is

$$c_{1-\alpha} = \inf\left\{c' > 0 : \frac{1}{L}\sum_{l=1}^{L} \mathbf{1}\left\{\max_{1 \leqslant k \leqslant m} T_{nk}(\bar{X}_k; \pi) + \min\{\hat{\lambda}_k, 0\} > c'\right\} \leqslant \alpha\right\}.$$

Our proposed test is to reject if and only if $\phi_n = 1$ for

$$\phi_n \equiv \mathbf{1}\{Q_n(\pi) > c_{1-\alpha}\}.$$

Based on our simulations, we suggest choosing

$$(R_n, b_n) = (n^{0.4}, n^{0.4}).$$

Remark 6. For the case $m = 1$, dropping the subscript k, we have that $Q_n(\pi) - \hat{\lambda} \xrightarrow{d} \mathcal{N}(0, 1)$ directly from Theorem 3. We can therefore use the computationally simpler test $\tilde{\phi}_n = \mathbf{1}\{Q_n(\pi) - \min\{\hat{\lambda}, 0\} > q_{1-\alpha}\}$, where $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution.

We next show that the test uniformly controls size. Let $\lambda_{\min}(A)$ denote the smallest eigenvalue of a matrix A, and $\Sigma_P = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{E}_P[(X_i - \mu_0)(X_i - \mu_0)']/n$, where $\mathbf{E}_P[\cdot]$ denotes the expectation under the data-generating process (DGP) P.

Theorem 2. *Let $\mathbf{P}_0$ be the set of all DGPs any sequence of DGPs $\{P_n\}_{n\in\mathbb{N}}$ in $\mathbf{P}_0$, we have*

(a) $\sum_{i=1}^{n}(X_i - \mu_0)/\sqrt{n} = O_{P_n}(1)$.

(b) There exists an estimator $\hat{\Sigma}$ consistent for Σ_{P_n}. Additionally, $\limsup_{n\to\infty} ||\Sigma_{P_n}|| < \infty$ *and* $\liminf_{n\to\infty} \lambda_{min}(\Sigma_{P_n}) > 0$.

(c) $\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} ||X_i||^3 < \infty$.

If $R_n \to \infty$, $b_n^{-1} = O(1)$, and $R_n b_n/n = o(1)$ as $n \to \infty$, then $\sup_{P\in\mathbf{P}_0} \mathbf{E}[\phi_n] \to \alpha$.

The theorem follows fairly directly from the next proposition, which also provides results on the power of the test.

Proposition 1. *Suppose assumptions (a)-(c) of Theorem 1 hold, and the rate conditions on R_n, b_n in Theorem 2 hold. For $k = 1, \ldots, m$, let $\delta_k^* = \lim_{n\to\infty} \sqrt{R_n b_n}\Sigma_{kk}^{-1/2}\mu_{0k}$, where μ_{0k} is the kth component of μ_0.*

(a) If $\max_k \delta_k^* = -\infty$, then $\mathbf{E}[\phi_n] \to \alpha$.

(b) If $\max_k \delta_k^*$ is constant and weakly negative, then $\mathbf{E}[\phi_n] \to \alpha$.

(c) If $\max_k \delta_k^* = \infty$ for all k, then $\mathbf{E}[\phi_n] \to 1$.

(d) If $\max_k \delta_k^*$ is constant and strictly positive, then $\mathbf{E}[\phi_n] \to \beta \in (0, 1)$.

Note that Theorem 1 allows $\mathbf{X}$ to be a triangular array, which we also allow here. Parts (a) and (b) show that the test controls size and is asymptotically exact. Parts (c) and (d) describe the test's power, (d) showing that the test has power against local alternatives that vanish no faster than rate $(R_n b_n)^{-1/2}$. Since we need $R_n b_n/n \to 0$, this implies that it does not have power against $\sqrt{n}$ local alternatives, which is the same as conventional subsampling.

We sketch the proof of the proposition for the simple case where $m = 1$. Dropping the dependence on k, we have

$$Q_n(\pi) = T_n(\bar{X}; \pi) + \sqrt{R_n b_n}\hat{\Sigma}^{-1/2}(\bar{X} - \mu_0) + \sqrt{R_n b_n}\hat{\Sigma}^{-1/2}\mu_0. \tag{9}$$

The first term on the right-hand side is asymptotically distributed $\mathcal{N}(0, 1)$ by the proof of Theorem 3. The second term is $o_p(1)$ by the assumptions of the proposition. Hence, for $\delta^* = \lim_{n\to\infty} \sqrt{R_n b_n}\Sigma^{-1/2}\mu_0$, $Q_n(\pi)$ is approximately distributed

$$T_n(\bar{X}; \pi) + \delta^* \sim \mathcal{N}(0, 1) + \delta^*.$$

Critical values are instead constructed using the distribution of

$$T_n(\bar{X}; \pi) + \min\left\{\sqrt{R_n b_n}\hat{\Sigma}^{-1/2}(\bar{X} - \mu_0) + \sqrt{R_n b_n}\hat{\Sigma}^{-1/2}\mu_0, 0\right\}, \tag{10}$$

which has asymptotic distribution $\mathcal{N}(0, 1) + \min\{\delta^*, 0\}$. This is the same asymptotic distribution as that of (9) only under the null. Moreover, these arguments hold under any sequence of DGPs $\{P_n\}_{n\in\mathbb{N}}$ in $\mathbf{P}_0$.

Remark 7. Suppose $m = 1$, and consider the "conventional" moment-inequalities setting in which $\mathbf{X}$ is i.i.d. and the test statistic is $\sum_{i=1}^n \hat{\Sigma}^{-1/2}X_i/\sqrt{n}$. The problem with constructing critical values for this statistic is that while

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\Sigma}^{-1/2}(X_i - \mu_0) \xrightarrow{d} \mathcal{N}(0, I_m),$$

it is impossible to consistently estimate $\sqrt{n}\Sigma^{-1/2}\mu_0$. Much of the moment-inequalities literature boils down to finding clever ways to bound this nuisance parameter from

above (Canay and Shaikh, forthcoming). In contrast, in our setting the nuisance parameter is $\sqrt{R_n b_n}\Sigma^{-1/2}\mu_0$, which can be consistently estimated by $\sqrt{R_n b_n}\hat{\Sigma}^{-1/2}\bar{X}$, since $R_n b_n/n \to 0$.

Remark 8. Consider the case $m = 1$ and drop the subscript k. The bias-variance trade-off appears in the moment inequalities context when δ^* is a negative constant. To see this, recall that $Q_n(\pi)$ satisfies (9), whereas critical values are constructed using (10). Hence, if $R_n b_n = n$, then the bias term $\sqrt{R_n b_n}\hat{\Sigma}^{-1/2}(\bar{X} - \mu_0) \xrightarrow{d} \mathcal{N}(0, 1)$, and clearly the size of the test is incorrect. Thus, validity of the test requires $R_n b_n/n \to 0$, and the rate of convergence is $(R_n b_n)^{-1/2}$, which is the same type of bias-variance trade-off faced by the mean-type test in the equality-testing case.

5 Empirical Application

Jackson and Rogers (2007) propose a model of network formation that generates a parametric degree distribution, where a parameter r interpolates between the exponential and power law distributions. Their model provides microfoundations for the different distributions. When $r \to \infty$, the network is formed primarily through random meetings, and the distribution is exponential. When $r \to 0$, the network is formed primarily through "network-based meetings," as nodes are more likely to meet friends of nodes that were previously met randomly. High-degree nodes are more likely to be met through network-degree meetings, which corresponds to a "rich-get-richer" or "preferential-attachment" mechanism that generates a power law degree distribution.

Jackson and Rogers (2007) estimate the parameters of the degree distribution using data on six distinct social networks and informally assess the extent to which the estimated distributions depart from a power law. In this section, we use the same datasets to implement the test described in §2.4 and §4. This formally tests the null that an exponential degree distribution fits the data at least as well as a power law.

The test requires estimates of two parameters of the power law distribution, the lower support point and the power law exponent. For a fixed value of the lower support point, we estimate the exponent by (pseudo) maximum likelihood. We estimate the lower support point by minimizing the Kolmogorov-Smirnov distance between the empirical distribution and the power law distribution with exponent equal to the maximum likelihood estimate (see e.g. Clauset et al., 2009, p. 672). Based on our simulation results, we set the randomized subsampling parameters to be $(R_n, b_n) = (n^{0.4}, n^{0.4})$.

Jackson and Rogers (2007) describe the six datasets as follows: "the links among Web sites at Notre Dame University (*labeled 'WWW' below*), the network of coauthorship relations among economists publishing in journals listed by EconLit in the 1990s (*labeled 'Coauthor'*), a citation network of research articles stemming from Milgram's

1960 paper... (*labeled 'Citation'*), a friendship network among 67 prison inmates in the 1950s (*labeled 'Prison'*), a network of ham radio calls during a one-month period (*labeled 'Ham Radio'*), and, finally, a network of romantic relationships among high-school students (*labeled 'Romance'*)."

Table 1 displays the results of the tests. Row "Exponent" displays the estimated power law exponent, while row "xmin" displays the estimated lower support point. Row "LL" gives the log likelihood ratio. We reject the null for the WWW and citation networks and do not reject for the others. This is qualitatively consistent with the findings in Table 1 of Jackson and Rogers (2007). They estimate r to be close to zero only for the WWW and citation networks (respectively 0.57 and 0.63). The coauthor and ham radio networks yield respective estimates of 4.7 and 5.0. As the authors note, this means network-based meetings are eight times more common in the WWW network compared to the coauthor network, so their degree distributions should be closer to exponential than power law, which is consistent with our test results. Lastly, they estimate r to be infinite for the prison and romance networks, indicating that the distribution is very close to exponential for these networks. This is also consistent with our results.

Table 1: Power Law Tests.

	Coauthor	Ham Radio	Prison	Romance	Citation	WWW
# Nodes	56639	44	67	572	396	325729
Exponent	4.03	1.46	1.63	1.94	2.12	1.95
xmin	9	1	1	1	4	2
LL	-0.11	-1.69	-5.80	-13.75	0.92	25.44
Reject	N	N	N	N	Y	Y

6 Monte Carlo

This section presents results from four simulation studies exploring the finite-sample properties of the proposed tests, each corresponding to one of the applications in §2. For tests of moment equalities, we only present results for the U-type statistic using test (2) and asymptotic critical values. Our tests of moment inequalities only involve single inequalities, so we use the asymptotic critical value given in Remark 6. For both tests, due to space constraints, we only display results for pairs of tuning parameters that imply the same rate of convergence.

6.1 Cluster Dependence

Let c index cities, f index families, and i index individuals. We generate outcomes according to

$$Y_{ifc} = \theta_0 + \alpha_f + \varepsilon_{ifc},$$

where $\alpha_f \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\varepsilon_{ifc} \overset{iid}{\sim} \mathcal{N}(0,1)$. Thus, the true level of clustering is at the family level.

We first test the null that $\theta_0 = \theta$, for values of θ specified below, against a two-sided alternative at the 5 percent level. We present results for randomized subsampling and t-tests using Liang and Zeger (1986) clustered standard errors at each possible level of clustering. For these equality tests, we set the true $\theta_0 = 1$.

Table 2 displays simulation results for the size and power of the test, computed using 5000 simulations. The sample size is denoted by n, the number of individuals; the number of families is $n/2$; and the number of cities 20. The tuning parameters (R_n, b_n) are given by the floors of $(n^{\kappa_1}, n^{\kappa_2})$. Rows "RS $(\theta = \alpha)$" display results for the randomized subsampling test for $\theta = \alpha$. Rows "cluster α" display results for the t-test, where $\alpha = c$ means clustering at the city level, $\alpha = f$ the family level, and $\alpha = i$ the individual level.

The results show that the t-test overrejects when clustering at too coarse a level and the number of clusters is small (clustering at the city level). It also overrejects when clustering at too fine a level (clustering at the individual level), intuitively because this assumes more independence in the data than is warranted. Randomized subsampling controls size well when $(R_n, b_n) = (n^{2/3}, n^{1/3})$. When b_n is chosen larger ($(\kappa_1, \kappa_2) = (1/3, 1/2)$), the test tends to overreject, and when chosen smaller ($(\kappa_1, \kappa_2) = (1, 1/6)$), the test is underpowered. Note that the latter two choices of tuning parameters have the same rate of convergence as $(n^{2/3}, n^{1/3})$.

Table 2: Equality Test

n	400			4000		
(κ_1, κ_2)	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$
RS $(\theta = 1)$	5.24	1.24	6.98	4.92	2.18	7.48
RS $(\theta = 1.5)$	88.20	55.10	85.64	100	100	100
cluster c		7.34			7.16	
cluster f $(\theta = 1)$		5.72			4.88	
cluster f $(\theta = 1.5)$		99.98			100	
cluster i		11.48			10.88	

$\theta_0 = 1$. 5000 simulations. Significance level: 5%.

Next, we test the null that $\theta_0 \leqslant 0$ at the 5 percent level using the moment inequalities test in §4. Table 3 reports results computed using 5000 simulations. The

rows indicate the true value of θ_0. All choices of tuning parameters displayed satisfy $R_n b_n = n^{0.8}$, and all perform well.

Table 3: Inequality Test

n	400			4000		
(κ_1, κ_2)	$(0.2, 0.6)$	$(0.4, 0.4)$	$(0.6, 0.2)$	$(0.2, 0.6)$	$(0.4, 0.4)$	$(0.6, 0.2)$
$\theta_0 = -1$	4.54	4.62	5.24	4.98	5.00	4.80
$\theta_0 = -(R_n b_n)^{-1/2}$	5.18	5.20	5.88	5.24	5.38	5.10
$\theta_0 = (R_n b_n)^{-1/2}$	22.22	20.84	22.46	21.26	20.06	19.76
$\theta_0 = 1$	100	100	100	100	100	100

$H_0 : \theta_0 \leqslant 1$. 5000 simulations. Significance level: 5%.

6.2 Network Statistics

We generate a network according to a strategic model of network formation, following the Monte Carlo design in Leung (2017c) with $\theta = (0, 0.25, 0.25, 1)$. We are interested in two statistics that are functions of the network, the average clustering coefficient (defined in §2.2) and the average degree. Leung and Moon (2017) prove $\sqrt{n}$-consistency of the sample statistics for their population analogs.

We test the null that the expected value of the statistic is equal to θ against a two-sided alternative at the 5 percent level, for both average clustering (Table 4 in the appendix) and average degree (Table 5). The tables report rejection rates, with n denoting the network size and tuning parameters (R_n, b_n) given by the floors of $(n^{\kappa_1}, n^{\kappa_2})$. Row "RS $(\theta = \theta_0)$" displays results from the randomized subsampling test when θ is set to the true expectation θ_0. Row "t-test" displays rejection rates for $\theta = \theta_0$ for the naive t-test that assumes i.i.d. data. Rejection rates are computed by averaging across 5000 simulations.

The results show that setting $(R_n, b_n) = (n^{2/3}, n^{1/3})$ works well across the different sample sizes, producing rejection percentages fairly close to the nominal level. The t-test unsurprisingly substantially overrejects. Similar to the cluster-dependence results in the previous subsection, when b_n is chosen larger ($(\kappa_1, \kappa_2) = (1/3, 1/2)$), the test substantially overrejects, and when chosen smaller ($(\kappa_1, \kappa_2) = (1, 1/6)$), the test is very underpowered. Note that the latter two choices of tuning parameters have the same rate of convergence as $(n^{2/3}, n^{1/3})$.

In unreported results, we compute rejection rates under two different models of network formation and obtain similar results. One is the Erdős-Rényi model, where link-formation is i.i.d., and the linking probability is set to $7/n$, chosen to obtain a sparse network with limiting average degree of seven. The other is the following random geometric graph model: nodes are endowed with positions $X_i \overset{iid}{\sim} U([0, 1]^2)$, and nodes i and j form a link if and only if $||X_i - X_j|| \leqslant r_n$, where $||\cdot||$ is the

Euclidean norm and $r_n = (7/(n\pi))^{1/2}$. (The parameter r_n is chosen to obtain a sparse network with limiting average degree of seven.)

6.3 Treatment Effects with Network Spillovers

Consider the setup in §2.3. We assign units to treatment with probability 0.3, and draw the network from a strategic model of network formation following the Monte Carlo design in Leung (2017c) with $\theta = (-1, 0.25, 0.25, 1)$. We generate outcomes according to the linear model

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 T_i + \beta_4 \gamma_i + \varepsilon_i,$$

where $\varepsilon_i = \sum_j G_{ij}\nu_j / \sum_j G_{ij}$, with $\nu_j \overset{iid}{\sim} \mathcal{N}(0,1)$. This represents exogenous peer effects in unobservables. We set $(\beta_1, \beta_2, \beta_4) = (1, 2, 0.5)$, and β_3 is specified below.

Unless the network is very large, nonparametric estimators suffer from small effective sample sizes. We therefore consider a linear regression estimator of Y_i on $(1, D_i, T_i, \gamma_i)$.

For equality tests, we set the true β_2 equal to -2. We test the null that $\beta_3 = \beta$ at the 5 percent level. Table 6 in the appendix displays simulation results for the size and power of the test, computed using 5000 simulations. The number of nodes is denoted by n. The tuning parameters (R_n, b_n) are given by the floors of $(n^{\kappa_1}, n^{\kappa_2})$. The row labels display the hypothesized value of β_2 being tested. Hence, the first row concerns the size of the test and the second row the power. The results are fairly similar to those in the previous two subsections, although our preferred choice of tuning parameters $(n^{2/3}, n^{1/3})$ slightly underrejects, and $(n^{1/3}, n^{1/2})$ is closer to the nominal level.

We next test the null that the ASE is weakly negative at the 5 percent level using the moment inequalities test in §4. Table 7 in the appendix reports results computed using 5000 simulations. The row labels display the true value of β_2. Hence, the first two such rows concern the size of the test and the next two rows the power. Similar to the simulation experiments under cluster dependence, the displayed tuning parameter choices all perform similarly.

6.4 Testing for Power Laws

Following the notation in §4, the data $\{Z_i\}_{i=1}^n$ consists of the degrees of n nodes generated according to one of two network formation models: the Erdős-Rényi or preferential attachment model. In the former model, links are i.i.d. Bernoulli$(10/n)$, and thus the degree distribution has exponential tails. The preferential attachment model is the standard Barabási and Albert (1999) model, where the out-degree of a node equals 10, and that generates a limiting power law exponent of 3. Resnick and Samorodnitsky (2016) (Theorem 4.1) proves that the empirical degree distribution of a variant of the Barabási-Albert model is $\sqrt{n}$-consistent.

We test the null that the log-likelihood ratio in §2.4 is at most zero at the 5 percent level, where the null distribution is exponential and the alternative is the power law. Details for maximum likelihood estimation can be found in §5. Table 8 in the appendix reports rejection percentages, which are obtained by averaging across 5000 simulations, with $n = 5000$ and tuning parameters (R_n, b_n) given by the floors of $(n^{\kappa_1}, n^{\kappa_2})$. Row "Exp" is the average estimated power law exponent, "xmin" the average estimated lower support, "LL" the likelihood ratio, and "Reject" the rejection rate of the test. All choices of the tuning parameters displayed satisfy $R_n b_n = n^{0.8}$ and all perform well.

7 Conclusion

In this paper, we provide general and computationally simple methods for constructing confidence intervals and hypothesis tests that are robust to dependence. The basic procedure, first proposed by Song (2016), is to compute a resampled test statistic by averaging over random subsamples of the data, which does not require knowledge of the dependence structure. We extend Song's results, showing that randomized subsampling is valid for general forms of weakly dependent data. To illustrate the broad applicability of the results, we discuss in detail applications to clustering when the group structure is unknown, spatial data when locations are unobserved, and network data. We also devise a randomized subsampling test for moment inequalities. An interesting theoretical result is that, while conventional moment-inequalities procedures have to account for a nuisance parameter that cannot be consistently estimated, the nuisance parameter is actually known when using randomized subsampling, which leads to a simple and asymptotically exact test.

A Appendix

A.1 Proofs

Proof of Theorem 1. This is a corollary of Theorems 3 and 4 below. ■

Theorem 3 (Mean-Type Statistic)**.** *Suppose the following conditions hold.*

(a) *As $n \to \infty$, we have $R_n \to \infty$, $b_n^{-1} = O(1)$, and $R_n b_n / n = o(1)$.*

(b) $\sum_{i=1}^n (X_i - \mu_0)/\sqrt{n} = O_p(1)$.

(c) *Suppose $\Sigma = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{E}[(X_i - \mu_0)(X_i - \mu_0)']/n$ is finite and positive definite. There exists an estimator $\hat{\Sigma}$ consistent for Σ.*

(d) $\frac{1}{n}\sum_{i=1}^n ||X_i||^3 = O_p(1)$.

Then

$$T_n(\mu_0;\pi) \xrightarrow{d} \mathcal{N}(0, I_m) \quad \textit{and} \quad T_n(\bar{X};\pi) \xrightarrow{d} \mathcal{N}(0, I_m).$$

PROOF. Notice $T_n(\bar{X}) = T_n(\mu_0) + e_n$, where

$$e_n = \sqrt{\frac{R_n b_n}{n}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1/2}(X_i - \mu_0). \tag{11}$$

By assumptions (a)-(c), $e_n = o_p(1)$. Hence, it remains to establish asymptotic normality of $T_n(\mu_0)$.

Recall the definition of $\tilde{X}_{T,r}$ from (6), and consider the decomposition

$$T_n(\mu_0) = \underbrace{\frac{1}{\sqrt{R_n}} \sum_{r=1}^{R_n} \left(\tilde{X}_{T,r} - \mathbf{E}[\tilde{X}_{T,r} \,|\, \mathbf{X}] \right)}_{[I]} + \underbrace{\frac{1}{\sqrt{R_n}} \sum_{r=1}^{R_n} \mathbf{E}[\tilde{X}_{T,r} \,|\, \mathbf{X}]}_{[II]}.$$

By definition of π_r,

$$\begin{aligned}
[II] &= \sqrt{\frac{R_n}{b_n}} \hat{\Sigma}^{-1/2} \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \sum_{i=1}^{b_n} (X_{\pi(i)} - \mu_0) \\
&= \sqrt{\frac{R_n}{b_n}} \hat{\Sigma}^{-1/2} \frac{1}{n!} \sum_{i=1}^{b_n} \sum_{\pi \in \Pi} (X_{\pi(i)} - \mu_0) \\
&= \sqrt{\frac{R_n}{b_n}} \hat{\Sigma}^{-1/2} \frac{1}{n!} b_n \sum_{i=1}^{n} (X_i - \mu_0) \binom{n-1}{b_n - 1} (b_n - 1)!(n - b_n)! \\
&= \sqrt{\frac{R_n}{b_n}} \hat{\Sigma}^{-1/2} \frac{(n - b_n)!}{n!} b_n \sum_{i=1}^{n} (X_i - \mu_0) \binom{n-1}{b_n - 1} (b_n - 1)! \\
&= \sqrt{\frac{R_n b_n}{n}} \hat{\Sigma}^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_0).
\end{aligned} \tag{12}$$

To understand the third line, note that the number of times $X_i - \mu_0$ appears in the sum $\sum_{\pi \in \Pi}(X_{\pi(i)} - \mu_0)$ is equal to the number of permutation-subsamples of size b_n for which X_i is assigned the label of 1. The number of such permutation-subsamples is the number of ways one can choose the remaining $b_n - 1$ elements of the subsample from $n - 1$ units times the number of possible orderings of the subsample $(b_n - 1)!$ times the number of possible orderings of indices not included in the subsample $(n - b_n)!$. Since the right-hand side of (12) equals (11), we have established that $[II] = o_p(1)$.

It remains to show that $[I] \xrightarrow{d} \mathcal{N}(0, I_m)$. We will apply a martingale difference CLT, where the conditioning σ-algebra is that generated by $\mathbf{X}$ (e.g. Hall and Heyde, 2014, Corollary 3.1). First, we show that the conditional variance of $[I]$ converges in probability unconditionally to I_m. Note that

$$\begin{aligned}
\mathrm{Var}([I] \,|\, \mathbf{X}) &= \mathbf{E}[\tilde{X}_{T,r} \tilde{X}_{T,r}' \,|\, \mathbf{X}] - \mathbf{E}[\tilde{X}_{T,r} \,|\, \mathbf{X}] \mathbf{E}[\tilde{X}_{T,r} \,|\, \mathbf{X}]' \\
&= \mathbf{E}[\tilde{X}_{T,r} \tilde{X}_{T,r}' \,|\, \mathbf{X}] + o_p(1)
\end{aligned}$$

by our argument for $[II]$. Now,

$$\begin{aligned}
\mathbf{E}[\tilde{X}_{T,r}\tilde{X}'_{T,r}\,|\,\mathbf{X}] &= \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\frac{1}{b_n}\sum_{i=1}^{b_n}\sum_{j=1}^{b_n}\hat{\Sigma}^{-1/2}(X_{\pi(i)}-\mu_0)(X_{\pi(j)}-\mu_0)'(\hat{\Sigma}^{-1/2})' \\
&= \frac{1}{n!}\left[\sum_{i=1}^{n}\hat{\Sigma}^{-1/2}(X_i-\mu_0)(X_i-\mu_0)'(\hat{\Sigma}^{-1/2})'(n-1)!\right. \\
&\qquad\left. + (b_n-1)\sum_{i=1}^{n}\sum_{j\neq i}\hat{\Sigma}^{-1/2}(X_i-\mu_0)(X_j-\mu_0)'(\hat{\Sigma}^{-1/2})'(n-2)!\right] \\
&= \left(1-\frac{b_n}{n(n-1)}\right)\frac{1}{n}\sum_{i=1}^{n}\hat{\Sigma}^{-1/2}(X_i-\mu_0)(X_i-\mu_0)'(\hat{\Sigma}^{-1/2})' \\
&\qquad + \frac{b_n-1}{n(n-1)}\hat{\Sigma}^{-1/2}(\bar{X}-\mu_0)(\bar{X}-\mu_0)'(\hat{\Sigma}^{-1/2})'.
\end{aligned}$$

The first term in the last equation converges in probability to I_m by assumptions (b) and (c). The second term is $o_p(1)$ by assumptions (a)-(c). Thus, we obtain

$$\mathrm{Var}(T_n(\mu_0)\,|\,\mathbf{X}) \xrightarrow{p} I_m.$$

It remains to check the Lindeberg condition. Let μ_{0s} denote the sth component of μ_0. Since $\mathbf{E}[\tilde{X}_{T,r}\,|\,\mathbf{X}] = o_p(1)$ by (12), it is enough to show that

$$\mathbf{E}\left[\left(\frac{1}{\sqrt{b_n}}\sum_{i=1}^{b_n}\left(X_{\pi_r(i)s}-\mu_{0s}\right)\right)^3\,\middle|\,\mathbf{X}\right] = O_p(1)$$

for any $s=1,\ldots,m$. Abbreviate $\tilde{X}_{is} = X_{is}-\mu_{0s}$. Then

$$\begin{aligned}
&\mathbf{E}\left[\left(\frac{1}{\sqrt{b_n}}\sum_{i=1}^{b_n}\tilde{X}_{\pi_r(i)s}\right)^3\,\middle|\,\mathbf{X}\right] = \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\frac{1}{b_n^{3/2}}\sum_{i,j,k}^{b_n}\tilde{X}_{\pi(i)s}\tilde{X}_{\pi(j)s}\tilde{X}_{\pi(k)s} \\
&= \frac{1}{n!}\left[\frac{1}{\sqrt{b_n}}\sum_{i=1}^{n}\tilde{X}_{is}^3\frac{(n-1)!}{(n-b_n)!(b_n-1)!}(b_n-1)!(n-b_n)! + 3\frac{b_n-1}{\sqrt{b_n}}\sum_{i\neq j}^{n}\tilde{X}_{is}^2\tilde{X}_{js}(n-2)!\right. \\
&\qquad\left. \frac{(b_n-1)(b_n-2)}{\sqrt{b_n}}\sum_{i\neq j\neq k}^{n}\tilde{X}_{is}\tilde{X}_{js}\tilde{X}_{ks}(n-3)!\right].
\end{aligned}$$

This is the same asymptotic order as

$$\frac{1}{\sqrt{b_n}}\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{is}^3 + 3\sqrt{\frac{b_n}{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{is}^2\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{X}_{is}\right) + \left(\frac{b_n}{n}\right)^{3/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{X}_{is}\right)^3,$$

which is $O_p(1)$ by assumptions (a)-(d). ∎

Theorem 4 (U-Type Statistic)**.** *Define*

$$\tilde{S}_n(\mu_0;\pi) = \frac{1}{\sqrt{R_n}} \sum_{r=1}^{R_n} \frac{1}{b_n} \sum_{i=1}^{b_n} \sum_{j=1,j\neq i}^{b_n} (X_{\pi_r(i)} - \mu_0)' \hat{\Sigma}^{-1} (X_{\pi_r(j)} - \mu_0). \tag{13}$$

Suppose the assumptions of Theorem 3 hold, except with (a) *replaced with*

$$R_n \to \infty, \quad b_n^{-1} = O(1), \quad \frac{\sqrt{R_n} b_n}{n} = o(1), \tag{14}$$

as $n \to \infty$*. Then*

$$S_n(\mu_0;\pi) \xrightarrow{d} \mathcal{N}(0,2) \quad \textit{and} \quad S_n(\bar{X};\pi) \xrightarrow{d} \mathcal{N}(0,2).$$

PROOF. **Step 1.** We first show that $\tilde{S}_n(\bar{X};\pi) = \tilde{S}_n(\mu_0;\pi) + A_n + o_p(1)$ for some term A_n defined below. We have

$$\begin{aligned}
&\tilde{S}_n(\bar{X};\pi) \pm \tilde{S}_n(\mu_0;\pi) = \tilde{S}_n(\mu_0;\pi) \\
&+ \frac{1}{\sqrt{R_n} b_n} \sum_{r=1}^{R} \sum_{i=1}^{b_n} \sum_{j\neq i} \Big(-(\bar{X} - \mu_0)' \hat{\Sigma}^{-1} X_{\pi_r(j)} - X_{\pi_r(i)}' \hat{\Sigma}^{-1} (\bar{X} - \mu_0) + \bar{X}' \hat{\Sigma}^{-1} \bar{X} - \mu_0' \hat{\Sigma}^{-1} \mu_0 \Big).
\end{aligned}$$

From the right-hand side, add and subtract

$$\frac{1}{\sqrt{R_n} b_n} \sum_{r=1}^{R} \sum_{i=1}^{b_n} \sum_{j\neq i} \Big(-(\bar{X} - \mu_0)' \hat{\Sigma}^{-1} (-\mu_0) - (-\mu_0)' \hat{\Sigma}^{-1} (\bar{X} - \mu_0) \Big)$$

to obtain

$$\begin{aligned}
\tilde{S}_n(\bar{X};\pi) = \tilde{S}_n(\mu_0;\pi) - (\bar{X} - \mu_0)' \hat{\Sigma}^{-1} \frac{1}{\sqrt{R_n} b_n} \sum_{r=1}^{R} \sum_{i=1}^{b_n} \sum_{j\neq i} \big((X_{\pi_r(i)} - \mu_0) + (X_{\pi_r(j)} - \mu_0) \big) \\
+ \frac{\sqrt{R_n}(b_n - 1)}{n} \sqrt{n}(\bar{X} - \mu_0)' \hat{\Sigma}^{-1} \sqrt{n}(\bar{X} - \mu_0).
\end{aligned}$$

By assumption (14) of this theorem and assumption (b) of Theorem 3, the third term on the right-hand side is $o_p(1)$. Call the second term on the right-hand side A_n.

Step 2. We show that $A_n = o_p(1)$, from which it follows that $\tilde{S}_n(\bar{X};\pi)$ and $\tilde{S}_n(\mu_0;\pi)$ have the same asymptotic distribution. Notice that A_n equals -1 times the sum of two virtually identical terms (they are equal up to an $o_p(1)$ term), one of which is

$$\begin{aligned}
(\bar{X} - \mu_0)' \hat{\Sigma}^{-1} \frac{1}{\sqrt{R_n} b_n} \sum_{r=1}^{R} \sum_{i=1}^{b_n} \sum_{j\neq i} (X_{\pi_r(i)} - \mu_0) = \frac{\sqrt{R_n}(b_n - 1)}{n} \sqrt{n}(\bar{X} - \mu_0)' \hat{\Sigma}^{-1} \sqrt{n}(\bar{X} - \mu_0) \\
+ \frac{\sqrt{b_n}(b_n - 1)}{b_n} (\bar{X} - \mu_0)' \hat{\Sigma}^{-1} \frac{1}{\sqrt{R_n} b_n} \sum_{r=1}^{R_n} \sum_{i=1}^{b_n} (X_{\pi_r(i)} - \bar{X}).
\end{aligned}$$

The first term on the right-hand side was shown to be $o_p(1)$ in step 1. The second term equals

$$\sqrt{\frac{\sqrt{R_n}b_n}{n}}\frac{b_n-1}{b_n}\sqrt{n}(\bar{X}-\mu_0)'\hat{\Sigma}^{-1/2}\underbrace{\frac{1}{R_n^{1/4}}\frac{1}{\sqrt{R_n b_n}}\sum_{r=1}^{R_n}\sum_{i=1}^{b_n}\hat{\Sigma}^{-1/2}(X_{\pi_r(i)}-\bar{X})}_{B_n}.$$

Following the exact same steps as the proof of Theorem 3, we can show that $B_n = o_p(1)$. In particular, by assumption (14) of this theorem and assumptions (b) and (c) of Theorem 3, we have

$$e_n = [II] = \sqrt{\frac{\sqrt{R_n}b_n}{n}}\hat{\Sigma}^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu_0) = o_p(1).$$

Furthermore, $[I]$ is now multiplied by an extra $R_n^{-1/4}$ term and is therefore $o_p(1)$. Hence, $A_n = o_p(1)$.

Step 3. Decompose

$$\tilde{S}_n(\mu_0;\pi) = (\tilde{S}_n(\mu_0;\pi) - \mathbf{E}[\tilde{S}_n(\mu_0;\pi)\,|\,\mathbf{X}]) + \mathbf{E}[\tilde{S}_n(\mu_0;\pi)\,|\,\mathbf{X}]. \tag{15}$$

We show that $\mathbf{E}[\tilde{S}_n(\mu_0;\pi)\,|\,\mathbf{X}] \xrightarrow{p} 0$:

$$\begin{aligned}
\mathbf{E}[\tilde{S}_n(\mu_0;\pi)\,|\,\mathbf{X}] &= \frac{1}{\sqrt{R_n}b_n}\sum_{r=1}^{R_n}\sum_{i=1}^{b_n}\sum_{j\neq i}\mathbf{E}\left[(X_{\pi_r(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi_r(j)}-\mu_0)\,\middle|\,\mathbf{X}\right] \\
&= \frac{1}{|\Pi|}\sum_{\pi\in\Pi}\sqrt{R_n}\frac{1}{b_n}\sum_{i=1}^{b_n}\sum_{j\neq i}(X_{\pi(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi(j)}-\mu_0) \\
&= \sqrt{R_n}\frac{1}{n!b_n}\sum_{i=1}^{b_n}\sum_{j\neq i}\sum_{\pi\in\Pi}(X_{\pi(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi(j)}-\mu_0) \\
&= \sqrt{R_n}\frac{(n-b_n)!}{n!}(b_n-1)\frac{(n-2)!}{(n-b_n)!}\sum_{i=1}^{b_n}\sum_{j\neq i}\sum_{\pi\in\Pi}(X_{\pi(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi(j)}-\mu_0) \\
&= \sqrt{R_n}(b_n-1)\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_j-\mu_0).
\end{aligned}$$

From the last line, add and subtract

$$\sqrt{R_n}(b_n-1)\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0)$$

to obtain

$$\frac{\sqrt{R_n}(b_n-1)}{n-1}\left(\sqrt{n}(\bar{X}-\mu_0)'\hat{\Sigma}^{-1}\sqrt{n}(\bar{X}-\mu_0) - \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0)\right).$$

Assumptions (b) and (c) of Theorem 3 imply that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0)\xrightarrow{p}1.$$

Assumption (14) then establishes the claim.

Step 4. In view of (15) and step 3, it remains to show that

$$\tilde{S}_n(\mu_0;\pi)-\mathbf{E}[\tilde{S}_n(\mu_0;\pi)\,|\,\mathbf{X}]\xrightarrow{d}\mathcal{N}(0,2).$$

We will apply a martingale difference CLT, where the conditioning σ-algebra is that generated by $\mathbf{X}$ (e.g. Hall and Heyde, 2014, Corollary 3.1). In this step, we show that the conditional variance converges in probability to 2. Let

$$\tilde{X}_r=\frac{1}{b_n}\sum_{i=1}^{b_n}\sum_{j\neq i}(X_{\pi_r(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi_r(j)}-\mu_0).$$

Then $\mathrm{Var}(\tilde{S}_n(\mu_0;\pi)\,|\,\mathbf{X})=\mathbf{E}[\tilde{X}_r^2\,|\,\mathbf{X}]-\mathbf{E}[\tilde{X}_r\,|\,\mathbf{X}]^2$, where the second term on the right-hand side is $o_p(1)$ by step 3 above. Now,

$$\begin{aligned}\mathbf{E}[\tilde{X}_r^2\,|\,\mathbf{X}]&=\frac{1}{b_n^2}\frac{1}{|\Pi|}\sum_{\pi\in\Pi}\sum_{i=1}^{b_n}\sum_{j\neq i}\sum_{k=1}^{b_n}\sum_{l\neq k}(X_{\pi_r(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi_r(j)}-\mu_0)(X_{\pi_r(k)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi_r(l)}-\mu_0)\\&=\frac{1}{b_n^2}\frac{1}{n!}\left(2b_n!\sum_{i=1}^{n}\sum_{j\neq i}\left[(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_j-\mu_0)\right]^2\frac{(n-2)!}{(n-b_n)!}\right.\\&\quad+4\frac{b_n!}{(b_n-3)!}\sum_{i=1}^{n}\sum_{j\neq k\neq i}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_j-\mu_0)(X_k-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0)\frac{(n-3)!}{(b_n-2)!}\\&\quad\left.+b_n!\sum_{i=1}^{n}\sum_{j\neq k\neq l\neq i}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_j-\mu_0)(X_k-\mu_0)'\hat{\Sigma}^{-1}(X_l-\mu_0)\frac{(n-4)!}{(b_n-4)!}\right).\end{aligned}\tag{16}$$

The first term on the right-hand side equals

$$2\frac{b_n-1}{b_n}\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)'\frac{1}{n-1}\sum_{j\neq i}\hat{\Sigma}^{-1}(X_j-\mu_0)(X_j-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0),$$

which converges in probability to 2 by assumption (c) of Theorem 3. The second term on the right-hand side of (16) equals

$$4\frac{b_n-1}{b_n}\frac{b_n-2}{n-2}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i}(X_j-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0)(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_k-\mu_0).$$

Under the assumptions of this theorem, this is of the same asymptotic order as

$$4\frac{b_n}{n}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(X_j-\mu_0)'\right)\hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)(X_i-\mu_0)'\right)\hat{\Sigma}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(X_k-\mu_0)\right),$$

which is $o_p(1)$. Lastly, the third term on the right-hand side of (16) is of the same asymptotic order as

$$\frac{b_n^2}{n^2}\left[\sqrt{n}(\bar{X}-\mu_0)\hat{\Sigma}^{-1}\sqrt{n}(\bar{X}-\mu_0)\right]^2 = o_p(1).$$

Step 5. It remains to verify the Lindeberg condition. Since $\mathbf{E}[\tilde{X}_r \,|\, \mathbf{X}] = O_p(1)$ by step 4, it is enough to show that

$$\mathbf{E}\left[\left(\frac{1}{b_n}\sum_{i=1}^{b_n}\sum_{j\neq i}(X_{\pi_r(i)}-\mu_0)'\hat{\Sigma}^{-1}(X_{\pi_r(j)}-\mu_0)\right)^3 \,\middle|\, \mathbf{X}\right] = O_p(1)$$

Similar to step 4, we break this expectation into a sum of multiple summations and show that each is $o_p(1)$. This is tedious algebra, and we only show below calculations for a few representative terms. One term is a summation over six distinct indices, which, following calculations in step 4, equals

$$\frac{1}{b_n^3}\frac{b_n!}{n!}\sum_{i\neq j\neq k\neq l\neq m\neq p}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_j-\mu_0) \times (X_k-\mu_0)'\hat{\Sigma}^{-1}(X_l-\mu_0)(X_m-\mu_0)'\hat{\Sigma}^{-1}(X_p-\mu_0)\frac{(n-6)!}{(b_n-6)!}.$$

This is of the same asymptotic order as

$$\frac{b_n^3}{n^3}\left[\sqrt{n}(\bar{X}-\mu_0)'\hat{\Sigma}^{-1}\sqrt{n}(\bar{X}-\mu_0)\right]^3 = O_p(1).$$

Several terms will be summations over five distinct indices (so that, e.g. index i equals index p and all others are distinct). These are of the same asymptotic order as

$$\frac{b_n^2}{n^2}\sqrt{n}(\bar{X}-\mu_0)'\hat{\Sigma}^{-1}\sqrt{n}(\bar{X}-\mu_0)\hat{\Sigma}^{-1} \times \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0)\sqrt{n}(\bar{X}-\mu_0)'\hat{\Sigma}^{-1}\sqrt{n}(\bar{X}-\mu_0) = O_p(1).$$

Several terms will be summations over 4 distinct indices (so that, e.g. index $i=k=m$ and all others are distinct), equal to

$$\frac{1}{b_n^3}\frac{b_n!}{n!}(n-b_n)!\sum_{i\neq j\neq l\neq p}(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_j-\mu_0) \times (X_i-\mu_0)'\hat{\Sigma}^{-1}(X_l-\mu_0)(X_i-\mu_0)'\hat{\Sigma}^{-1}(X_p-\mu_0)\frac{(n-4)!}{(b_n-4)!}.$$

This is of the same asymptotic order as

$$\frac{b_n}{n^{3/2}}\sqrt{n}(\bar{X}-\mu_0)\hat{\Sigma}^{-1}\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)(X_i-\mu_0)'\hat{\Sigma}^{-1}\sqrt{n}(\bar{X}-\mu_0)\sqrt{n}(\bar{X}-\mu_0)'\hat{\Sigma}^{-1}(X_i-\mu_0),$$

which is $O_p(1)$ using (d) of Theorem 3 to account for the presence of three $(X_i - \mu_0)$ in the summation over i.

The last terms we will consider are the summations over two indices (so that, e.g. indices $i = k = m$ and $j = l = p$ but they are otherwise distinct), equal to

$$\frac{1}{b_n^3} \frac{b_n!}{n!} \sum_{i \neq j} \left[(X_i - \mu_0)' \hat{\Sigma}^{-1} (X_j - \mu_0) \right]^3 \frac{(n-2)!}{(b_n - 2)!}.$$

This is of the same asymptotic order as

$$\frac{1}{b_n} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(X_i - \mu_0)' \hat{\Sigma}^{-1} (X_j - \mu_0) \right]^3.$$

The average over i, j is $O_p(1)$ under assumptions (c) and (d) of Theorem 3. Hence, the term is $O_p(1)$ by assumption (14). ■

PROOF OF THEOREM 2. We first establish that $\sup_{P \in \mathbf{P}_0} \mathbf{E}[\phi_n] \to \alpha' \leqslant \alpha$. If not, then we can find some sequence $\{P_n\}_{n \in \mathbb{N}}$ satisfying the assumptions of the theorem under which $\liminf_{n \to \infty} \mathbf{E}[\phi_n] > \alpha$. This contradicts conclusions (a) and (b) of Proposition 1. Lastly, the test is asymptotically exact because $\mathbf{P}_0$ includes DGPs in which $\mu_0 = 0$, which falls under conclusion (b) of Proposition 1. ■

PROOF OF PROPOSITION 1. Note that the setup of this Proposition allows $\mathbf{X}$ to be a triangular array, as in §3. We have that

$$(T_{nk}(0; \pi); k = 1, \ldots, m) = (T_{nk}(\bar{X}_k; \pi) + \hat{\lambda}_k; k = 1, \ldots, m), \tag{17}$$

and by the assumptions of the theorem,

$$\hat{\lambda}_k = \sqrt{R_n b_n} \hat{\Sigma}_{kk}^{-1/2} \mu_{0k} + o_p(1).$$

Thus,

$$Q_n(\pi) = \max_k \left\{ \hat{\Sigma}_{kk}^{-1/2} \left(\frac{1}{\sqrt{R_n b_n}} \sum_{r=1}^{R_n} \sum_{i=1}^{b_n} (X_{\pi_r(i),k} - \bar{X}_k) + \sqrt{R_n b_n} \mu_{0k} + o_p(1) \right) \right\}.$$

By assumption, $\hat{\Sigma}_{kk}^{1/2} / \Sigma_{kk}^{1/2} \xrightarrow{p} 1$ for all k. By the proof of Theorem 3, for any $\tilde{\lambda} \in \mathbb{R}_-^m$,

$$\left(\frac{1}{\sqrt{R_n b_n}} \sum_{r=1}^{R_n} \sum_{i=1}^{b_n} (X_{\pi_r(i),k} - \bar{X}_k) + \tilde{\lambda}_k; k = 1, \ldots, m \right) \xrightarrow{d} \mathcal{N}(\tilde{\lambda}, \Sigma). \tag{18}$$

Now consider cases (a) and (b). Let $\mathcal{Z} \sim \mathcal{N}(0, \tilde{\Sigma})$ for $\tilde{\Sigma}_{jk} = \Sigma_{jk} / (\Sigma_{jj} \Sigma_{kk})^{1/2}$. By Slutsky's theorem and Polya's theorem, the CDF of

$$\max_k \{ T_{nk}(\bar{X}_k; \pi) + \hat{\Sigma}_{kk}^{-1/2} \tilde{\lambda}_k \}$$

converges in sup norm to the CDF of

$$\max_k\{\mathcal{Z}_k + \Sigma_{kk}^{-1/2}\tilde{\lambda}_k\}.$$

Moreover, this convergence holds uniformly in $\tilde{\lambda} \in \mathbb{R}^m_-$ (see e.g. Romano and Shaikh, 2008). It therefore holds for $\tilde{\lambda} = \sqrt{R_n b_n}\mu_0$. Then (a) and (b) follow from the fact $\min\{\hat{\lambda}_k, 0\} - \hat{\lambda}_k \xrightarrow{p} 0$ for all k and the max function is continuous and strictly increasing in the largest component of its input.

For case (b), we have (17) $\xrightarrow{p} \infty$ and $\min\{\hat{\lambda}_k, 0\} \xrightarrow{p} 0$ for each k. The claim follows from the fact that the max function is continuous and strictly increasing in the largest component of its input.

For case (c), without loss of generality suppose that δ_k^* is finite and positive for $k = 1, \ldots, \ell_l$ and finite and negative for $k = \ell_1 + 1, \ldots, \ell_2$. Then

$$\max_k (17) \xrightarrow{d} \max_{k=1,\ldots,\ell}\{\mathcal{N}(0, \tilde{\Sigma}) + \delta_k^*\},$$

$\min\{\hat{\lambda}_k, 0\} \xrightarrow{p} 0$ for $k = 1, \ldots, \ell_1$, and $\min\{\hat{\lambda}_k, 0\} \xrightarrow{p} \delta_k^*$ for $k = \ell_1 + 1, \ldots, \ell_2$. Thus,

$$\mathbf{E}[\phi_n] \to \mathbf{P}\left(\max_{k=1,\ldots,\ell}\{\mathcal{N}(0, \tilde{\Sigma}) + \delta_k^*\} > q_{1-\alpha}\right),$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of $\max_{k=1,\ldots,\ell}\{\mathcal{N}(0, \tilde{\Sigma}) + \tilde{\delta}\}$, where $\tilde{\delta}_k = 0$ for $k = 1, \ldots, \ell$ and $\tilde{\delta}_k = \delta_k^*$ for $k = \ell_1 + 1, \ldots, \ell_2$. The claim follows. ■

A.2 Tables

Table 4: Average Clustering.

n	100			500			1000		
(κ_1, κ_2)	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$
RS test ($\theta = \theta_0$)	6.26	1.86	9.68	6.04	1.34	8.66	5.60	3.34	7.98
RS test ($\theta = \theta_0 + 0.1$)	51.00	37.72	54.86	99.62	94.52	98.74	100	100	92.16
t-test ($\theta = \theta_0$)		18.24			18.36			19.58	

5000 simulations. Significance level: 5%.

Table 5: Average Degree.

n	100			500			1000		
(κ_1, κ_2)	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$
RS test ($\theta = \theta_0$)	7.28	2.84	10.22	5.88	1.14	8.22	6.30	2.72	8.82
RS test ($\theta = \theta_0 + 1$)	43.48	34.94	49.18	89.98	65.62	87.92	99.34	98.94	98.44
t-test		21.72			19.66			19.48	

5000 simulations. Significance level: 5%.

Table 6: Treatment Spillovers (Equality Test)

n	100			500			1000		
(κ_1, κ_2)	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$	$(2/3, 1/3)$	$(1, 1/6)$	$(1/3, 1/2)$
$H_0 : \beta_2 = -2$	2.72	1.32	3.44	3.56	1.84	4.64	3.90	2.42	4.90
$H_0 : \beta_2 = -3$	15.40	6.42	25.18	66.88	22.30	68.16	93.30	90.52	88.48

True $\beta_2 = -2$. 5000 simulations. Significance level: 5%.

Table 7: Treatment Spillovers (Inequality Test)

n	500			1000			2500		
(κ_1, κ_2)	$(0.6, 0.2)$	$(0.4, 0.4)$	$(0.2, 0.6)$	$(0.6, 0.2)$	$(0.4, 0.4)$	$(0.2, 0.6)$	$(0.6, 0.2)$	$(0.4, 0.4)$	$(0.2, 0.6)$
$\beta_2 = -2$	4.12	3.08	2.68	4.14	3.84	3.26	3.96	4.24	3.80
$\beta_2 = -(R_n b_n)^{-1/2}$	4.94	4.72	4.30	4.70	4.78	4.30	5.16	5.24	4.90
$\beta_2 = (R_n b_n)^{-1/2}$	7.62	6.96	6.16	7.08	7.26	6.64	7.20	7.50	7.42
$\beta_2 = 2$	27.84	29.66	26.38	69.00	75.06	70.72	85.88	89.84	85.66

$H_0 : \beta_2 \leqslant 0$. 5000 simulations. Significance level: 5%.

Table 8: Power Law Test.

	Erdős-Rényi			Preferential Attachment		
(κ_1, κ_2)	$(0.2, 0.6)$	$(0.4, 0.4)$	$(0.6, 0.2)$	$(0.2, 0.6)$	$(0.4, 0.4)$	$(0.6, 0.2)$
Exponent		5.09			2.87	
xmin		10.00			11.32	
LL		-19.35			9.17	
Reject %	4.30	5.40	5.50	100	100	99.90

$n = 5000$. 5000 simulations. Significance level: 5%.

References

Acemoglu, D., C. Garcia-Jimeno, and J. Robinson, "State Capacity and Economic Development: A Network Approach," *American Economic Review*, 2015, *105* (8), 2364–2409.

Andrews, D. and G. Soares, "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 2010, *78* (1), 119–157.

Aronow, P. and C. Samii, "Estimating Average Causal Effects under General Interference, with Application to a Social Network Experiment," *Annals of Applied Statistics*, forthcoming.

Bakirov, N. and G. Székely, "Student's t-Test for Gaussian Scale Mixtures," *Journal of Mathematical Sciences*, 2006, *139* (3), 6497–6505.

Barabási, A., *Network Science*, Cambridge University Press, 2015.

Barabási, A. and R. Albert, "Emergence of Scaling in Random Networks," *Science*, 1999, *286* (5439), 509–512.

Barrios, T., R. Diamond, G. Imbens, and M. Kolesár, "Clustering, Spatial Correlations, and Randomization Inference," *Journal of the American Statistical Association*, 2012, *107* (498), 578–591.

Bertrand, M., E. Duflo, and S. Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *The Quarterly Journal of Economics*, 2004, *119* (1), 249–275.

Bester, C., T. Conley, and C. Hansen, "Inference with Dependent Data Using Clustering Covariance Matrix Estimators," *Journal of Econometrics*, 2011, *165* (2), 137–151.

Bickel, P., A. Chen, and E. Levina, "The Method of Moments and Degree Distributions for Network Models," *The Annals of Statistics*, 2011, *39* (5), 2280–2301.

Cameron, C. and D. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

_ , J. Gelbach, and D. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 2008, *90* (3), 414–427.

Canay, I. and A. Shaikh, "Practical and Theoretical Advances for Inference in Partially Identified Models," *Advances in Economics and Econometrics: 11th World Congress of the Econometric Society*, forthcoming.

_ , J. Romano, and A. Shaikh, "Randomization Tests Under an Approximate Symmetry Assumption," *Econometrica*, 2017, *85* (3), 1013–1030.

Clauset, A., C. Shalizi, and M. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, 2009, *51* (4), 661–703.

de Paula, A., S. Richards-Shubik, and E. Tamer, "Identifying Preferences in Networks with Bounded Degree," *Econometrica*, forthcoming.

Gabaix, X., "Power Laws in Economics and Finance," *Annual Review of Economics*, 2009, *1* (1), 255–294.

Hall, P. and C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, 2014.

Ibragimov, M., R. Ibragimov, and J. Walden, *Heavy-Tailed Distributions and Robustness in Economics and Finance*, Vol. 214, Springer, 2015.

Ibragimov, R. and U. Müller, "t-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business and Economic Statistics*, 2010, *28* (4), 453–468.

_ **and** _ , "Inference with Few Heterogeneous Clusters," *Review of Economics and Statistics*, 2016, *98* (1), 83–96.

Jackson, M., *Social and Economic Networks*, Princeton University Press, 2008.

_ **and B. Rogers**, "Meeting Strangers and Friends of Friends: How Random Are Social Networks?," *American Economic Review*, 2007, *97* (3), 890–915.

Jenish, N. and I. Prucha, "Central Limit Theorems and Uniform Laws of Large Numbers for Arrays of Random Fields," *Journal of Econometrics*, 2009, *150* (1), 86–98.

_ **and** _ , "On Spatial Processes and Asymptotic Inference Under Near-Epoch Dependence," *Journal of Econometrics*, 2012, *170* (1), 178–190.

Klaus, Andreas, Shan Yu, and Dietmar Plenz, "Statistical analyses support power law distributions found in neuronal avalanches," *PloS one*, 2011, *6* (5), e19779.

Lahiri, S., *Resampling Methods for Dependent Data*, Springer Science and Business Media, 2013.

Leung, M., "Treatment and Spillover Effects Under Network Interference," *working paper*, 2017.

_ , "A Weak Law for Moments of Pairwise-Stable Networks," *working paper*, 2017.

_ , "A Weak Law for Moments of Pairwise-Stable Networks," *working paper*, 2017.

_ **and R. Moon**, "Normal Approximation in Strategic Network Formation," *working paper*, 2017.

Liang, K. and S. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 1986, *73* (1), 13–22.

Newman, Mark EJ, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, 2005, *46* (5), 323–351.

Nicolau, J. and P. Rodrigues, "A New Regression-Based Tail Index Estimator," *working paper*, 2015.

Pesaran, H., "Global and Partial Non-Nested Hypotheses and Asymptotic Local Power," *Econometric Theory*, 1987, *3* (1), 69–97.

Politis, D., J. Romano, and M. Wolf, *Subsampling* Springer Series in Statistics, New York, NY: Springer New York, 1999.

Resnick, S. and G. Samorodnitsky, "Asymptotic Normality of Degree Counts in a Preferential Attachment Model," *Advances in Applied Probability*, 2016, *48* (A), 283–299.

Romano, J., A. Shaikh, and M. Wolf, "A Practical Two-Step Method for Testing Moment Inequalities," *Econometrica*, 2014, *82* (5), 1979–2002.

_ and _ , "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, 2008, *138* (9), 2786–2807.

Sheng, S., "Identification and Estimation of Network Formation Games," *working paper*, 2016.

Song, K., "Ordering-Free Inference from Locally Dependent Data," *arXiv preprint arXiv:1604.00447*, 2016.

Vuong, Q., "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 1989, pp. 307–333.