Limit Theorems for Small Worlds *

Arkadiusz Szydłowski†

University of Kent

[PLEASE SEE `https://arekszydlowski.github.io` FOR THE LATEST VERSION]

July 28, 2024

Abstract

We obtain central limit theorems for data coming from a sparse, small world network, i.e. a network with limited maximal degree and a relatively small (but growing) diameter, properties encountered in many social and economic networks. We do not impose restrictions on the strength of dependence between connected nodes and only assume that non-connected nodes are statistically independent. The limit theorems hold conditionally on the network evolution and the sufficient conditions vary between different dynamic network structures, from requiring that the largest network component grows only marginally slower than the number of nodes, N, to restricting it to grow slower than $\sqrt{N}$. These conditions translate into restrictions on the constant of growth of the network diameter relative to the maximal degree using the bound on the number of connected nodes from algebraic graph theory. We consider both means of node- and edge-specific characteristics and show that for the latter imposing weak dependence conditions may be necessary.

JEL: C10, C45

Keywords: Networks, CLT, m-dependence, Nonstationary data, Wild cluster bootstrap

*I would like to thank Bristol Econometric Study Group participants, Vadim Marmer, Nizar Allouch and Alfred Duncan for useful comments.

†School of Economics, University of Kent, Canterbury, CT2 7PE, UK. *E-mail address*: a.szydlowski@kent.ac.uk

1 Introduction

Many economic and social networks on top of being sparse exhibit a small-world property (Watts & Strogatz (1998)), namely that the network distance between each pair of connected nodes is small compared to the number of nodes in the network.[1] Formally, a small-world network has a diameter proportional to $\log N$, where N is the number of nodes. We combine this restriction and boundedness of the maximal degree of a node (i.e. sparsity) and ask a question if one can obtain a central limit theorem (CLT) for a network data where observations from connected nodes (through any path) can be arbitrarily dependent and only unconnected nodes are statistically independent.

The answer to our main question is affirmative under additional assumptions restricting the size of the diameter relative to the maximal degree for any given N, or, put differently, restricting the constant of proportionality relating the diameter to $\log N$. We proceed with suggesting estimators for the variance of the sample mean and investigate their performance in Monte Carlo simulations. Although a simple block-variance estimator leads to considerable undercoverage of the resulting confidence interval, the performance of a wild cluster bootstrap procedure is very promising and we recommend using it for practical application of our results to hypothesis testing and building confidence intervals.

We proceed with our analysis conditional on network evolution, thus we do not include uncertainty coming from network formation. Hence, our results apply to stable networks with network-mediated dependence as the main source of dependence. An example is a long-term friendship network where we are interested in labour market outcomes. These outcomes are likely to have been affected by network interactions (e.g. referrals) and our results suggest how to proceed with inference on means of such outcomes.

The conditions needed to obtain a CLT vary between different network structures, in particular on the number of growing components and the variation of their sizes. On one end, when the network consist of many components growing at the same rate it is enough that the largest component grows at a rate only marginally slower than N, on the other, when we have both large components growing in size and many fixed size components, the largest component may need to grow at a rate smaller

[1] The "small-world" property often also includes the characteristic that the network graph is much more clustered than a random graph (see e.g. Definition 4.1.3 in Watts (1999)). However, the latter property is not useful for the purpose of providing CLT in our context so we do not discuss it.

than $\sqrt{N}$.

In addition to node-specific means we also consider CLTs for edge-specific characteristics, where we distinguish between flows, i.e. purely characteristics of edges, and contrasts, i.e. functions of characteristics of nodes involved in an edge. We show that a CLT for the means of flows holds under relatively mild and natural strengthening of conditions needed for node-specific means. However, we require strong mixing conditions (with respect to network distance) to justify a CLT for the means of contrasts, which is considerably stronger than other conditions we impose.

Kojevnikov et al. (2021) provide a CLT for node-specific means assuming weak dependence in the form of ψ-dependence (see also Leung & Moon (2023)). They do not condition on the observed network in their analysis and provide some primitive conditions and examples of network formation processes consistent with ψ-dependence. They propose a HAC-type variance estimator. Leung (2023) shows, however, that for many networks a cluster-robust inference may perform better. We find that a wild cluster bootstrap works quite well in our setup. Similarly to our paper Ogburn et al. (2024) provide a CLT conditional on the network formation process but only allow dependence up to friends-of-friends, whereas we allow for any connected nodes (via any path) to be dependent, thus generalising their results.

There is a large literature on obtaining limit theorems with spatial networks (see Jenish & Prucha (2012), Kuersteiner & Prucha (2013), Kuersteiner (2019) among others). Although many social and economic phenomena could be modelled using these networks, most social and economic networks have relatively high clustering coefficients, which means common presence of cliques (see Jackson (2008)). But Kojevnikov et al. (2021) demonstrate that spatial networks have limitations in terms of accommodating nontrivial presence of cliques, thus restricting their usefulness in modelling observed networks.

Our results can be seen as extending results on limit theorems with m-dependence, where m can diverge to infinity, (Romano & Wolf (2000)) to the case where m is heterogenous and diverges at different rates for different groups (at the same time restricting dependence groups to be non-overlapping). From this perspective, our work is also related to the literature on normal approximations under local dependence (Baldi & Rinott (1989), Chen & Shao (2004)) with a difference that we allow the dependence neighbourhoods to grow with the sample size. Finally, our results are closely related to results in the clustering literature (see e.g. MacKinnon et al. (2023)

for a review) as the components in a network can be seen as different dependence clusters. In particular, Djogbenou et al. (2019) also use Lyapnuov condition to establish a CLT for clustered data and their Assumption 3 restricts the rate of growth of the largest cluster much like our conditions. However, unlike this article they do not consider specific cluster evolution structures nor they link the conditions to small-world properties of the data.

Other recent articles on inference using network data include Bickel et al. (2011), Bickel et al. (2013), Matsushita & Otsu (2023) among others.

2 Main idea

Let $\{Y_1, \ldots, Y_N\}$ denote mean zero random variables corresponding to nodes in a network G_N. We are interested in the central limit theorem for the sample mean:

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

assuming that only nodes not connected through any network path have statistically independent Y_i's and dependence between the remaining observations is not restricted.

Let the network consist of c_N separated components of size $N_c, c = 1, \ldots, c_N$ and the number of non-zero correlations among the nodes in a component is of order $N_c^{1+\gamma_c}, \gamma_c \in [0, 1]$. Then, if $Var(Y_i) < \infty$, we have:

$$Var\left(\sum_{i=1}^{N} Y_i\right) \sim \sum_{c=1}^{c_N} N_c^{1+\gamma_c}$$

Using this structure we provide sufficient conditions to verify the Lyapunov condition, which involve bounds on the rates of growth of N_c's. Next the bound on the maximal number of nodes in a sparse network with a given maximal degree and diameter (Pineda-Villavicencio & Wood (2015)) is used to translate these conditions to the parameters of a small world network.

For simplicity assume that the network consists of equal sized components, then Lyapunov's

condition is satisfied (see below) if:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \sum_{c=1}^{c_N} \frac{N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} = c_N^{-\frac{\delta}{2}}$$

converges to zero for any $\delta > 0$. This requires that the number of components grows with N, in other words, $N_c/N \to 0$. Now the result from the algebraic graph theory bounds N_c by a quantity of order $d_{max}^{\Delta_N}$, where d_{max} denotes the maximal degree and Δ_N denotes the network's diameter.

Since our conditions imply N_c is of order smaller than N, the results do not apply to networks with a giant component involving almost all nodes or, alternatively, networks with a giant component growing at rate N. This happens, for example, in the Facebook network where 99.91% of individuals belong to the largest connected component (Ugander et al. (2011)). Still, our limit theorems would apply if the giant component is of order arbitrarily smaller than N or if one can reasonably divide the giant component into statistically independent sub-components satisfying this condition (e.g. the links between some Facebook groups do not generate any cross-traffic, thus the groups can be viewed as independent). Finally, we note that if one is willing to assume weak dependence between connected nodes the results of Kojevnikov et al. (2021) would apply to such networks.

Recently, for a related problem, Kojevnikov & Song (2023) showed that consistent estimation of the mean in clustered samples, without intra-cluster dependence restrictions, requires presence of at least two large clusters, which implies that the largest cluster has to be of order smaller than $O(N)$. Thus, our findings are in line with that result.

3 Central limit theorems for small-world networks

Recall that $d_{max} = \max_{i \in \mathcal{N}_N} d_i$ is the maximal degree in network G_N, where $\mathcal{N}_N = \{1, 2, \ldots, N\}$. Define $l(i,j)$ to be the network distance on the shortest path between i and j and set $l(i,j) = \infty$ if i and j are not connected by any network path. The diameter of network G_N is now formally defined as $\Delta_N = \max_{i,j \in \mathcal{N}_N : l(i,j) < \infty} l(i,j)$. All our results hold conditionally on network evolution $\{G_N\}_{N=1}^{\infty}$ and, thus, take network formation as given.

3.1 Node specific means

Consider the sample mean $\overline{Y}$ defined above, let $C_c \subset \mathcal{N}_N$ enumerate nodes in network component c and make the following assumptions:

Assumption 1. *(a)* $d_{max} \geq 2, d_{\max} = O(1)$.

(b) $\Delta_N \leq \log_a(bN)$ *for some constants* $a > 1, b > 0$.

(c) $l(i,j) = \infty$ *implies* $Y_i \perp Y_j$.

(d) There exist $\{\gamma_c\}_{c=1}^{c_N} : 0 \leq \gamma_c \leq 1, \underline{\sigma}^2 > 0, \bar{\sigma}^2 > 0, \delta > 0$ *such that:*

$$\underline{\sigma}^2 N_c^{1+\gamma_c} \leq Var\left(\sum_{i \in C_c} Y_i\right) \leq \bar{\sigma}^2 N_c^{1+\gamma_c},$$
$$E\left(\sum_{i \in C_c} Y_i\right)^{2+\delta} \leq K_N^{\frac{2+\delta}{2}} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}.$$

for all C_c *and* N, *where* $K_N = O(1)$.

Assumption 1(a) imposes sparsity of the network. The lower bound on the maximal degree has a technical nature and rules out the case of networks formed only of connected pairs of nodes. Note that sparsity is often imposed as a condition on average degree, namely $1/N \sum_{i=1}^{N} d_i = O(1)$, which is implied by our condition. However, for all practical purposes both formulations can be seen as equivalent. Assumption 1(b) imposes the small-world property, namely that the diameter of the network is (at most) proportional to $\log N$ and part (c) states that unconnected nodes are statistically independent.

The last condition has two parts. The first part assumes that the number of non-zero correlations in a component of size N_c is a power function of N_c. Though, certainly imposing some structure on the number of non-zero correlations, it accommodates a variety of cases, in particular, all correlations being non-zero ($\gamma_c = 1, \underline{\sigma}^2 < 1, \bar{\sigma}^2 > 1$), the fraction of non-zero correlations being constant ($\gamma_c = 1, \underline{\sigma}^2 < 1, \bar{\sigma}^2 < 1$) and decreasing toward zero with N_c ($\gamma_c < 1$). We note that the power form is chosen for convenience and clarity of the CLT conditions and one may impose other correlation structures and follow similar arguments as in our proofs to obtain alternative conditions.[2] Also,

[2] As $N_c^{1+\gamma_c}$ does not need to be an integer, the bounds $\underline{\sigma}^2, \bar{\sigma}^2$ also accommodate rounding.

one does not need to know γ_c's in practice and can assume $\gamma_c = 1$, for all c, for the purpose of the conditions given below. Similarly, Romano & Wolf (2000) impose conditions that imply $Var(\sum_{i=1}^{N} Y_i) \sim N^{1+\gamma}$ for some $-1 \leq \gamma < 1$. The second part of (d) assumes existence of moments for the component-wise averages. For example, with $\delta = 2$, this condition holds if Y_i's have finite fourth moments and the number of nonzero within-component pairwise correlations is proportional to $N_c^{1+\gamma_c}$.

Further conditions needed to obtain a central limit theorem depend on the network evolution scheme, in particular, how many growing network components there are and what their relative growth rates are. Before we proceed, we need to clarify some notation. As the network evolves both the existing components grow in size and new components arise. Thus, we will make the dependence of the size of a component on the number of nodes explicit, i.e. write $N_c(N)$, and keep in mind that c_N is a function of N. Without loss of generality, let the first component ($c = 1$) be a component for which the number of non-zero correlations, i.e. $N_1^{1+\gamma_1}$, grows at the fastest rate.

Theorem 1. *Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of mean zero random variables and define $B_N^2 = Var(\sqrt{N}\overline{Y})$. Under Assumption 1:*

$$\frac{\sqrt{N}\overline{Y}}{B_N} \to^D N(0,1)$$

as $N \to \infty$ (conditionally on network evolution) if either of the following holds:

(a) all components grow at the same rate, γ_c's are all equal and $\log_{d_{max}-1} a > 1$;

(b) all components grow (possibly at different rates) and we have for all c:

(i) $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_c(N)}{N_1(N)} \to 0$,

(ii) $\frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}}$ is weakly decreasing in N,

(iii) there exists $M < \infty$ such that $\frac{N_{c_{\tilde{N}}}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N}(N)^{\gamma_N}} < \left(\frac{N_1(\tilde{N})}{N_1(N)}\right)^{\gamma_1}$ for all $(N, \tilde{N})$ such that $c_{\tilde{N}} = c_N + 1$ and $N > M, \tilde{N} > M$,

(iv) $\log_{d_{max}-1} a > (1+\gamma_c)\frac{2+\delta}{\delta}$.

(c) c_k components grow with N and remaining $c_N - c_k$ components have fixed size, c_k is fixed and $\forall c$:

$$\log_{d_{max}-1} a > 1 + \gamma_c,$$

(d) c_k *components grow with* N *and remaining* $c_N - c_k$ *components have fixed size,* $\frac{c_N}{N} \to s > 0$, $c_k \to \infty$ *and one of the following conditions is satisfied:*

(i) $\log_{d_{max}-1} a > (1+\gamma_c)\frac{2+\delta}{\delta}, \forall c,$

(ii) $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_c(N)}{N_1(N)} \to 0$, $\frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}}$ *is weakly decreasing in* N, *there exists* $M < \infty$ *such that* $\frac{N_{c_{\tilde{N}}}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N}(N)^{\gamma_N}} < \left(\frac{N_1(\tilde{N})}{N_1(N)}\right)^{\gamma_1}$ *for all* $(N, \tilde{N})$ *such that* $c_{\tilde{N}} = c_N + 1$ *and* $N > M, \tilde{N} > M$, *and* $\log_{d_{max}-1} a > 1 + \gamma_c, \forall c,$

(iii) components $\{1, \ldots, c_k\}$ *grow at the same rate,* $\{\gamma_c\}_{c=1}^{c_k}$ *are all equal and* $\log_{d_{max}-1} a > 1 + \gamma_c \frac{2+\delta}{\delta}, \forall c,$

(e) c_k *components grow with* N *and remaining* $c_N - c_k$ *components have fixed size,* $\frac{c_N}{N} \to 0$ *and either all components* $\{1, \ldots, c_k\}$ *grow at the same rate (with* $\{\gamma_c\}_{c=1}^{c_k}$ *all equal) and* $c_N^{(2+\delta)/2}/c_k^{1+\delta} \to 0$, *or we have for all* c*:*

(i) $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_c(N)}{N_1(N)} \to 0,$

(ii) $\frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}}$ *is weakly decreasing in* N,

(iii) there exists $M < \infty$ *such that* $\frac{N_{c_{\tilde{N}}}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N}(N)^{\gamma_N}} < \left(\frac{N_1(\tilde{N})}{N_1(N)}\right)^{\gamma_1}$ *for all* $(N, \tilde{N})$ *such that* $c_{\tilde{N}} = c_N + 1$ *and* $N > M, \tilde{N} > M$,

(iv) $\log_{d_{max}-1} a > (1+\gamma_c)\frac{2+\delta}{\delta}$ *when* $\frac{c_k}{c_N} \to 1$ *or* $\log_{d_{max}-1} a > \frac{2+\delta}{\delta}$ *otherwise.*

A natural corollary, providing sufficient conditions for a CLT irrespective of network structure and values of γ_c, follows:

Corollary 1. *Let* $\{Y_i\}_{i=1}^{\infty}$ *be a sequence of mean zero random variables and define* $B_N^2 = Var(\sqrt{NY})$. *Under Assumption 1:*

$$\frac{\sqrt{NY}}{B_N} \to^D N(0,1)$$

as $N \to \infty$ *(conditionally on network evolution) if either of the following holds:*

(a) all c_k *growing components (where* $c_k \leq c_N$*) satisfy:*

(i) $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_c(N)}{N_1(N)} \to 0,$

(ii) $\frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}}$ *is weakly decreasing in* N,

(iii) there exists $M < \infty$ *such that* $\frac{N_{c_{\tilde{N}}}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N}(N)^{\gamma_N}} < \left(\frac{N_1(\tilde{N})}{N_1(N)}\right)^{\gamma_1}$ *for all* $(N, \tilde{N})$ *such that* $c_{\tilde{N}} = c_N + 1$ *and* $N > M, \tilde{N} > M$,

and $\log_{d_{max}-1} a > \frac{2(2+\delta)}{\delta}$,

(b) all c_k growing components (where $c_k \leq c_N$) grow at the same rate and either $\log_{d_{max}-1} a > 1 + \frac{2+\delta}{\delta}$. *or* $c_N^{(2+\delta)/2}/c_k^{1+\delta} \to 0$.

The proof of Theorem 1 is given in the Appendix and follows the argument outlined in the previous section. Note that the presence of the factor $\sqrt{N}$ in the statement of theorem does not imply that we obtain a square root rate of convergence as in general B_N will not be $O(1)$. The result can be restated as a result conditional on common shocks affecting all the nodes in the network just as in Kojevnikov et al. (2021), and would then apply to networks where there is some dependence between unconnected nodes and the dependence can be modelled through observables.

Remark 1. *Theorem 1 provides only sufficient conditions for the CLT theorem to hold. Still, it covers a wide range of network evolution setups, including networks with a number of dominant components and a number of small-sized components, a structure commonly encountered in social networks.*

Remark 2. *The first condition in (b) requires that the component(s) with the fastest growing number of non-zero between-nodes correlations is(are) also the fastest growing component(s). It is trivially satisfied if all γ_c's are equal. Additionally, note that together conditions (b)(i)-(ii) imply* $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Leftrightarrow \frac{N_c(N)}{N_1(N)} \to 0$.

Remark 3. *The conditions (b)(ii) and (b)(iii) require presence of a dominant component(s) in which the number of non-zero correlations grows at the fastest rate, as well as that the new components cannot grow at a faster rate that this dominant component. Note that this allows some components to grow at the same rate.*

Remark 4. *The conditions imposed on* $\log_{d_{max}-1} a$ *restrict the scaling factor for the network diameter relative to the maximal degree and require that, for given N, the diameter is not too large relative to the maximal degree (in other words,* $1/\log(a)$ *cannot be too large). Part (a) assumes that $a > d_{max} - 1$, which is also the case in part (e)(iv) if all moments of Y exist, i.e. $\delta = \infty$. Finally, note that the condition on the diameter in (d)(iii) is weaker than the one in (d)(i), but both become equivalent to conditions in (b)(iv) and (c) when $\delta \to \infty$.*

Remark 5. *To the best of our knowledge, there is no precise formula linking the parameters of a small world model like the Watts-Strogatz small world (SW) model (Watts & Strogatz (1998)) and*

the Barabási-Albert (BA) preferential attachment model (Barabási & Albert (1999)) to the constants of proportionality in the diameter so it is difficult to translate our conditions on the constant a *to the parameters of these models.*[3] *For a random graph* $G(N,p)$*, the diameter is bounded when* $Np \to \infty$ *and* $Np \to 0$ *so our conditions on the size of* a *will be satisfied as long as* N *is large enough. When* $Np \to \lambda > 1$ *we have* $\Delta_N = \log N / \log \lambda$*. In such random graph let us truncate the degree at some* d_{max} *and, on top of that, artificially split the largest component, which is of order* N *in this case, such that it grows at some slower rate. This way the generated network dynamics will fit our setup in part (b) and condition (b)(iv) becomes* $\log \lambda > \log(d_{max}-1)\frac{2+\delta}{\delta} \sup_{c\geq 1}(1+\gamma_c)$ *which gives* $\lambda > 9$ *when* $d_{max} = 4, \gamma_c = 1, \delta = \infty$*.*

Remark 6. *Let us compare our conditions to Assumption 3 in Djogbenou et al. (2019). Adapting from their regression setup to a simple sample mean setup, if* η_N *denotes the rate of divergence of* $Var(\sum_{i=1}^N Y_i)$*, they require (see MacKinnon et al. (2023)):* $\left(\frac{\sqrt{\eta_N}}{N}\right)^{-(2+\delta)/(1+\delta)} \frac{N_1}{N} \to 0$*. Let us compare this condition to the ones in parts (a) and (c). For the former case, it is easy to derive that both our and their conditions require* $c_N \to \infty$*, in other words* $N_1/N \to 0$*. For the latter case, one can see from the proof of Theorem 1 that we practically need* $N_1^2/N \to 0$ *(assuming* $\gamma_1 = 1$*), whereas their condition implies that we need:*[4]

$$\left(\frac{N^2}{\sum_{c=1}^{c_k} N_c^2 + (c_N - c_k)O(1)}\right)^{\frac{2+\delta}{2(1+\delta)}} \frac{N_1}{N} = \left(\frac{N^{-\frac{\delta}{2+\delta}} N_1^{\frac{2(1+\delta)}{2+\delta}}}{N^{-1}\sum_{c=1}^{c_k} N_c^2 + O(1)}\right)^{\frac{2+\delta}{2(1+\delta)}}$$

to converge to zero, which is satisfied if $N_1^{2(1+\delta)/\delta}/N \to 0$*. Note that this is a stronger condition (e.g. requires* $N_1^3/N \to 0$ *when* $\delta = 2$*) than the one we impose and both are equivalent only if all the moments exist (i.e.* $\delta = \infty$*).*

[3] Even if these models are precisely defined. See Bollobás & Riordan (2002) for a discussion of the mathematical definitions of small-world networks.

[4] Note that the second part of our Assumption 1(d) (with $\gamma_c = 1$) is implied by an equivalent of their Assumption 1 (see e.g. their Lemma A.2).

4 Variance estimation

In this section we suggest estimators of the variance B_N^2 that can be used for inference together with Theorem 1. Since Y_i's have zero mean:

$$B_N^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} E(Y_i Y_j)\mathbb{1}\{l(i,j) < \infty\}$$

thus a natural estimator arises:

$$\hat{B}_N^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} Y_i Y_j \mathbb{1}\{l(i,j) < \infty\}.$$

Note that this estimator can be viewed as a block-variance estimator where the blocks correspond to different unconnected components of the network and grow in size with N.

Theorem 2. *Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of mean zero random variables, Assumptions 1 (a)-(c) hold and $E|Y_i|^4$ is bounded for all i. Then:*

$$Var(\hat{B}_N - B_N) \to 0.$$

as $N \to \infty$ (conditionally on network evolution) if:

(a) all components grow at the same rate and $\log_{d_{max}-1} a > 3$,

(b) all components grow (possibly at different rates) and $\log_{d_{max}-1} a > 4$,

(c) c_k components grow with N and remaining $c_N - c_k$ components have fixed size, c_k is fixed and $\log_{d_{max}-1} a > 2$,

(d) c_k components grow with N and remaining $c_N - c_k$ components have fixed size, $c_k \to \infty$ and either $\log_{d_{max}-1} a > 4$ or all c_k components grow at the same rate and $\log_{d_{max}-1} a > 3$.

The theorem implies consistency of the proposed estimator. As the estimator only uses cross-products corresponding to observations in the same network component (alternatively, block) we coin it the block-variance estimator. Recall that, when all components grow at the same rate with N, Theorem 1 allows the size of the largest connected component to grow at a rate arbitrarily

close to N but here the allowed rate is not higher than $N^{1/3}$. Again, this is in line with findings in Kojevnikov & Song (2023) for clustered samples. They show that one requires much stricter conditions for variance estimation then for consistent discrimination of the mean.[5]

Although consistent, this estimator does not work well in practice if there is a lot of dependence between Y_i's. In our Monte Carlo simulations we show that a confidence interval using the block-variance estimator severely undercovers even for the sample size $N = 10000$ when there is strong dependence between observations belonging to the same network component. This is in line with simulations for a related HAC estimator in Kojevnikov et al. (2021) when the "autoregressive" parameter is close to 0.5.

Since our setup is similar to the problem of estimating variance with a few large and growing in size clusters (Cameron et al. (2008)), as an alternative to the block-variance estimator we consider the wild clustered bootstrap and find that it performs much better in our Monte Carlo simulations. Let $c = 1, \dots, c_N$ enumerate separate components of network G_N. The bootstrap procedure is as follows:

1. For each connected component draw $v_c = -1$ or 1 with probability 1/2.
2. Calculate $\overline{Y}^* = \frac{1}{N} \sum_{i=1}^{N} Y_i v_{c(i)}$ where $c(i)$ denotes the component that i belongs to.
3. Estimate B_N^2 by variance of $\sqrt{N}\overline{Y}^*$ across bootstrap samples.

As an alternative one may consider randomisation tests of Canay et al. (2017).

5 Means of edge-specific characteristics

In this section we provide limit theorems for means of characteristics of edges between nodes. Applications include means of input-output flows in production networks (see e.g. Acemoglu et al. (2012)) or mean difference in socio-economic status between individuals belonging to the same local community (see e.g. Chetty et al. (2022)). Note that the edge characteristics in these two examples have a different structure – in the former they are nonparametric functions of a node pair (i, j) ("flows") whereas in the latter they are known functions of characteristics of a node (i, j) involved in

[5]Note that the Hansen and Lee condition (Hansen & Lee (2019)) that they require for consistent estimation of variance is satisfied with clusters of size $N^{1/3}$.

an edge ("contrasts"). These differences lead to distinct analysis, in particular a CLT for contrasts requires stronger conditions.

5.1 Flows

Let Y_{ij} denote the characteristic of an edge between nodes i and j and assume that there are no flows between separate components, i.e. $Y_{ij} = 0$ if nodes i and j are not connected (by any path). With this structure we effectively have c_N components with $N_{c,f} = N_c(N_c - 1)$ outcome pairs Y_{ij} and the analysis resembles the one for node-specific means, but now the effective sample size is $N_f = \sum_{c=1}^{c_N} N_{c,f}$. We can define the edge-specific mean by:

$$\overline{Y}_f = \frac{1}{N_f} \sum_{i=1}^{N} \sum_{j \neq i} Y_{ij} \mathbb{1}\{l(i,j) < \infty\}.$$

Similarly to node-specific means we will assume that flows in separate network components are statistically independent and modify Assumption 1(d) to the present context:

Assumption 1(c)'. *$l(i,k) = \infty$ implies $Y_{ij} \perp Y_{kl}$.*

Assumption 1(d)'. *There exist $\{\gamma_c\}_{c=1}^{c_N} : 0 \leq \gamma_c \leq 1, \underline{\sigma}^2 > 0, \bar{\sigma}^2 > 0, \delta > 0$ such that:*

$$\underline{\sigma}^2 N_{c,f}^{1+\gamma_c} \leq Var\left(\sum_{i \in C_c} \sum_{j \in C_c : j \neq i} Y_{ij} \right) \leq \bar{\sigma}^2 N_{c,f}^{1+\gamma_c},$$
$$E\left(\sum_{i \in C_c} \sum_{j \in C_c : j \neq i} Y_{ij} \right)^{2+\delta} \leq K_N^{\frac{2+\delta}{2}} N_{c,f}^{(1+\gamma_c)\frac{2+\delta}{2}}.$$

for all C_c and N, where $K_N = O(1)$.

We have the following result:

Theorem 3. *Let $\{Y_{ij}\}_{i,j=1}^{\infty}$ be a sequence of mean zero random variables and define $B_{N,f}^2 = Var(\sqrt{N_f}\overline{Y}_f)$. Under Assumptions 1 (a), (b), 1(c)' and 1(d)' :*

$$\frac{\sqrt{N_f}\overline{Y}_f}{B_{N,f}} \to^D N(0,1)$$

as $N \to \infty$ (conditionally on network evolution) if

(a) condition (a) in Theorem 1 holds,

(b) conditions (b)(i)-(iii) in Theorem 1 hold and $\log_{d_{max}-1} a > 2(1+\gamma_c)\frac{2+\delta}{\delta}, \forall c$,

(c) condition (c) in Theorem 1 holds with $\log_{d_{max}-1} a > 2(1+\gamma_c), \forall c$,

(d) condition (d) in Theorem 1 holds with (i)-(iii) replaced by:

(i)' $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_c(N)}{N_1(N)} \to 0$, $\frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}}$ *is weakly decreasing in* N*, there exists* $M < \infty$ *such that* $\frac{N_{c_{\tilde{N}}}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N}(N)^{\gamma_N}} < \left(\frac{N_1(\tilde{N})}{N_1(N)}\right)^{\gamma_1}$ *for all* $(N, \tilde{N})$ *such that* $c_{\tilde{N}} = c_N + 1$ *and* $N > M, \tilde{N} > M$*, and* $\log_{d_{max}-1} a > 2(1+\gamma_c), \forall c$,

(ii)' *components* $\{1, \ldots, c_k\}$ *grow at the same rate,* $\{\gamma_c\}_{c=1}^{c_k}$ *are all equal and* $\log_{d_{max}-1} a > 2(\frac{1+\delta}{\delta} + \gamma_c \frac{2+\delta}{\delta}), \forall c$,

(e) condition (e) in Theorem 1 holds with part (iv) replaced by: $\log_{d_{max}-1} a > 2(1+\gamma_c)\frac{2+\delta}{\delta}$ *when* $\frac{c_k}{c_N} \to 1$ *or* $\log_{d_{max}-1} a > \frac{2(2+\delta)}{\delta}, \forall c$.

Theorem 3 can be used for inference once an estimator of $B_{N,f}$ is available. One would expect that an analogous estimator to the block-variance estimator or a wild cluster bootstrap described in Section 4 would work by the same reasoning as for node-specific means. Note that Theorem 3 strengthens conditions of Theorem 1 due to the fact that in the current setup each network "component" contains (up to) $N_c(N_c - 1)$ correlated elements compared to N_c before.

5.2 Contrasts

Let h be a symmetric function and define the edge-specific mean as:[6]

$$\overline{Y}_c = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} h(Y_i, Y_j) \mathbb{1}\{l(i,j) < \infty\}.$$

A leading example would be $h(Y_i, Y_j) = |Y_j - Y_i|$ with Y_i denoting a measure of socio-economic status like income (Chetty et al. (2022)), in which case the statistic would measure average differences in income among neighbourhoods ("economic connectedness") and our results would provide a starting

[6]The definition could be extended to functions of characteristics of triples, quadruples etc. of nodes, which can be used to study clique characteristics. The treatment of such statistics would follow similar lines. Hence, for the sake of exposition, we do not analyse them in detail.

point for conducting inference which takes into account network-dependence between connected units. We point out that, when there is non-negligible dependence between connected individuals, even large sample sizes may not guarantee statistical significance of the findings as in such case the "effective" sample size may be small.

Under stationarity of Y_i we get the following Hoeffding decomposition:

$$\frac{\sqrt{N}\overline{Y}_c}{B_{N,c}} = B_{N,c}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} h_1(Y_i)\frac{N_c(i)-1}{N-1} + B_{N,c}^{-1}\frac{\sqrt{N}}{N(N-1)}\sum_{i<j} h_2(Y_i,Y_j)\mathbb{1}\{l(i,j)<\infty\}, \quad (1)$$

where $h_1(y) = E_Y[h(y,Y)], h_2(y_1,y_2) = h(y_1,y_2) - h_1(y_1) - h_1(y_2)$, $N_c(i)$ denotes the number of nodes in the component to which i belongs and $B_{N,c}^2 = Var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} h_1(Y_i)\frac{N_c(i)-1}{N-1}\right)$. In order to obtain a CLT we need to show that the second term in the decomposition converges to zero in probability and show that a CLT holds for the triangular array $\left\{h_1(Y_i)\frac{N_c(i)-1}{N-1}\right\}_{i,N}$.

The variance of the second term in (1) (up to a scaling factor) can be written as:

$$\sum_{c=1}^{c_N}\sum_{i\in C_c}\sum_{j\in C_c, j\neq i}\sum_{k\in C_c}\sum_{l\in C_c, l\neq j} E[h_2(Y_i,Y_j)h_2(Y_k,Y_l)]$$

and under Assumption 1(d) this term is of order $\sum_{c=1}^{c_N} N_c^{3+\gamma_c}$, which is the same as the order of $B_{N,c}^2$. This shows a difficulty in obtaining a central limit theorem for contrasts without some further restrictions on dependence between Y_i's. Thus, we impose a strong mixing condition with respect to the network distance $l(\cdot,\cdot)$ following e.g. Kojevnikov et al. (2021).

For σ-fields $\mathcal{F},\mathcal{G}$, let $\alpha(\mathcal{F},\mathcal{G}) = \sup_{F\in\mathcal{F},G\in\mathcal{G}} |P(F\cap G) - P(F)P(G)|$ and define the component-specific mixing coefficients by:

$$\alpha_{c,N}(s) = \sup\{\alpha(\sigma(Y_A),\sigma(Y_B)) : A,B\subset C_c, l(A,B)\geq s\}$$

where $Y_A = \{Y_i\}_{i\in A}$ and $l(A,B) = \min_{i\in A}\min_{i'\in B} l(i,i')$. Further, note that the data $\{Y_i\}_{i=1}^N$ is α-mixing with $\alpha_N(s) = \max_{c\in\{1,\ldots,c_N\}}\alpha_c(s)$. Let $c_{N_c}(s,m;k)$ be the quantity capturing the network's denseness defined on p. 891 in Kojevnikov et al. (2021). We impose the following assumption.

Assumption 2. *For all $c\in\mathbb{N}$, $\{Y_i\}_{i\in C_c}$ is a stationary strong mixing process with $E[h(Y_i,Y_j)] = 0$ and we have for $p>4$:*

(a) h is a bounded Lipchitz function satisfying: $E|h(Y_i, Y_j)|^p < \infty$.

(b) $\frac{1}{N^3 B_{N,c}^2} \sum_{c=1}^{c_N} N_c \sum_{s \geq 0} c_{N_c}(s, N_c; 2)\alpha_c(s) \to 0$.

(c) There exists a positive sequence m_N such that for $k = 1, 2$:

$$\frac{1}{N^{\frac{k}{2}} B_{N,c}^{2+k}} \sum_{s \geq 0} c_N(s, m_N; k)\alpha_N(s)^{1-\frac{2+k}{p}} \to 0,$$

$$\frac{N^{3/2}\alpha_N(m_N)^{1-\frac{1}{p}}}{B_{N,c}} \to 0.$$

Remark 7. *Assumptions (a) and (c) are needed for the asymptotic normality of the first term in (1). Assumptions (a) and (b) are used to show that the second term vanishes. Note that if $p \to \infty$ and $c_{N_c}(s, N_c, 2) \leq c_N(s, m_N, 2)$, then (c) implies (b), however, in general, $c_N(s, m, 2)$ is not monotone in the second argument so this does not follow.*

Remark 8. *Lipschitz continuity in part (a), though relatively strong, is satisfied trivially for our leading example of $h(y_1, y_2) = |y_2 - y_1|$.*

Remark 9. *Using the techniques in the proof of Theorem 1 it is easy to derive that the following set of conditions is sufficient for part (b):*

$$\log_{d_{max}-1} a > 1/2,$$

$$\sum_{c=1}^{c_N} \sum_{s \geq 0} c_{N_c}(s, N_c; 2)\alpha_c(s) = O(B_{N,c}^2).$$

We are now ready to state the CLT theorem for contrasts.

Theorem 4. *Under Assumptions 1(c) and 2 we have:*

$$\frac{\sqrt{N}\overline{Y}_c}{B_{N,c}} \to^D N(0, 1)$$

as $N \to \infty$ (conditionally on network evolution).

Compared to the previous results, this theorem imposes significantly stricter assumptions since restrictions on the structure of the network and the largest component do not suffice here, as

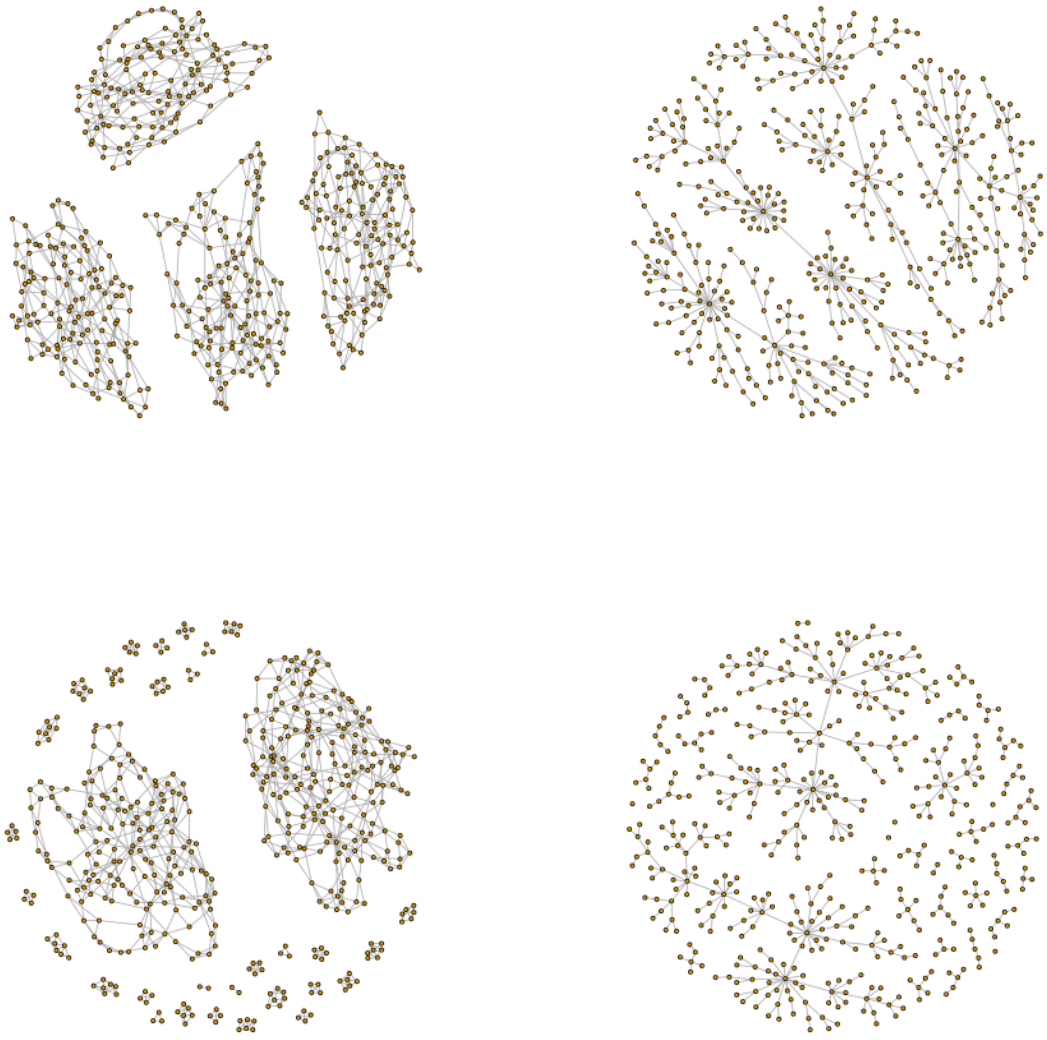
mentioned above. The small-world property and sparsity conditions will affect the network denseness function, $c_N(s, m; k)$. However, establishing a precise link between the diameter, maximal degree and this function is a question for further research.

6 Monte Carlo simulations

We consider two network generating algorithms: the Watts-Strogatz small world (SW) model (Watts & Strogatz (1998)) and the Barabási-Albert (BA) preferential attachment model (Barabási & Albert (1999)). The first model generates networks with diameters proportional to $\log N$ whereas the second model produces diameters proportional to $\log N$ or $\log N / \log \log N$ depending on the parameters (Bollobás & Riordan (2004)). For most parameter values the BA model implies that the maximal degree of a node grows with N, thus we further "prune" the graph to make sure that the maximal degree is stable: (1) we start with a node with the highest degree and randomly erase superfluous edges, (2) check if the maximal degree satisfies the imposed bound, (3) if not, we go back to step (1) and repeat the procedure.

In terms of the architecture of the network, we consider both (approx.) equal-sized components and growing + fixed components. For the former case we start with four connected components for $N = 500$ and add one component for each increase in the sample size above that, hence ending up with seven components for $N = 10000$. For the latter case, we allocate 30% of all nodes to the fixed components and we draw fixed component sizes from the binomial distribution with mean size of 5 nodes and maximal size of 10. We start with two growing components when $N = 500$ and add one more for each increase in the sample size, such that these components grow (approx.) at rates $N^{\{0.45,0.25,0.15,0.1,0.1\}}$, respectively. We perform 1000 MC repetitions and use 1000 replications for the bootstrap procedures. Figure 1 shows four examples of networks generated by SW and BA models (top panel: equal components, bottom panel: growing + fixed).

Figure 1: Examples of Monte Carlo designs, SW model (left) and BA model (right), $N = 500$.



6.1 Node-specific means

Let $C(i)$ denote the network component containing node i and $N_c(i)$, as before, denote the number of connected nodes in this component. The data is generated from the following process

$$Y_i = \frac{1}{\sqrt{N_c(i)-1}} \sum_{j \neq i, j \in C_{c(i)}} \varepsilon_j$$

where ε_j's are i.i.d., drawn from a standardised uniform distribution. In other words, node i's outcome is equal to the average of ε's of all the nodes that i is connected to, which implies strong dependence between outcomes belonging to the same network component. We consider coverage of confidence intervals built using known variance ("oracle") , block-variance estimator $\hat{B}_N$ ("estim.") and wild cluster bootstrap ("boot.") introduced in Section 4.

Table 1 contains the simulation result for means of node-specific characteristics. The BA model is parametrised by: m – the number of edges added in each step of building the graph, z_a - appeal of nodes that do not have any connections. The algorithm for building an SW network starts with a circle (or, more generally, lattice) graph and "rewires" some of the connections between neighbouring nodes to some more distant nodes, thus is parametrised by: p - probability of rewiring an edge, k - the number of edges per vertex in the initial circle graph. Different values of these parameters produce graphs with different maximal degrees and diameters.

When we use the known variance the coverage is close to the nominal 95% level across the designs and parameter values, thus confirming that the CLT holds for small world networks. However, once we use the estimated variance $\hat{B}_N^2$ the coverage deteriorates substantially, with values somehow close to the nominal values only in the three top left panels of Table 1 for which the networks are pretty sparse with a small degree and a large diameter. This shows the difficulty of precisely estimating the variance with strong dependence between observations in a network setting, a phenomenon also occurring in Kojevnikov et al. (2021) (see their simulation results with large values of the "autoregressive" parameter γ).[7]

The wild cluster bootstrap works reasonably well for networks with equal components, besides the small sample size $N = 500$, with coverage values only slightly above the nominal 95% across all

[7]Kojevnikov et al. (2021) use a HAC estimator with kernel weighting but the idea behind our $\hat{B}_N$ and their estimator is similar.

Table 1: Simulated coverage, node-specific means, 95% level

	BA model				Coverage			SW model				Coverage		
N	m	z_a	d_{max}	Δ_N	oracle	estim.	boot.	p	k	d_{max}	Δ_N	oracle	estim.	boot.
	Equal Components													
500	1	0	10	12	0.954	0.844	0.988	0.05	2	6	12	0.958	0.814	0.998
1000	1	0	10	11	0.964	0.930	0.954	0.05	2	6	15	0.950	0.854	0.975
5000	1	0	10	14	0.938	0.941	0.954	0.05	2	7	20	0.949	0.860	0.978
10000	1	0	10	16	0.952	0.950	0.942	0.05	2	7	20	0.960	0.897	0.972
500	1	1	10	16	0.954	0.869	0.983	0.05	5	13	5	0.949	0.836	0.994
1000	1	1	10	16	0.947	0.870	0.968	0.05	5	13	6	0.932	0.862	0.987
5000	1	1	10	18	0.957	0.928	0.955	0.05	5	14	7	0.954	0.876	0.965
10000	1	1	10	20	0.956	0.937	0.956	0.05	5	15	8	0.954	0.893	0.968
500	1	2	10	15	0.942	0.850	0.979	0.05	10	24	3	0.943	0.811	0.998
1000	1	2	10	17	0.955	0.869	0.951	0.05	10	24	4	0.946	0.828	0.983
5000	1	2	10	25	0.950	0.923	0.969	0.05	10	26	5	0.945	0.858	0.971
10000	1	2	10	21	0.941	0.920	0.968	0.05	10	26	5	0.941	0.904	0.957
500	2	0	20	6	0.959	0.818	0.994	0.10	2	8	9	0.947	0.815	0.995
1000	2	0	20	7	0.948	0.847	0.977	0.10	2	7	10	0.954	0.851	0.986
5000	2	0	20	10	0.956	0.876	0.967	0.10	2	8	14	0.951	0.862	0.968
10000	2	0	20	10	0.951	0.860	0.958	0.10	2	9	15	0.957	0.886	0.969
500	2	1	20	7	0.953	0.801	0.999	0.10	5	14	4	0.951	0.803	0.997
1000	2	1	20	7	0.950	0.831	0.980	0.10	5	14	5	0.943	0.836	0.970
5000	2	1	20	10	0.959	0.879	0.971	0.10	5	16	6	0.954	0.857	0.962
10000	2	1	20	10	0.948	0.875	0.970	0.10	5	16	7	0.961	0.883	0.963
500	2	2	20	6	0.951	0.809	0.995	0.10	10	25	3	0.953	0.800	0.995
1000	2	2	20	7	0.952	0.854	0.983	0.10	10	25	4	0.942	0.855	0.975
5000	2	2	20	9	0.955	0.884	0.968	0.10	10	28	4	0.946	0.874	0.975
10000	2	2	20	10	0.955	0.888	0.974	0.10	10	30	5	0.951	0.873	0.963
	Growing + Fixed													
500	1	0	10	10	0.942	0.895	0.964	0.05	2	6	13	0.941	0.673	0.999
1000	1	0	10	13	0.945	0.902	0.963	0.05	2	7	15	0.950	0.782	0.996
5000	1	0	10	12	0.955	0.938	0.959	0.05	2	8	19	0.949	0.768	0.996
10000	1	0	10	13	0.958	0.953	0.956	0.05	2	7	21	0.949	0.774	0.994
500	1	1	10	11	0.952	0.908	0.968	0.05	5	13	6	0.952	0.670	0.997
1000	1	1	10	14	0.953	0.917	0.955	0.05	5	14	6	0.947	0.768	0.997
5000	1	1	10	16	0.945	0.938	0.959	0.05	5	14	8	0.949	0.789	0.998
10000	1	1	10	17	0.953	0.932	0.952	0.05	5	15	9	0.951	0.796	0.992
500	1	2	10	16	0.956	0.712	0.999	0.05	10	24	4	0.940	0.650	0.999
1000	1	2	10	17	0.949	0.818	0.990	0.05	10	24	4	0.948	0.772	0.998
5000	1	2	10	23	0.943	0.918	0.965	0.05	10	27	5	0.941	0.786	0.997
10000	1	2	10	22	0.941	0.934	0.965	0.05	10	26	6	0.964	0.796	0.991
500	2	0	20	7	0.953	0.661	0.999	0.10	2	7	10	0.953	0.658	0.996
1000	2	0	20	8	0.956	0.804	1.000	0.10	2	7	10	0.949	0.792	0.998
5000	2	0	20	9	0.954	0.788	0.994	0.10	2	8	14	0.948	0.801	0.998
10000	2	0	20	10	0.950	0.790	0.994	0.10	2	7	17	0.956	0.816	0.987
500	2	1	20	7	0.958	0.670	0.999	0.10	5	15	5	0.945	0.668	0.997
1000	2	1	20	7	0.936	0.762	0.993	0.10	5	15	5	0.951	0.761	0.998
5000	2	1	20	10	0.953	0.785	0.997	0.10	5	15	7	0.959	0.798	0.998
10000	2	1	20	10	0.951	0.799	0.988	0.10	5	16	7	0.963	0.820	0.994
500	2	2	20	6	0.947	0.658	0.997	0.10	10	26	3	0.956	0.677	0.998
1000	2	2	20	7	0.961	0.791	0.997	0.10	10	26	4	0.954	0.761	0.995
5000	2	2	20	9	0.956	0.786	0.997	0.10	10	28	5	0.941	0.763	0.999
10000	2	2	20	9	0.963	0.810	0.991	0.10	10	29	4	0.960	0.957	0.957

Note: 1000 Monte Carlo simulations, 1000 bootstrap replications. "Oracle" – known variance, "estim." – variance estimator $\hat{B}_N$, "boot." – wild cluster bootstrap.

designs. However, we also see that it provides very conservative inference when we have unequal-sized components, especially for the SW model. Still it provides valid inference, thus we recommend using the wild cluster bootstrap for data coming from sparse small world networks.[8]

6.2 Edge-specific means

As the case of means of flows discussed in Section 5.1 is very similar to the case of node-specific means, we only run simulations for the means of contrasts. As Theorem 4 requires weak dependence within components, we follow the design in Kojevnikov et al. (2021) and generate outcomes as:

$$Y_i = \sum_{s \geq 0} \frac{\rho^s}{|L_i(s)|} \sum_{j \in L_i(s)} \varepsilon_j$$

where $L_i(s)$ denotes the set of nodes at distance s from i, we set $\rho = 0.5$ and again ε follows the standardised uniform distribution.[9] Further, following our main example, we take:

$$h(Y_i, Y_j) = |Y_j - Y_i|.$$

As variance estimation in the U-statistic setup is more involved than with simple means and we do not provide variance estimators above, we only provide coverage values with known variance.

Table 2 shows that for all specifications of the network formation model the coverage probabilities are close to 95% even for $N = 500$, which is in line with our CLT in Theorem 4.

7 Conclusion

Many social and economic networks are sparse and are small-world. We show that data coming from such networks satisfies a central limit theorem under the additional assumption restricting the constant of proportionality of the diameter to $\log N$, even without imposing weak dependence between connected nodes.

Our result can be seen as a "possibility" theorem showing that a CLT applies quite generally

[8] The overcoverage noted here is interesting in comparison with the simulation results in Djogbenou et al. (2019), which show in a regression setup that the wild cluster bootstrap tests tend to undercover in most of the cases of clustering.

[9] We have also ran simulations with normal errors and the results are very similar. See Appendix G.

Table 2: Simulated coverage, edge-specific means (contrasts), 95% level

	BA model					SW model				
N	m	z_a	d_{max}	Δ_N	Coverage oracle	p	k	d_{max}	Δ_N	Coverage oracle
	Equal Components									
500	1	0	10	10	0.946	0.05	2	6	12	0.952
1000	1	0	10	11	0.939	0.05	2	6	15	0.939
5000	1	0	10	13	0.951	0.05	2	7	20	0.958
10000	1	0	10	14	0.933	0.05	2	7	20	0.950
500	1	1	10	16	0.953	0.05	5	13	5	0.956
1000	1	1	10	16	0.944	0.05	5	13	6	0.935
5000	1	1	10	23	0.945	0.05	5	14	7	0.934
10000	1	1	10	19	0.937	0.05	5	15	8	0.937
500	1	2	10	14	0.951	0.05	10	24	3	0.957
1000	1	2	10	15	0.937	0.05	10	24	4	0.950
5000	1	2	10	18	0.934	0.05	10	26	5	0.948
10000	1	2	10	21	0.95	0.05	10	26	5	0.937
500	2	0	20	7	0.946	0.1	2	8	9	0.945
1000	2	0	20	8	0.941	0.1	2	7	10	0.944
5000	2	0	20	11	0.953	0.1	2	8	14	0.945
10000	2	0	20	11	0.947	0.1	2	9	15	0.944
500	2	1	20	7	0.962	0.1	5	14	4	0.961
1000	2	1	20	7	0.944	0.1	5	14	5	0.930
5000	2	1	20	9	0.929	0.1	5	16	6	0.952
10000	2	1	20	11	0.957	0.1	5	16	7	0.948
500	2	2	20	6	0.954	0.1	10	25	3	0.954
1000	2	2	20	7	0.950	0.1	10	25	4	0.954
5000	2	2	20	9	0.948	0.1	10	28	4	0.940
10000	2	2	20	9	0.956	0.1	10	30	5	0.948
	Growing + Fixed									
500	1	0	10	10	0.958	0.05	2	6	12	0.961
1000	1	0	10	13	0.957	0.05	2	6	15	0.952
5000	1	0	10	12	0.950	0.05	2	7	20	0.957
10000	1	0	10	13	0.951	0.05	2	7	20	0.957
500	1	1	10	11	0.947	0.05	5	13	5	0.952
1000	1	1	10	14	0.965	0.05	5	13	6	0.950
5000	1	1	10	16	0.954	0.05	5	14	7	0.957
10000	1	1	10	17	0.946	0.05	5	15	8	0.943
500	1	2	10	16	0.947	0.05	10	24	3	0.952
1000	1	2	10	17	0.946	0.05	10	24	4	0.948
5000	1	2	10	23	0.942	0.05	10	26	5	0.944
10000	1	2	10	22	0.936	0.05	10	26	5	0.940
500	2	0	20	7	0.943	0.1	2	8	9	0.947
1000	2	0	20	8	0.942	0.1	2	7	10	0.934
5000	2	0	20	9	0.953	0.1	2	8	14	0.954
10000	2	0	20	10	0.944	0.1	2	9	15	0.954
500	2	1	20	7	0.951	0.1	5	14	4	0.959
1000	2	1	20	7	0.946	0.1	5	14	5	0.947
5000	2	1	20	10	0.962	0.1	5	16	6	0.958
10000	2	1	20	10	0.952	0.1	5	16	7	0.955
500	2	2	20	6	0.947	0.1	10	25	3	0.943
1000	2	2	20	7	0.965	0.1	10	25	4	0.959
5000	2	2	20	9	0.948	0.1	10	28	4	0.963
10000	2	2	20	9	0.941	0.1	10	30	5	0.939

Note: 1000 Monte Carlo simulations, 1000 bootstrap replications. "Oracle" – known variance.

to network-dependent data with the largest component of size N^α where $\alpha < 1$, so imposing weak dependence conditions within connected components seem not strictly necessary here. We also provide some evidence that the wild cluster bootstrap can be used successfully (though, anti-conservatively) to estimate the variance in this strongly dependent network setting.

We consider a simple setup of undirected unweighted networks but the results should extend naturally to directed networks and networks in which we can assign (bounded) weights to covariances of characteristics between two connected nodes. If these weights would vanish or decrease sufficiently fast between large connected components, then one could potentially be able to extend our results to networks with a giant component of size $O(N)$, i.e. larger than allowed in our current setup.

Appendix

A Proofs

A.1 Useful lemmas

Lemma 1. *(Pineda-Villavicencio, Wood (2015)) Every graph with minimum degree d_{min}, maximum degree d_{max} and diameter Δ_N has at most $2d_{min}(d_{max}-1)^{\Delta_N-1}+1$ vertices.*

Lemma 2. *(Kojevnikov et al. (2021), Prop. 2.2) Let f and g belong to a collection of bounded Lipschitz real functions and sets (A,B) of nodes be such that $l(A,B)\geq s$. If $\{Y_{i,N}\}_{i=1}^{\infty}$ is a strong mixing triangular array (w.r.t. network distance $l(\cdot,\cdot)$) with mixing coefficients $\{\alpha_N(s): s\geq 0\}$, then we have:*

$$|cov(Y_A,Y_B)|\leq 4\|f\|_\infty\|g\|_\infty\alpha_N(s).$$

B General CLT for networked data

The following theorem gives high level conditions for the networked data to satisfy CLT. Our main theorem, Theorem 1, will follow from verifying these conditions for different network evolution structures.

Theorem 5. *Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of mean zero random variables and define $B_N^2=Var(\sqrt{N}\overline{Y})$. Let Assumption 1 hold and:*

(a) $B_N^2\geq L_N\frac{1}{c_N}\sum_{c=1}^{c_N}N_c^{\gamma_c}$,

(b) $\frac{K_N}{L_N}=O(1)$,

(c) $\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}}\frac{\sum_{c=1}^{c_N}N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N}N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}}\to 0.$

Then we have:

$$\frac{\sqrt{N}\overline{Y}}{B_N}\to^D N(0,1)$$

as $N\to\infty$ (conditionally on network evolution).

Proof. Firstly, note that under Assumption 1(c) partial sums from different network components are independent and we can write $B_N^2 = 1/N \sum_{c=1}^{c_N} Var\left(\sum_{i \in C_c} Y_i\right)$. Thus, in order to obtain a CLT it will suffice to verify Lyapunov's condition for the partial sums, i.e. show that

$$\frac{\sum_{c=1}^{c_N} E\left|\sum_{i \in C_c} Y_i\right|^{2+\delta}}{(NB_N)^{2+\delta}} \to 0.$$

But this follows from Assumption 1(d) and the conditions of the theorem by:

$$\frac{\sum_{c=1}^{c_N} E\left|\sum_{i \in C_c} Y_i\right|^{2+\delta}}{(NB_N)^{2+\delta}} \leq \frac{K_N^{\frac{2+\delta}{2}} \sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(L_N \frac{N}{c_N} \sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} = \left(\frac{K_N}{L_N}\right)^{\frac{2+\delta}{2}} \left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} = o(1)$$

□

C Proof of Theorem 1

As mentioned in the text N_c's are really a function of N but, to economise on notation, we will only make this explicit when necessary.

Part (a). We have $N_c = N_1, \forall c$, and $c_N N_1/N = 1$. Using that and Assumption 1(d) we obtain:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq \frac{\underline{\sigma}^2 N_1^{1+\gamma_1} c_N}{N N_1^{\gamma_1}} = \frac{\underline{\sigma}^2 N_1 c_N}{N} = O(1) \tag{2}$$

which verifies conditions (a) and (b) of Theorem 5 with $L_N = O(1)$.

With equal rates we have:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} = \frac{c_N N_1^{\frac{2+\delta}{2}}}{N^{\frac{2+\delta}{2}}} = c_N^{-\frac{\delta}{2}}$$

which converges to zero if the number of components grows, which is implied by $N_1/N \to 0$.

Finally, using the bound in Lemma 1 and taking $\log_{d_{max}-1}$, we get that $N_1/N \to 0$ is implied by

$$\log_a N - \log_{d_{max}-1} N < 0$$

which can be rewritten as:

$$\log N \frac{\log(d_{max}-1)-\log a}{\log a \log(d_{max}-1)} < 0.$$

This inequality is implied by the condition in part (a) of Theorem 1. Thus, the result follows from Theorem 5.

Part (b). Recall that N_1 denotes a component for which the number of non-zero correlations grows at the fastest rate, i.e. $N_c^{1+\gamma_c}/N_1^{1+\gamma_1} \to 0$ or $\to 1$ for all $c > 1$. For sufficiently large N, we get $N/c_N \leq N_1$ under condition (b)(i). Thus, we have:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O(1)\frac{c_N N_1}{N}\frac{\sum_{c=1}^{c_N} N_c^{1+\gamma_c}}{N_1 \sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O(1)\frac{\sum_{c=1}^{c_N} \frac{N_c^{1+\gamma_c}}{N_1^{1+\gamma_1}}}{\sum_{c=1}^{c_N} \frac{N_c^{\gamma_c}}{N_1^{\gamma_1}}}$$

Note that both the numerator and the denominator in the last expression are positive so the ratio will be bounded away from zero if the sum in the denominator converges.

In order to properly analyse the convergence of the series, let us define the "jump points" in c_N by $N_c^J = \{N : c_N = c_{N-1} + 1, c_N = c\}$. Now by condition (b)(ii) we have:

$$\sum_{c=1}^{c_N} \frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}} \leq \sum_{c=1}^{c_N} \frac{N_c(N_c^J)^{\gamma_c}}{N_1(N_c^J)^{\gamma_1}}$$

which gives an infinite sum indexed by c. For this sum to be finite we apply the ratio test (see e.g. Theorem 3.34 in Rudin (1976)), which requires:

$$\lim_{c_N \to \infty} \frac{N_{c_N+1}(N_{c_N+1}^J)^{\gamma_{c_N+1}}}{N_1(N_{c_N+1}^J)^{\gamma_1}} \Big/ \frac{N_{c_N}(N_{c_N}^J)^{\gamma_{c_N}}}{N_1(N_{c_N}^J)^{\gamma_1}} < 1.$$

This latter condition is satisfied (for sufficiently large N) by condition (b)(iii). This verifies assumptions (a) and (b) of Theorem 5 with $L_N = O(1)$.

In order to verify condition (c) of Theorem 5 note that:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq \left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{c_N N_1^{(1+\gamma_1)\frac{2+\delta}{2}}}{c_N^{\frac{2+\delta}{2}} N^{\epsilon\frac{2+\delta}{2}}} \leq \frac{c_N}{N}\left(\frac{N_1^{(1+\gamma_1)\frac{2+\delta}{\delta}}}{N}\right)^{\frac{\delta}{2}}$$

for some $\epsilon > 0$ by the fact that all components are growing at some positive rate. The last expression converges to zero since $c_N/N \to 0$ and condition (b)(iv) implies that $N_1^{(1+\gamma_1)\frac{2+\delta}{\delta}}/N \to 0$ by the same reasoning as for part (a).

Part (c). Here we have the number of fixed size components equal $c_N - c_k \leq N - \sum_{c=1}^{c_k} N_c$ which together with $N_c/N \to 0$ implies that $c_N/N = O(1)$. Furthermore:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O(1)\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c} + c_N - c_k}{\sum_{c=1}^{c_k} N_c^{\gamma_c} + (c_N - c_k)\bar{N}^u} = O(1)\frac{\sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N} + O(1)}{\sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N} + O(1)} \geq O(1)$$

where $\bar{N}^u$ denotes an upper bound on the number of nodes in a fixed size component. The last inequality follows from $N_c^{1+\gamma_c}/N \to 0$, which is implied by the condition in part (c) of the theorem (see reasoning in the proof of part (a)).

Similarly:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq O(1)\frac{\sum_{c=1}^{c_k}\left(\frac{N_c^{1+\gamma_c}}{N}\right)^{\frac{2+\delta}{2}} + O(N^{-\frac{\delta}{2}})}{\left(\sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N} + O(1)\right)^{\frac{2+\delta}{2}}} \to 0$$

which completes the proof of this part.

Part (d). As in the previous part:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O(1)\frac{\sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N} + O(1)}{\sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N} + O(1)} \geq O(1)$$

Note that $c_k/N \to 0$ and we have $(c_N - c_k)/N \leq (1 - \sum_{c=1}^{c_k} N_c/N) \leq 1$. This together with $(c_N - c_k)/N \to s > 0$ implies that $\sum_{c=1}^{c_k} N_c/N = O(1)$. Since $\sum_{c=1}^{c_k} N_c^{1+\gamma_c}/N \geq \sum_{c=1}^{c_k} N_c/N \geq \sum_{c=1}^{c_k} N_c^{\gamma_c}/N$, we obtain $\frac{B_N^2}{\frac{N}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O(1)$.

Moreover:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq O(1)\frac{\sum_{c=1}^{c_k}\left(\frac{N_c^{1+\gamma_c}}{N}\right)^{\frac{2+\delta}{2}} + O(N^{-\frac{\delta}{2}})}{\left(\sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N} + O(1)\right)^{\frac{2+\delta}{2}}} \tag{3}$$

as in the previous part, however now the expression involves infinite sums. Using the fact that the

slowest growing component grows at a rate N^{ϵ} for some $\epsilon > 0$ we can bound $c_k \leq N^{1-\epsilon}$:

$$\sum_{c=1}^{c_k} \left(\frac{N_c^{1+\gamma_c}}{N} \right)^{\frac{2+\delta}{2}} \leq N^{1-\epsilon} \left(\frac{N_1^{1+\gamma_1}}{N} \right)^{\frac{2+\delta}{2}} \leq \left(\frac{N_1^{(1+\gamma_1)\frac{2+\delta}{\delta}}}{N} \right)^{\frac{\delta}{2}}$$

and the last expression converges to zero by condition (d)(i), which finalises verification of (c) in Theorem 5.

Alternatively, condition (c) in Theorem 5 will be satisfied if $\sum_{c=1}^{c_k} \left(\frac{N_c^{1+\gamma_c}}{N} \right)^{\frac{2+\delta}{2}} = o(1)$. First, note that $\log_{d_{max}-1} a > 1 + \gamma_c, \forall c$, implies that $N_1^{1+\gamma_1}/N \to 0$ and we can write:

$$\sum_{c=1}^{c_k} \left(\frac{N_c^{1+\gamma_c}}{N} \right)^{\frac{2+\delta}{2}} \leq \left(\frac{N_1^{1+\gamma_1}}{N} \right)^{\frac{\delta}{2}} \sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N} = o(1) \sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N}$$

so it is enough to show that the latter sum converges using the ratio test. To show that first note that for N large enough we have:

$$\sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N} = \sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N_1^{1+\gamma_1}} \frac{N_1^{1+\gamma_1}}{N} \leq \sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N_1^{1+\gamma_1}} = \sum_{c=1}^{c_k} \frac{N_c}{N_1} \frac{N_c^{\gamma_c}}{N_1^{\gamma_1}} \leq \sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N_1^{\gamma_1}}$$

Now $\sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N_1^{\gamma_1}} = O(1)$ follows from the same argument as the one in part (b).

Finally, if all c_k components grow at the same rate N_1, the expression in (3) simplifies to:

$$O(1) \frac{c_k \left(\frac{N_1^{1+\gamma_1}}{N} \right)^{\frac{2+\delta}{2}} + O(N^{-\frac{\delta}{2}})}{\left(c_k \frac{N_1^{\gamma_1}}{N} + O(1) \right)^{\frac{2+\delta}{2}}}$$

and now $c_k \sim N/N_1$ which gives:

$$c_k \left(\frac{N_1^{1+\gamma_1}}{N} \right)^{\frac{2+\delta}{2}} \simeq \left(\frac{N_1^{1+\gamma_1\left(\frac{2+\delta}{\delta}\right)}}{N} \right)^{\frac{\delta}{2}}$$

and the last expression converges to zero under condition (d)(iii).

Part (e). Here we have $c_N/N \to 0$. We will distinguish three cases: 1) c_k components grow at the same rate, 2) $c_k/c_N \to 1$, 3) $c_k/c_N \to 0$.

1) Consider the case when all N_c's are equal. Note that we have $N/c_N \leq N_1$ and $c_k N_1/N \to 1$,

which implies $c_k N_1^{1+\gamma_1}/N \geq O(1)$. Thus:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O(1)\frac{c_k N_1^{1+\gamma_1} + (c_N - c_k)O(1)}{c_k N_1^{1+\gamma_1} + \frac{N}{c_N}(c_N - c_k)O(1)} = O(1)\frac{1 + \left(\frac{c_N}{c_k} - 1\right) O\left(N_1^{-(1+\gamma_1)}\right)}{1 + \left(1 - \frac{c_k}{c_N}\right) O\left(N c_k^{-1} N_1^{-(1+\gamma_1)}\right)} = O(1)$$

since $(\frac{c_N}{c_k} - 1)N_1^{-(1+\gamma_1)} = \frac{c_N - c_k}{N}\frac{N}{c_k N_1^{1+\gamma_1}} = o(1)$.

It remains to verify the condition (c) of Theorem 5, which follows by:

$$\begin{aligned}\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} &\leq \left(\frac{c_N N_1}{N}\right)^{\frac{2+\delta}{2}} \frac{c_k N_1^{(1+\gamma_1)\frac{2+\delta}{2}} + (c_N - c_k)O(1)}{\left(c_k N_1^{1+\gamma_1} + (c_N - c_k)O(N_1)\right)^{\frac{2+\delta}{2}}} \\ &= \left(\frac{c_N N_1}{N}\right)^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \frac{1 + \left(\frac{c_N}{c_k} - 1\right) O\left(N_1^{-(1+\gamma_1)\frac{2+\delta}{2}}\right)}{\left(1 + \left(\frac{c_N}{c_k} - 1\right) O\left(N_1^{-\gamma_1}\right)\right)^{\frac{2+\delta}{2}}} = \frac{c_N^{\frac{2+\delta}{2}}}{c_k^{1+\delta}} o(1) = o(1)\end{aligned}$$

where the second equality follows from $N \geq c_k N_1$ and the last one is due to the condition stated in the Theorem. Note that $c_N^{\frac{2+\delta}{2}}/c_k^{1+\delta} \to 0$ is implied by $c_k/c_N \to 1$.

2) We have:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq \frac{\frac{c_N}{N}\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} + \frac{c_N}{N}\frac{c_N/c_k - 1}{1/c_k \sum_{c=1}^{c_k} N_c^{\gamma_c}} O(1)}{1 + \frac{c_N/c_k - 1}{1/c_k \sum_{c=1}^{c_k} N_c^{\gamma_c}} O(1)} = \frac{\frac{c_N}{N}\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} + o(1)}{1 + o(1)}$$

since $1/c_k \sum_{c=1}^{c_k} N_c^{\gamma_c} > 0$. Assumptions (e)(i)-(ii) imply the first component is growing the fastest. Thus, using $c_N N_1/N \geq 1$:

$$\frac{c_N}{N}\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} \geq \frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_1 N_c^{\gamma_c}} = \frac{\sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N_1^{1+\gamma_1}}}{\sum_{c=1}^{c_k} \frac{N_c^{\gamma_c}}{N_1^{\gamma_1}}} \geq O(1)$$

where the last inequality follows from conditions (e)(ii)-(iii) following the same arguments as in the proof of part (b).

Furthermore, we obtain for some $\epsilon > 0$:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_c^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_c^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq \left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} N_1^{\frac{2+\delta}{2}} \frac{c_k N_1^{(1+\gamma_1)\frac{2+\delta}{2}} + (c_N - c_k)O(1)}{\left(c_k N_1 N^{\epsilon} + N_1(c_N - c_k)O(1)\right)^{\frac{2+\delta}{2}}}$$

$$= \left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} N_1^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \frac{\left(\frac{N_1^{\gamma_1}}{N^{\epsilon}}\right)^{\frac{2+\delta}{2}} + o(1)}{1 + o(1)}$$

which will converge to zero if $\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} N_1^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \left(\frac{N_1^{\gamma_1}}{N^{\epsilon}}\right)^{\frac{2+\delta}{2}} \to 0$. But:

$$\left(\frac{c_N}{N}\right)^{\frac{2+\delta}{2}} N_1^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \left(\frac{N_1^{\gamma_1}}{N^{\epsilon}}\right)^{\frac{2+\delta}{2}} = \left(\frac{c_N}{c_k}\right)^{\frac{\delta}{2}} c_N \left(\frac{N_1^{1+\gamma_1}}{N^{1+\epsilon}}\right)^{\frac{2+\delta}{2}} \leq \left(\frac{c_N}{c_k}\right)^{\frac{\delta}{2}} \frac{c_N}{N} \left(\frac{N_1^{(1+\gamma_1)\frac{2+\delta}{\delta}}}{N}\right)^{\frac{\delta}{2}}$$

which converges to zero if $N_1^{(1+\gamma_1)\frac{2+\delta}{\delta}}/N \to 0$, which is implied by condition (e)(iv) (recall that $c_N/N \to 0$ and $c_k/c_N \to 1$).

3) Consider the case $c_k/c_N \to 0$ now. We have:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_c^{\gamma_c}} \geq O\left(\frac{c_N}{N}\right) \frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c} + (c_N - c_k)O(1)}{\sum_{c=1}^{c_k} N_c^{\gamma_c} + (c_N - c_k)O(1)} \tag{4}$$

and we can either have the first or the second term in the denominator diverging faster.

Consider first the case when $1/c_N \sum_{c=1}^{c_k} N_c^{\gamma_c} \leq O(1)$. Now we can rewrite (4) as:

$$O(1)\frac{\sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N} + \left(\frac{c_N}{N}\right)\left(1 - \frac{c_k}{c_N}\right)O(1)}{\frac{1}{c_N}\sum_{c=1}^{c_k} N_c^{\gamma_c} + \left(1 - \frac{c_k}{c_N}\right)O(1)} \geq O(1)$$

using $\sum_{c=1}^{c_k} \frac{N_c^{1+\gamma_c}}{N} \geq \sum_{c=1}^{c_k} \frac{N_c}{N} \to 1$, where the last limit follows from $c_N/N \to 0$ (note that $\sum_{c=1}^{c_k} N_c = N - (c_N - c_k)O(1)$). Furthermore, for the case $1/c_N \sum_{c=1}^{c_k} N_c^{\gamma_c} \to \infty$ we write (4) as:

$$O(1)\frac{\frac{c_N}{N}\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} + \frac{c_N(c_N - c_k)}{N\sum_{c=1}^{c_k} N_c^{\gamma_c}}O(1)}{1 + \frac{c_N - c_k}{\sum_{c=1}^{c_k} N_c^{\gamma_c}}O(1)} = O(1)\frac{\frac{c_N}{N}\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} + o(1)}{1 + o(1)} \geq O(1)$$

where the last inequality follows by the same argument as in the proof of part 2) above, with the help of conditions (e)(ii)-(iii).

Finally, we can write:

$$\frac{c_N}{N}\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_c^{\gamma_k}} \leq \left(\frac{c_N N_1}{N}\right)^{\frac{2+\delta}{2}} c_N^{-\frac{\delta}{2}} \frac{\left(\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{N_1 \sum_{c=1}^{c_k} N_c^{\gamma_c}}\right)^{\frac{2+\delta}{2}} + \frac{c_N - c_k}{N_1 \sum_{c=1}^{c_k} N_c^{\gamma_c}} O(1)}{\left(1 + \frac{c_N - c_k}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} O(1)\right)^{\frac{2+\delta}{2}}}$$

$$\leq \frac{c_N}{N}\left(\frac{N_1^{\frac{2+\delta}{\delta}}}{N}\right)^{\frac{\delta}{2}} \frac{1 + o(1)}{\left(1 + \frac{c_N - c_k}{\sum_{c=1}^{c_k} N_c^{\gamma_c}} O(1)\right)^{\frac{2+\delta}{2}}} = o(1)$$

where we have used:

$$\frac{\sum_{c=1}^{c_k} N_c^{1+\gamma_c}}{N_1 \sum_{c=1}^{c_k} N_c^{\gamma_c}} \leq 1$$

$$\frac{1}{c_N} N_1 \sum_{c=1}^{c_k} N_c^{\gamma_c} \geq \frac{1}{c_N} \sum_{c=1}^{c_k} N_c^{1+\gamma_c} \geq \frac{N}{c_N} \frac{\sum_{c=1}^{c_k} N_c}{N} \to \infty$$

and $N_1^{\frac{2+\delta}{\delta}}/N \to 0$ follows from condition (e)(iv). Hence, this part of Theorem 1 follows from Theorem 5.

D Proof of Theorem 2

We have:

$$\begin{aligned} N^2 Var(\hat{B}_N - B_N) &= E\left(\sum_{i=1}^{N}\sum_{j}(Y_i Y_j - E(Y_i Y_j))\mathbb{1}\{l(i,j) < \infty\}\right)^2 \\ &= \sum_{i=1}^{N}\sum_{j}\sum_{k}\sum_{l} E[(Y_i Y_j - E(Y_i Y_j))(Y_k Y_l - E(Y_k Y_l))]\mathbb{1}\{l(i,j) < \infty\}\mathbb{1}\{l(k,l) < \infty\} \\ &= \sum_{i=1}^{N}\sum_{j}\sum_{k}\sum_{l} Cov(Y_i Y_j, Y_k Y_l)\mathbb{1}\{l(i,j) < \infty\}\mathbb{1}\{l(k,l) < \infty\}. \end{aligned} \tag{5}$$

But $\mathbb{1}\{l(i,j) < \infty\}\mathbb{1}\{l(k,l) < \infty\} = 0$ unless i and j belong to the same network component and same happens for k and l. But for such pairs of (i,j) and (k,l) we have $Cov(Y_i Y_j, Y_k Y_l) \neq 0$ only when (i,j) and (k,l) belong to the same network component (see Assumption 1(c)'). Thus, we can

rewrite (5) as:

$$(5) = \sum_{c=1}^{c_N} \sum_{i \in C_c} \sum_{j \in C_c} \sum_{k \in C_c} \sum_{l \in C_c} Cov(Y_i Y_j, Y_k Y_l) \leq M \sum_{c=1}^{c_N} N_c^4$$

because $Cov(Y_i Y_j, Y_k Y_l) \leq M$ for bounded M by the assumption that $E|Y_i|^4$ is bounded (in the statement of the theorem) and Cauchy-Schwartz inequality.

Now consider the three cases in the theorem:

(a) $\sum_{c=1}^{c_N} N_1^4 = c_N N_1^4 = N N_1^3 = o(N^2)$ by the condition $\log_{d_{max}-1} a > 3$,

(b) $\sum_{c=1}^{c_N} N_c^4 \leq c_N N_1^4 = \frac{c_N}{N} N N_1^4 = o(N^2)$ by $c_N/N = o(1)$ and the condition $\log_{d_{max}-1} a > 4$,

(c) $\sum_{c=1}^{c_N} N_c^4 \leq c_k N_1^4 + (c_N - c_k) O(1) = o(N^2)$ by $c_N = o(N^2)$ and the condition $\log_{d_{max}-1} a > 2$,

(d) $\sum_{c=1}^{c_N} N_c^4 \leq c_k N_1^4 + o(N^2)$ by the same argument as above. Now, with equal components, $c_k N_1 = O(N)$, and $\log_{d_{max}-1} a > 3$ implies that the final expression is $o(N^2)$. Otherwise, $c_k N_1^4 = o(N) N_1^4 = o(N^2)$ where the last equality follows from $\log_{d_{max}-1} a > 4$.

E Proof of Theorem 3

First note that the statements (i) $\frac{N_c(N)^{1+\gamma_c}}{N_1(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_c(N)}{N_1(N)} \to 0$, (ii) $\frac{N_c(N)^{\gamma_c}}{N_1(N)^{\gamma_1}}$ is weakly decreasing in N, (iii) there exists $M < \infty$ such that $\frac{N_{c_{\tilde{N}}}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N}(N)^{\gamma_N}} < \left(\frac{N_1(\tilde{N})}{N_1(N)}\right)^{\gamma_1}$ for all $(N, \tilde{N})$ such that $c_{\tilde{N}} = c_N + 1$ and $N > M$, are equivalent (respectively) to statements:

(i)' $\frac{N_{c,f}(N)^{1+\gamma_c}}{N_{1,f}(N)^{1+\gamma_1}} \to 0 \Rightarrow \frac{N_{c,f}(N)}{N_{1,f}(N)} \to 0$,

(ii)' $\frac{N_{c,f}(N)^{\gamma_c}}{N_{1,f}(N)^{\gamma_1}}$ is weakly decreasing in N,

(iii)' there exists $M < \infty$ such that $\frac{N_{c_{\tilde{N}},f}(\tilde{N})^{\gamma_{\tilde{N}}}}{N_{c_N,f}(N)^{\gamma_N}} < \left(\frac{N_{1,f}(\tilde{N})}{N_{1,f}(N)}\right)^{\gamma_1}$ for all $(N, \tilde{N})$ such that $c_{\tilde{N}} = c_N + 1$ and $N > M$,

For most parts the proof follows by the same argument as in the proof of Theorem 1 above, with N_f replacing N, $N_{c,f}$ replacing N_c and $B_{N,f}$ replacing B_N. Thus, we present only arguments that differ. W.l.o.g. we often write N_c^2 instead of $N_c(N_c - 1)$ as these are of the same order.

Part (b). Condition (a). of Theorem 5 is satisfied by the same argument as in the proof of Theorem 1. Note that $N_f = \sum_{c=1}^{c_N} N_c(N_c - 1)$ is minimised subject to $\sum_{c=1}^{c_N} N_c = N$ by setting

$N_c = N/c_N$, which implies $N_f \geq N(N-1)/c_N$. For condition (c), by the same reasoning as above, we have:

$$\left(\frac{c_N}{N_f}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_{c,f}^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_{c,f}^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq \frac{c_N N_{1,f}^{(1+\gamma_1)\frac{2+\delta}{2}}}{N_f^{\frac{2+\delta}{2}}}$$

and the last expression is bounded by:

$$\frac{c_N^{2+\frac{\delta}{2}} N_1^{(1+\gamma_1)(2+\delta)}}{N^{2+\delta}} = \left(\frac{c_N}{N}\right)^{2+\frac{\delta}{2}} \left(\frac{N_1^{2(1+\gamma_1)\frac{2+\delta}{\delta}}}{N}\right)^{\frac{\delta}{2}} = o(1)$$

where the last equality is implied by $c_N/N \to 0$ and condition (b) in the statement of the theorem.

Part (c). Here the proof follows the same lines as in Theorem 1 with a difference that now we require $N_{1,f}^{1+\gamma_1}/N \to 0$ which is implied by the condition in the statement of the theorem.

Part (d). Consider unequal components case first. Note that we have $\sum_{c=1}^{c_k} N_{c,f}^{\gamma_1}/N_f \leq \sum_{c=1}^{c_k} N_{c,f}/N_f = O(1) \leq \sum_{c=1}^{c_k} N_{c,f}^{1+\gamma_1}/N_f$, which implies:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_{c,f}^{\gamma_c}} \geq O\left(\frac{c_N}{N}\right) \frac{\sum_{c=1}^{c_k} N_{c,f}^{1+\gamma_c}/N_f + o(1)}{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}/N + O(1)} \geq \frac{O(1)}{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}/N + O(1)}$$

so it remains to show that $\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}/N = O(1)$. Since:

$$\frac{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}}{N} = \frac{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}}{N_{1,f}^{\gamma_1}} \frac{N_{1,f}^{\gamma_1}}{N} \leq \frac{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}}{N_{1,f}^{\gamma_1}}$$

for N large enough, this follows from assumption (d)(i') in the theorem, noting that we also have $N_1^{2\gamma_1}/N \to 0$. Now to verify condition (c) of Theorem 5, by the same reasoning as in the proof of Theorem 1 we need $\sum_{c=1}^{c_k} \left(N_{c,f}^{1+\gamma_c}/N\right)^{\frac{2+\delta}{2}} \to 0$. Noting that:

$$\sum_{c=1}^{c_k} \left(\frac{N_{c,f}^{1+\gamma_c}}{N}\right)^{\frac{2+\delta}{2}} \leq \left(\frac{N_{1,f}^{1+\gamma_1}}{N}\right)^{\frac{\delta}{2}} \sum_{c=1}^{c_k} \frac{N_{c,f}^{1+\gamma_c}}{N}$$

condition (d)(i') implies both that $N_{1,f}^{1+\gamma_1}/N \to 0$ and $\sum_{c=1}^{c_k} N_{c,f}^{1+\gamma_c}/N$ converges, which proves the needed claim.

Consider now the case where c_k components grow at the same rate. Firstly, using $c_N/N_f \geq 1/N_{1,f}$:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_{c,f}^{\gamma_c}} \geq \frac{c_k N_{1,f}^{1+\gamma_1} + (c_N - c_k)O(1)}{c_k N_{1,f}^{1+\gamma_1} + (c_N - c_k)O(N_f/c_N)} = \frac{1 + \frac{c_N - c_k}{c_k N_{1,f}^{1+\gamma_1}} O(1)}{1 + \frac{(1-c_k/c_N)N_f}{c_k N_{1,f}^{1+\gamma_1}} O(1)}$$

and the last expression is bounded away from zero since $c_k N_{1,f}^{1+\gamma_1}/N_f \geq O(1)$ (see above) and:

$$\frac{c_N - c_k}{c_k N_{1,f}^{1+\gamma_1}} = \frac{c_N - c_k}{N_f} \frac{N_f}{c_k N_{1,f}^{1+\gamma_1}} = o(1).$$

Now to verify condition (c) of Theorem 5, just as before we need $c_k (N_{1,f}^{1+\gamma_1}/N)^{(2+\delta)/2} \to 0$ and using $c_k \sim N/N_1$ this expression becomes $N_1^{1+\delta+\gamma_1(2+\delta)}/N^{\delta/2} = (N_1^{2(1+\delta)/\delta + 2\gamma_1(2+\delta)/\delta}/N)^{\delta/2}$ and converges to zero by assumption (d)(ii)' of the theorem.

Part (e). First consider the case with c_k components growing at the same rate:

1) Using $c_N/N_f \geq 1/N_{1,f}$ and rearranging we have:

$$\frac{B_N^2}{\frac{1}{c_N}\sum_{c=1}^{c_N} N_{c,f}^{\gamma_c}} \geq O(1) \frac{1 + \frac{c_N - c_k}{c_k N_{1,f}^{1+\gamma_1}} O(1)}{1 + \left(1 - \frac{c_k}{c_N}\right) \frac{N_f}{c_k N_{1,f}^{1+\gamma_1}} O(1)}$$

Now $N_f = c_k N_{1,f} + (c_N - c_k)O(1)$ and $(c_N - c_k)/N_f \to 0$ imply $c_k N_{1,f}/N_f \to 1$, which gives $c_k N_{1,f}^{1+\gamma_1}/N_f \geq O(1)$ and $\frac{c_N - c_k}{c_k N_{1,f}^{1+\gamma_1}} = \frac{c_N - c_k}{N_f} \frac{N_f}{c_k N_{1,f}^{1+\gamma_1}} = o(1)$. This verifies condition (a) of Theorem 5.

Now proceeding as in the proof of Theorem 1 we get:

$$\left(\frac{c_N}{N_f}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_{c,f}^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_{c,f}^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq \left(\frac{c_N N_{1,f}}{N_f}\right)^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \frac{1 + O(1)(c_N - c_k)/(c_k N_{1,f}^{(1+\gamma_1)\frac{2+\delta}{2}})}{\left(1 + O(1)(c_N - c_k)/(c_k N_{1,f}^{\gamma_1})\right)^{\frac{2+\delta}{2}}} = \frac{c_N^{\frac{2+\delta}{2}}}{c_k^{1+\delta}} o(1) = o(1)$$

where the first equality follows from $\frac{c_N - c_k}{N_f} \frac{N_f}{c_k N_{1,f}^{1+\gamma_1}} = o(1)$ in the previous paragraph and $N_f \geq c_k N_{1,f}$ (which implies $N_{1,f}/N_f \leq 1/c_k$). The final equality follows from condition (e) of the theorem.

2) Now consider the case $c_k/c_N \to 1$. First part of the proof follows by the same reasoning as

in Theorem 1 using assumptions (e)(ii)-(iii). Also, by similar reasoning as the one there:

$$\left(\frac{c_N}{N_f}\right)^{\frac{2+\delta}{2}} \frac{\sum_{c=1}^{c_N} N_{c,f}^{(1+\gamma_c)\frac{2+\delta}{2}}}{\left(\sum_{c=1}^{c_N} N_{c,f}^{\gamma_c}\right)^{\frac{2+\delta}{2}}} \leq \left(\frac{c_N}{N_f}\right)^{\frac{2+\delta}{2}} N_{1,f}^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \frac{\left(\frac{N_{1,f}^{\gamma_1}}{N^{\epsilon}}\right)^{\frac{2+\delta}{2}} + o(1)}{(1+o(1))^{\frac{2+\delta}{2}}}$$

Since:

$$\left(\frac{c_N}{N_f}\right)^{\frac{2+\delta}{2}} N_{1,f}^{\frac{2+\delta}{2}} c_k^{-\frac{\delta}{2}} \left(\frac{N_{1,f}^{\gamma_1}}{N^{\epsilon}}\right)^{\frac{2+\delta}{2}} \leq \left(\frac{c_N}{c_k}\right)^{\frac{\delta}{2}} \frac{c_N}{N} \left(\frac{N_{1,f}^{1+\gamma_1} N^{\frac{2}{2+\delta}}}{N_f}\right)^{\frac{2+\delta}{2}}$$

A sufficient condition for the last expression to converge to zero is $N_{1,f}^{1+\gamma_1} N^{\frac{2}{2+\delta}}/N_f \to 0$. As in the proof of part (b), we have $N_f \geq N(N-1)/c_N$, which gives:

$$\frac{N_{1,f}^{1+\gamma_1} N^{\frac{2}{2+\delta}}}{N_f} \leq \frac{c_N}{N} \left(\frac{N_1^{2(1+\gamma_1)\frac{2+\delta}{\delta}}}{N}\right)^{\frac{\delta}{2+\delta}}$$

and the latter converges to zero by condition $\log_{d_{max}-1} a > 2(1+\gamma_c)\frac{2+\delta}{\delta}$.

3) Here $c_k/c_N \to 0$. Condition (a) of Theorem 5 follows by an argument mirroring the one in the proof of Theorem 1. To verify condition (c) note that:

$$\begin{aligned}\frac{c_N}{N_f} \frac{\sum_{c=1}^{c_k} N_{c,f}^{1+\gamma_c}}{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_k}} &\leq \left(\frac{c_N N_{1,f}}{N_f}\right)^{\frac{2+\delta}{2}} c_N^{-\frac{\delta}{2}} \frac{\left(\frac{\sum_{c=1}^{c_k} N_{c,f}^{1+\gamma_c}}{N_{1,f}\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}}\right)^{\frac{2+\delta}{2}} + \frac{c_N - c_k}{N_{1,f}\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}} O(1)}{\left(1 + \frac{c_N - c_k}{\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}} O(1)\right)^{\frac{2+\delta}{2}}} \\ &= O(1)\left(\frac{c_N N_{1,f}}{N_f}\right)^{\frac{2+\delta}{2}} c_N^{-\frac{\delta}{2}} \leq O(1)\frac{c_N}{N}^{2+\frac{\delta}{2}} \left(\frac{N_1^{\frac{2(2+\delta)}{\delta}}}{N}\right)^{\frac{\delta}{2}} = o(1)\end{aligned}$$

where we have used:

$$\frac{\sum_{c=1}^{c_k} N_{c,f}^{1+\gamma_c}}{N_{1,f}\sum_{c=1}^{c_k} N_{c,f}^{\gamma_c}} \leq 1, \quad \frac{1}{c_N} N_{1,f} \sum_{c=1}^{c_k} N_{c,f}^{\gamma_c} \to \infty, \quad N_f \geq N(N-1)/c_N$$

and the last equality follows from $\log_{d_{max}-1} a > \frac{2(2+\delta)}{\delta}$. The result now follows by Theorem 5.

F Proof of Theorem 4

For a sequence of random variables $\{W_1, \ldots, W_N\}$ define the U_N operator as:

$$U_N h = \frac{1}{N(N-1)} \sum_{i \neq j} h(W_i, W_j) \mathbb{1}\{l(i,j) < \infty\}.$$

where h is a symmetric kernel. By Hoeffding decomposition:

$$\frac{\sqrt{N}Y_c}{B_{N,c}} = B_{N,c}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_1(Y_i) \frac{N_c(i) - 1}{N-1} + \frac{1}{2} B_{N,c}^{-1} \sqrt{N} U_N h_2 \tag{6}$$

where $h_2(y_1, y_2) = h(y_1, y_2) - h_1(y_1) - h_1(y_2)$. Let us first show that $B_{N,c}^{-1}\sqrt{N}U_N h_2 = o_p(1)$.

We have:

$$\begin{aligned} Var(\sqrt{N}U_N h_2) &= \frac{1}{N(N-1)^2} \sum_{i=1}^{N} \sum_{j \neq i} \sum_{k=1}^{N} \sum_{l \neq k} E[h_2(Y_i, Y_j) h_2(Y_k, Y_l)] \mathbb{1}\{l(i,j) < \infty\} \mathbb{1}\{l(k,l) < \infty\} \\ &= \frac{1}{N(N-1)^2} \sum_{c=1}^{c_N} \sum_{i \in C_c} \sum_{j \in C_c, j \neq i} \sum_{k \in C_c} \sum_{l \in C_c, l \neq k} E[h_2(Y_i, Y_j) h_2(Y_k, Y_l)] \end{aligned}$$

since the term under the sum is only nonzero if (i,j,k,l) belong to the same component (note that $Eh_2(Y_i, Y_j) = 0$).

As in the proof of Theorem 3.1 in Kojevnikov et al. (2021), let $H_{N_c}(s,m)$ be defined as the sets of nodes $\{i,j,k,l\}$ where $\{i,j\}$ and $\{k,l\}$ are both in the m-neighbourhood from each other and the network distance between $\{i,j\}$ and $\{k,l\}$ is at least s, formally: $H_{N_c}(s,m) = \{(i,j,k,l) : l(i,j) \leq m, l(k,l) \leq m, l(\{i,j\},\{k,l\}) \geq s\}$. We have $H_{N_c}(s,m) \leq 4N_c c_{N_c}(s,m;2)$ (ibid.). Now by Lemma 2 we can bound:

$$\begin{aligned} N^3 Var(\sqrt{N}U_N h_2) &= \sum_{c=1}^{c_N} \sum_{s \geq 0} \sum_{\substack{\{i,j,k,l\} \in H_{N_c}(s,N_c) \\ j \neq i, l \neq k}} E[h_2(Y_i, Y_j) h_2(Y_k, Y_l)] \leq \sum_{c=1}^{c_N} \sum_{s \geq 0} |H_{N_c}(s, N_c)| \alpha_c(s) \\ &\leq 4 \sum_{c=1}^{c_N} N_c \sum_{s \geq 0} c_{N_c}(s, N_c; 2) \alpha_c(s). \end{aligned}$$

Now Assumption 2(b) implies that $Var(B_{N,c}^{-1}\sqrt{N}U_N h_2) = o(1)$.

Finally, the asymptotic normality of the first element in (6) follows from Theorem 3.2 in Ko-

jevnikov et al. (2021). To see that define $X_{i,N} = h_1(Y_i)\frac{N_c(i)-1}{N-1}$ note that the triangular array $\{X_{i,N}\}_{i=1}^{\infty}$ is strong mixing with coefficients $\alpha_N(\cdot)$ and the conditions of their theorem are implied by Assumptions 2(a), (c) (note that $N_c/N \leq 1$).

G Additional MC simulations

Table 3: Simulated coverage, edge-specific means (contrasts), normal errors, 95% level

	BA model					SW model				
N	m	z_a	d_{max}	Δ_N	Coverage oracle	p	k	d_{max}	Δ_N	Coverage oracle
	Equal Components									
500	1	0	10	10	0.944	0.05	2	6	12	0.957
1000	1	0	10	11	0.953	0.05	2	6	15	0.951
5000	1	0	10	13	0.953	0.05	2	7	20	0.946
10000	1	0	10	14	0.942	0.05	2	7	20	0.948
500	1	1	10	16	0.946	0.05	5	13	5	0.931
1000	1	1	10	16	0.941	0.05	5	13	6	0.940
5000	1	1	10	23	0.950	0.05	5	14	7	0.945
10000	1	1	10	19	0.963	0.05	5	15	8	0.939
500	1	2	10	14	0.953	0.05	10	24	3	0.946
1000	1	2	10	15	0.941	0.05	10	24	4	0.935
5000	1	2	10	18	0.947	0.05	10	26	5	0.948
10000	1	2	10	21	0.955	0.05	10	26	5	0.949
500	2	0	20	7	0.959	0.1	2	8	9	0.956
1000	2	0	20	8	0.958	0.1	2	7	10	0.946
5000	2	0	20	11	0.943	0.1	2	8	14	0.942
10000	2	0	20	11	0.951	0.1	2	9	15	0.956
500	2	1	20	7	0.953	0.1	5	14	4	0.953
1000	2	1	20	7	0.933	0.1	5	14	5	0.952
5000	2	1	20	9	0.941	0.1	5	16	6	0.955
10000	2	1	20	11	0.945	0.1	5	16	7	0.935
500	2	2	20	6	0.948	0.1	10	25	3	0.932
1000	2	2	20	7	0.955	0.1	10	25	4	0.949
5000	2	2	20	9	0.940	0.1	10	28	4	0.948
10000	2	2	20	9	0.936	0.1	10	30	5	0.933
	Growing + Fixed									
500	1	0	10	10	0.956	0.05	2	6	12	0.949
1000	1	0	10	13	0.952	0.05	2	6	15	0.952
5000	1	0	10	12	0.953	0.05	2	7	20	0.947
10000	1	0	10	13	0.941	0.05	2	7	20	0.962
500	1	1	10	11	0.958	0.05	5	13	5	0.944
1000	1	1	10	14	0.960	0.05	5	13	6	0.941
5000	1	1	10	16	0.961	0.05	5	14	7	0.939
10000	1	1	10	17	0.955	0.05	5	15	8	0.946
500	1	2	10	16	0.962	0.05	10	24	3	0.948
1000	1	2	10	17	0.961	0.05	10	24	4	0.959
5000	1	2	10	23	0.954	0.05	10	26	5	0.959
10000	1	2	10	22	0.950	0.05	10	26	5	0.947
500	2	0	20	7	0.928	0.1	2	8	9	0.943
1000	2	0	20	8	0.949	0.1	2	7	10	0.940
5000	2	0	20	9	0.948	0.1	2	8	14	0.947
10000	2	0	20	10	0.933	0.1	2	9	15	0.944
500	2	1	20	7	0.952	0.1	5	14	4	0.958
1000	2	1	20	7	0.953	0.1	5	14	5	0.953
5000	2	1	20	10	0.955	0.1	5	16	6	0.951
10000	2	1	20	10	0.948	0.1	5	16	7	0.959
500	2	2	20	6	0.943	0.1	10	25	3	0.952
1000	2	2	20	7	0.945	0.1	10	25	4	0.944
5000	2	2	20	9	0.949	0.1	10	28	4	0.944
10000	2	2	20	9	0.953	0.1	10	30	5	0.947

Note: 1000 Monte Carlo simulations, 1000 bootstrap replications. “Oracle” – known variance.

References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A. & Tahbaz-Salehi, A. (2012), 'The network origins of aggregate fluctuations', *Econometrica* **80**(5), 1977–2016.

Baldi, P. & Rinott, Y. (1989), 'On normal approximations of distributions in terms of dependency graphs', *The Annals of Probability* **17**(4), 1646 – 1650.

Barabási, A.-L. & Albert, R. (1999), 'Emergence of scaling in random networks', *Science (American Association for the Advancement of Science)* **286**(5439), 509–512.

Bickel, P., Choi, D., Chang, X. & Zhang, H. (2013), 'Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels', *The Annals of Statistics* **41**(4), 1922–1943.

Bickel, P. J., Chen, A. & Levina, E. (2011), 'The method of moments and degree distributions for network models', *The Annals of Statistics* **39**(5), 2280 – 2301.

Bollobás, B. & Riordan, O. (2004), 'The diameter of a scale-free random graph', *Combinatorica* **24**(1), 5–34.

Bollobás, B. & Riordan, O. M. (2002), *Mathematical results on scale-free random graphs*, John Wiley & Sons, Ltd, chapter 1, pp. 1–34.

Cameron, A. C., Gelbach, J. B. & Miller, D. L. (2008), 'Bootstrap-based improvements for inference with clustered errors', *The Review of Economics and Statistics* **90**(3), 414–427.

Canay, I. A., Romano, J. P. & Shaikh, A. M. (2017), 'Randomization tests under an approximate symmetry assumption', *Econometrica* **85**(3), 1013–1030.

Chen, L. H. Y. & Shao, Q.-M. (2004), 'Normal approximation under local dependence', *The Annals of Probability* **32**(3), 1985–2028.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barberá, P., Bhole, M. & Wernerfelt, N. (2022), 'Social capital I: measurement and associations with economic mobility', *Nature* **608**(7921), 108–121.

Djogbenou, A. A., MacKinnon, J. G. & Ørregaard Nielsen, M. (2019), 'Asymptotic theory and wild bootstrap inference with clustered errors', *Journal of Econometrics* **212**(2), 393–412.

Hansen, B. E. & Lee, S. (2019), 'Asymptotic theory for clustered samples', *Journal of Econometrics* **210**(2), 268–290.

Jackson, M. O. (2008), *Social and economic networks*, Princeton University Press, Princeton, N.J. ; Woodstock.

Jenish, N. & Prucha, I. R. (2012), 'On spatial processes and asymptotic inference under near-epoch dependence', *Journal of Econometrics* **170**(1), 178–190.

Kojevnikov, D., Marmer, V. & Song, K. (2021), 'Limit theorems for network dependent random variables', *Journal of Econometrics* **222**(2), 882–908.

Kojevnikov, D. & Song, K. (2023), 'Some impossibility results for inference with cluster dependence with large clusters', *Journal of Econometrics* **237**(2, Part A), 105524.

Kuersteiner, G. M. (2019), Limit theorems for data with network structure. Working paper.

Kuersteiner, G. M. & Prucha, I. R. (2013), 'Limit theory for panel data models with cross sectional dependence and sequential exogeneity', *Journal of Econometrics* **174**(2), 107–126.

Leung, M. P. (2023), 'Network cluster-robust inference', *Econometrica* **91**(2), 641–667.

Leung, M. P. & Moon, H. R. (2023), Normal approximation in large network models. Working Paper.

MacKinnon, J. G., Ørregaard Nielsen, M. & Webb, M. D. (2023), 'Cluster-robust inference: A guide to empirical practice', *Journal of Econometrics* **232**(2), 272–299.

Matsushita, Y. & Otsu, T. (2023), 'Empirical likelihood for network data', *Journal of the American Statistical Association* **0**(0), 1–12.

Ogburn, E. L., Sofrygin, O., Díaz, I. & van der Laan, M. J. (2024), 'Causal inference for social network data', *Journal of the American Statistical Association* **119**(545), 597–611.

Pineda-Villavicencio, G. & Wood, D. R. (2015), 'The degree-diameter problem for sparse graph classes', *The Electronic Journal of Combinatorics* **22**(2), 1–20.

Romano, J. P. & Wolf, M. (2000), 'A more general central limit theorem for m-dependent random variables with unbounded m', *Statistics & Probability Letters* **47**(2), 115–124.

Rudin, W. (1976), *Principles of Mathematical Analysis*, Third edition edn.

Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. (2011), The anatomy of the Facebook social graph. Working Paper.

Watts, D. J. (1999), *Small worlds: the dynamics of networks between order and randomness*, Princeton studies in complexity, Princeton University Press, Princeton, N.J.

Watts, D. J. & Strogatz, S. H. (1998), 'Collective dynamics of "small-world" networks', *Nature* **393**(6684), 440–442.