Network Cluster-Robust Inference*

Michael P. Leung†

February 21, 2021

Abstract. Since network data commonly consists of observations on a single large network, researchers often partition the network into clusters in order to apply cluster-robust inference methods. All existing such methods require clusters to be asymptotically independent. We prove that for this requirement to hold, under certain conditions, it is necessary and sufficient for clusters to have low conductance, the ratio of edge boundary size to volume, which yields a measure of cluster quality. We show in simulations that, for important classes of networks lacking low-conductance clusters, cluster-robust methods can exhibit substantial size distortion. To assess the existence of low-conductance clusters and construct them, we draw on results in spectral graph theory showing a close connection between conductance and the spectrum of the graph Laplacian. Based on these results, we propose to use the spectrum to compute the number of low-conductance clusters and spectral clustering to compute the clusters. We illustrate our results and proposed methods in simulations and empirical applications.

JEL Codes: C12, C21, C38

Keywords: social networks, clustered standard errors, graph Laplacian, spectral clustering

*First draft: Jan. 2021.

†Department of Economics, University of Southern California. E-mail: leungm@usc.edu.

1 Introduction

Cluster-robust methods are widely used to account for cross-sectional dependence (Cameron and Miller, 2015). The standard model of cluster dependence partitions the set of observations into a large number of clusters such that observations across clusters are independent (Hansen and Lee, 2019). However, in many settings, observations cannot be divided into independent, mutually exclusive clusters. Temporally and spatially dependent data typically have the property that correlation between observations decays with distance but never exactly reaches zero. Recent econometric work develops cluster-robust methods applicable to data of this type (Bester et al., 2011; Canay et al., 2017, 2020; Ibragimov and Müller, 2010, 2016). We build on this literature to study the performance of these methods under *network* dependence.

Clustered standard errors are frequently used with network data. When there exists a large set of plausibly independent networks, for example, geographically isolated villages (e.g. Banerjee et al., 2013), the standard asymptotics are applicable. However, it is common to observe only a single large network, say the friendship network of a school or village, in which case it may be unclear how to partition the network into clusters. For example, Miguel and Kremer (2004) study a setting with many schools but allow for cross-school treatment spillovers. Since observations in different schools may then be correlated, it is not clear whether clustering, say, at the school level yields valid standard errors.

Several papers approach this problem by using a "community detection" or "network clustering" unsupervised learning algorithm to divide the network into disjoint subnetworks according to some criteria and then clustering standard errors on the subnetworks (e.g. Aral and Nicolaides, 2017; Aral and Zhao, 2019; Zacchia, 2020). However, it is unclear whether the criteria used to construct these subnetworks deliver a set of clusters that can be used for statistically valid inference. This is because the clusters are linked (hence the need for partitioning), so they are not generally independent. Additionally, these algorithms often depend on tuning parameters that can be chosen to mechanically increase the number of clusters the algorithm outputs, and there are no guidelines on how to choose this number for the purpose of inference.

An alternative to cluster-robust inference is to use HAC variance estimators, which are well-known methods of adjusting for spatial or temporal dependence. In the context of network dependence, the same estimators may be used by interchanging tem-

poral or spatial distance with network (path) distance (Kojevnikov, 2021; Kojevnikov et al., 2020). However, simulation evidence has shown that tests using these estimators tend to over-reject in smaller samples (Conley et al., 2018; Ibragimov and Müller, 2010).

To our knowledge, there is no theoretical justification for applying cluster-robust methods to network-dependent data. Zacchia (2020) invokes the work of Bester et al. (2011) to justify network clustering, but the latter's results are specific to spatial data. For HAC estimators, complications arise in the choice of bandwidth when switching from Euclidean to network distance (Kojevnikov et al., 2020; Leung, 2020). A motivating question for this paper is whether cluster-robust methods also encounter novel complications when applied to network-dependent data.

Contributions. We show that complications do exist. Whereas weakly dependent spatial data can always be partitioned into a set of "quality" clusters (we will define what we mean by "quality"), this turns out not to be the case for network data. We show that cluster-robust methods applied to networks that lack quality clusters can exhibit substantial size distortion. This motivates the methods provided in this paper for diagnosing whether quality clusters exist and how to construct them.

We derive conditions on the clusters and data-generating process under which cluster-robust inference procedures are valid under network dependence. To our knowledge, all existing such methods require asymptotic independence of clusters. Bester et al. (2011) provide primitive sufficient conditions for asymptotic independence for spatial data, the key assumption being a restriction on the growth rate of cluster boundaries. For network data, we show that, under certain conditions, it is necessary and sufficient for clusters to have low *conductance*, which is the ratio of a cluster's edge boundary size to its volume (defined in §2.1). This yields a simple $[0, 1]$-measure of cluster quality and suggests an (infeasible) objective for constructing clusters: choose the set that minimizes conductance. The importance of conductance connects the literature on cluster-robust inference to results in spectral graph theory and spectral clustering, which we draw on to feasibly construct clusters.

Due to the topology of Euclidean space, clusters satisfying the boundary condition can always be constructed for spatial data (under increasing domain asymptotics). However, we show that in the network setting, this may not be possible, depending on the underlying process that generates the network. As we discuss, for low-conductance clusters to exist, the network must have a sufficiently small (higher-order) *Cheeger*

constant, which is a well-known graph invariant that measures network segregation. Some classes of networks satisfy this condition, but many apparently do not.[1] We therefore require methods to determine the existence of low-conductance clusters and, if they exist, to construct them.

Computing the Cheeger constant or set of clusters minimizing conductance turns out to be infeasible. Fortunately, Cheeger inequalities imply that the lower eigenvalues of the graph Laplacian (defined in §2.2) are informative about the constant's magnitude. A simple argument (Proposition 1) shows that a set of L low-conductance clusters exists if only if the Lth smallest eigenvalue is small, providing a practical diagnostic for determining both the existence and number of good clusters. To then compute the clusters, we can apply k-means clustering to the eigenvectors, which corresponds to spectral clustering, a widely used unsupervised learning algorithm.

Our simulation results show that there are advantages to using cluster-robust methods for network data, relative to HAC estimators. We find that the randomization test of Canay et al. (2017) better controls size in smaller samples, provided clusters have low conductance. However, when no such clusters exist, the test exhibits substantial size distortion even in large samples, unlike the HAC estimator. Our theory suggests this is due to the fact that clusters in this case do not satisfy the requirement of asymptotic independence.

Based on these results, we make three recommendations for empirical practice in §2.3. These concern how to assess whether a given set of clusters is of sufficient "quality" (compute the conductance), how to assess whether quality clusters exist (compute the spectrum of the Laplacian), and how to compute such clusters if they exist (apply spectral clustering or other community detection algorithms).

Community detection algorithms can output a small number of clusters, as is the case in our simulation results. Conventional clustered standard errors can perform poorly in this setting (Cameron and Miller, 2015). Our results therefore utilize asymptotics sending the sizes of a fixed number of clusters to infinity (Bester et al., 2011; Canay et al., 2017, 2020; Ibragimov and Müller, 2010, 2016). Our formal results provide interpretable primitive conditions under which the key independence assumption imposed by these papers holds. We do not develop a new inference procedure; rather, we provide diagnostics to assess whether these existing procedures are valid when applied to network data and an asymptotic theory supporting these diagnostics.

[1] A simple example is a fully connected network, but we will discuss more realistic examples.

Related Literature. There is a large, cross-disciplinary literature on spectral clustering. von Luxburg (2007) is a well-known reference in computer science. A growing literature in statistics studies the performance of spectral clustering for estimating stochastic block models (e.g. Lei and Rinaldo, 2015; Rohe et al., 2011). The goal of this literature is to recover latent types ("communities"), and theoretical results concern convergence of the sample Laplacian to a population Laplacian that identifies the types. We are instead interested in spectral clustering from the graph conductance perspective of identifying small-boundary clusters, regardless of their association or lack thereof with some underlying parameter such as type. This perspective seems to be emphasized more in the computer science literature (see e.g. Trevisan, 2016).

Jochmans and Weidner (2019) establish a connection between the second-largest eigenvalue of the graph Laplacian and the rate of convergence of fixed-effect network regressions. Their results suggest the practical importance of computing the spectrum to assess estimator precision. Our paper highlights the usefulness of the spectrum for a different purpose, namely to assess the validity of cluster-robust methods under network dependence.

As noted by Conley et al. (2018), there is little work on how to best group observations into clusters even for non-network data. Ibragimov and Müller (2010) show that there is no data-dependent way to "optimally" construct clusters while maintaining uniform size control but nonetheless note that this is an important practical issue. Ibragimov and Müller (2016) make some progress along this direction by providing a test for whether a finer cluster partition is valid compared to a coarser one. Partitioning space into equally-sized rectangles satisfies the boundary condition of Bester et al. (2011), but with irregularly spaced data, it may be possible to do better. Recent work by Müller and Watson (2021) addresses this problem by constructing novel standard errors using the principal components of the variance of a baseline spatial model and selecting the number of components to minimize the expected length of the confidence interval.

The outline of the paper is as follows. In the next section, we state the model, summarize our main results and their intuition, and make recommendations for empirical practice. We present the asymptotic theory in §3. Then in §4, we discuss the use of spectral clustering for constructing clusters. We provide theoretical and simulation results on the spectra of various geometric and random graphs in §5, showing that some important classes of graphs lack quality clusters. In §6, we present sim-

ulation results comparing the randomization test to HAC estimators and apply our results to two empirical studies. Finally, §7 concludes.

2 Setup and Overview

We observe a set of units $\mathcal{N}_n = \{1, \ldots, n\}$, data $W_i \in \mathbb{R}^{d_w}$ associated with each unit i, and an undirected network or graph $\boldsymbol{A}$ on $\mathcal{N}_n$. We represent $\boldsymbol{A}$ as a binary, symmetric adjacency matrix with ijth entry $A_{ij} = 1$ signifying a link between i and j and $A_{ij} = 0$ signifying its absence. We assume no self-links, so that $A_{ii} = 0$ for all i. Remark 1 below discusses possible extensions to weighted, directed networks.

Our analysis treats $\boldsymbol{A}$ as fixed (conditioned upon), whereas $\{W_i\}_{i=1}^n$ is random and not necessarily identically distributed. Let $\theta_0 \in \mathbb{R}^{d_\theta}$ be the true parameter of interest and $g\colon \mathbb{R}^{d_w} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_g}$ a moment function such that

$$\mathbf{E}[g(W_i, \theta_0)] = \mathbf{0} \quad \forall\, i \in \mathcal{N}_n.$$

The goal is inference on θ_0. Define the standard generalized method of moments (GMM) estimator

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \hat{G}(\theta)' \Psi \hat{G}(\theta),$$

where $\hat{G}(\theta) = n^{-1} \sum_{i=1}^n g(W_i, \theta)$ is the sample moment vector and Ψ a weighting matrix. For example, to recover parameters of linear-in-means-type models, Aral and Nicolaides (2017) and Zacchia (2020) both use instrumental variables estimators, which are well known special cases of GMM.

Various papers cited in the introduction develop cluster-robust methods for this setting when $\{W_i\}_{i=1}^n$ satisfies weak temporal or spatial dependence. We instead employ a notion of weak network dependence, formally defined in §3, which is conceptually similar to mixing conditions used in time series and spatial econometrics. The key difference is the metric, which is path distance. For any two units i, j, their *path distance* $\ell_{\boldsymbol{A}}(i, j)$ is the length of the shortest path between them in $\boldsymbol{A}$.[2] Weak network dependence essentially demands that the correlation between W_i and W_j decays to zero as $\ell_{\boldsymbol{A}}(i, j) \to \infty$.

[2] A *path* between i, j is a sequence of links $A_{k_1 k_2}, A_{k_2 k_3}, \ldots, A_{k_{m-1} k_m} = 1$ such that $k_1 = i$, $k_m = j$, and $k_a \neq k_b$ for all $a, b \in \{1, \ldots, m\}$. The *length* $\ell_{\boldsymbol{A}}(i, j)$ of this path is $m - 1$. If $i \neq j$ and no path between i, j exists, then we define $\ell_{\boldsymbol{A}}(i, j) = \infty$. If $i = j$, we define $\ell_{\boldsymbol{A}}(i, j) = 0$.

Clusters. Cluster-robust methods take as input a partition of $\mathcal{N}_n$ into L clusters, which we denote by $\{\mathcal{C}_\ell\}_{\ell=1}^L$. Each $\mathcal{C}_\ell$ implicitly depends on the network $\boldsymbol{A}$ since different networks may be partitioned differently. Being a partition, $\cup_{\ell=1}^L \mathcal{C}_\ell = \mathcal{N}_n$ and $\mathcal{C}_\ell \cap \mathcal{C}_m = \varnothing$ for all $\ell \neq m$. In practice, clusters will typically constitute *connected subnetworks* in the sense that $\ell_{\boldsymbol{A}}(i,j) < \infty$ for all $i,j \in \mathcal{C}_\ell$ and any cluster ℓ because community detection algorithms usually output connected subnetworks. Also, a union of k disconnected subnetworks is better treated as k distinct clusters since, being disconnected, such subnetworks are uncorrelated under the weak dependence concept we use, and cluster-robust methods are more powerful with a larger number of clusters.

In general, the observed network may consist of multiple *components*, which are connected subnetworks that are disconnected (in the sense of infinite path distance) from the rest of the network. Under weak network dependence, observations in different components are uncorrelated, so components can therefore be treated as separate clusters. This implies that, if a network consists of many components, standard many-cluster asymptotics are applicable. However, a well-known stylized fact about real-world networks is that they typically possess a *giant component* containing the vast majority of units (formally, order n units), and all other components are small (formally, being $o(n)$ and typically $O(\log n)$ in size; see Barabási, 2015). For example, the giant of the Facebook graph contains 99.91 percent of all units, whereas its second-largest component only has about 2000 units (Ugander et al., 2011). Therefore, the key task for clustering is partitioning the giant, which is the main part of the network. All other components, being small, can be treated as individual clusters.

Inference Procedures. Suppose we have a set of clusters. Let $\hat{\theta}_\ell$ be the GMM estimator computed only using observations in cluster ℓ and $\hat{G}_\ell(\theta) = n_\ell^{-1} \sum_{i \in \mathcal{C}_\ell} g(W_i, \theta)$, the sample moment vector constructed only using these observations. Cluster-robust methods use estimates $(\hat{\theta}_\ell)_{\ell=1}^L$ or moments $(\hat{G}_\ell(\hat{\theta}_\ell))_{\ell=1}^L$ (possibly for constrained versions of $\hat{\theta}_\ell$) to construct tests. A commonly used method is a wild bootstrap procedure due to Cameron et al. (2008), whose formal properties under fixed-L asymptotics are studied by Canay et al. (2020). Their results, as well as those of Bester et al. (2011), require clusters to satisfy certain homogeneity conditions, which are not imposed by Canay et al. (2017) and Ibragimov and Müller (2010).

Cai et al. (2021) argue that the randomization test of Canay et al. (2017) has a number of attractive properties relative to the other alternatives. We next summarize this test since it will be the focus of our simulation study in §6. Let $S_n = (\sqrt{n}(\hat{\theta}_\ell -$

$\theta))_{\ell=1}^{L}$, and define the Wald statistic

$$T(S_n) = \left(\frac{1}{\sqrt{L}}\sum_{\ell=1}^{L}\sqrt{n}(\hat{\theta}_\ell - \theta)'\right)\left(\frac{1}{L}\sum_{\ell=1}^{L} n(\hat{\theta}_\ell - \theta)(\hat{\theta}_\ell - \theta)'\right)^{-1}\left(\frac{1}{\sqrt{L}}\sum_{\ell=1}^{L}\sqrt{n}(\hat{\theta}_\ell - \theta)\right),$$

where θ is the null value of θ_0. One can alternatively use subvectors of $\hat{\theta}_\ell$ and θ. For $\pi = (\pi_\ell)_{\ell=1}^{L} \in \{-1, 1\}^L$, let $\pi S_n = (\pi_\ell\sqrt{n}(\hat{\theta}_\ell - \theta))_{\ell=1}^{L}$. The test rejects if

$$T(S_n) > T^{(k)}(S_n), \tag{1}$$

where $k = \lceil 2^L(1-\alpha)\rceil$, α is the level of the test, and $T^{(k)}(S_n)$ is the kth largest value of $\{T(\pi S_n) \colon \pi \in \{-1, 1\}^L\}$.

2.1 Conductance

In §3, we provide conditions under which a given set of clusters can be used for asymptotically valid cluster-robust inference. We consider a sequence of networks with associated clusters indexed by the network size n and take n to infinity, while keeping the number of clusters L fixed. For economy of language, we often simply refer to a network rather than a sequence of networks when discussing asymptotic results.

Let $n_\ell = |\mathcal{C}_\ell|$, the cardinality of $\mathcal{C}_\ell$, and $\hat{G}_\ell(\theta) = n_\ell^{-1}\sum_{i\in\mathcal{C}_\ell} g(W_i, \theta)$. Under weak network dependence and standard regularity conditions, we establish that

$$\tfrac{1}{\sqrt{n}}\begin{pmatrix} n_1\hat{G}_1(\theta_0) \\ \vdots \\ n_L\hat{G}_L(\theta_0)\end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^*), \quad \mathbf{\Sigma}^* = \begin{pmatrix} \rho_1\mathbf{\Sigma}_1 & \sqrt{\rho_1\rho_2}\,\mathbf{\Sigma}_{12} & \cdots & \sqrt{\rho_1\rho_L}\,\mathbf{\Sigma}_{1L} \\ \sqrt{\rho_2\rho_1}\,\mathbf{\Sigma}_{21} & \rho_2\mathbf{\Sigma}_2 & \cdots & \sqrt{\rho_2\rho_L}\,\mathbf{\Sigma}_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{\rho_L\rho_1}\,\mathbf{\Sigma}_{L1} & \sqrt{\rho_L\rho_2}\,\mathbf{\Sigma}_{L2} & \cdots & \rho_L\mathbf{\Sigma}_L \end{pmatrix}, \tag{2}$$

where $\rho_\ell = \lim_{n\to\infty} n_\ell/n$, $\mathbf{\Sigma}_{\ell m} = \lim_{n\to\infty}\mathrm{Cov}(\sqrt{n_\ell}\hat{G}_\ell(\theta_0), \sqrt{n_m}\hat{G}_m(\theta_0))$, and $\mathbf{\Sigma}_\ell = \mathbf{\Sigma}_{\ell\ell}$. This is an elementary but key intermediate result for establishing that the vector of GMM estimates $(\sqrt{n}(\hat{\theta}_\ell - \theta_0))_{\ell=1}^{L}$ is asymptotically normal.

The cluster-robust methods previously cited all require asymptotic independence in the sense that the off-diagonal blocks $\mathbf{\Sigma}_{\ell m}$ are zero for all $\ell \neq m$ (e.g. Canay et al., 2020, Assumption 2.2(i)). Assumption 3.1(ii) of Canay et al. (2017) imposes symme-

try of the limit distribution, which, under the group of transformations considered in their §4 and our (1), corresponds to having off-diagonal blocks equal to zero. We therefore interpret zero off-diagonal blocks as the key requirement for the validity of cluster-robust methods. Our goal is to provide restrictions on the network and clusters under which this holds.

We next introduce some standard definitions from spectral graph theory (Chung, 1997; Trevisan, 2016).

Definition 1. For any $S \subseteq \mathcal{N}_n$, define its *edge boundary size* with respect to $\boldsymbol{A}$ as the number of links involving a unit in S and a unit not in S:

$$|\partial_{\boldsymbol{A}}(S)| = \sum_{i \in S} \sum_{j \in \mathcal{N}_n \setminus S} A_{ij}.$$

The *volume* of S is $\mathrm{vol}_{\boldsymbol{A}}(S) = \sum_{i \in S} \sum_{j=1}^{n} A_{ij}$, the sum of the *degrees* $\sum_{j=1}^{n} A_{ij}$ of units i in S. Finally, the *conductance* of S (assuming it has at least one link) is

$$\phi_{\boldsymbol{A}}(S) = \frac{|\partial_{\boldsymbol{A}}(S)|}{\mathrm{vol}_{\boldsymbol{A}}(S)}.$$

The conductance is the probability that a randomly chosen neighbor of a randomly chosen unit in S lies outside of S. The denominator is a trivial upper bound on the numerator since all units in S may only be connected to units outside of S. In addition to delivering a $[0, 1]$ measure, normalizing by the volume ensures that small sets do not have low conductance simply by virtue of having few members or few links.

Our main assumption for guaranteeing $\boldsymbol{\Sigma}_{\ell m} = \mathbf{0}$ for all $\ell \neq m$ is

$$\max_{1 \leqslant \ell \leqslant L} \phi_{\boldsymbol{A}}(\mathcal{C}_\ell) \to 0 \quad \text{as} \quad n \to \infty. \tag{3}$$

This says *the maximal conductance of the clusters is small*. Theorem 1 establishes sufficiency under additional conditions and Theorem 2 necessity. The intuition for (3) is as follows. As we will see, weak network dependence requires restrictions on the sizes of K*-neighborhoods*, where the K-neighborhood of a unit i is

$$\mathcal{N}_{\boldsymbol{A}}(i, K) = \{j \in \mathcal{N}_n \colon \ell_{\boldsymbol{A}}(i, j) \leqslant K\},$$

recalling that $\ell_{\boldsymbol{A}}(i, j)$ is path distance. In sparse networks, the size of this set is asymp-

totically bounded for any i, K. This ensures a type of increasing domain asymptotics as $n \to \infty$, meaning that units are minimally spaced apart in large networks. For example, it rules out a completely connected network where all units are distance 1 apart. Now, if (3) holds, this means that, for two clusters of order n size, the number of links connecting the clusters is $o(n)$. Consequently, given that units in the network are minimally spaced apart, most units in one cluster will be far from units in the other for large n, which will imply asymptotic independence of clusters since correlation decays with distance. This is the same intuition for temporally (Ibragimov and Müller, 2010) and spatially (Bester et al., 2011) correlated data, but it was not previously obvious how this extends to network data. As for necessity, as far as we are aware, there are no prior results, but the intuition is similar. If (3) fails, then we could have each unit in one cluster directly linked to a unit in the other, in which case the clusters can be strongly correlated.

Note that (3) does not mean the giant asymptotically fractures into L distinct components. For any given cluster, it requires the number of cross-cluster links to be small relative to the number of links emanating from units in the cluster. However, any cluster can still have many links to any other cluster. Figure 1 plots two random graphs from simulations in §5.2, coloring units by clusters obtained via spectral clustering (see §4). Connections within clusters are denser than connections across clusters, which means conductance is low. Indeed, the clusters in the left and right panels have maximal conductance 0.13 and 0.07, respectively.

2.2 Graph Invariants

The next natural question is whether a sequence of clusters satisfying (3) necessarily exists for any given network sequence. In the spatial setting, one can always construct clusters satisfying the required boundary condition under increasing domain asymptotics (Bester et al., 2011, Assumption 2(iv)), for example by partitioning $\mathbb{R}^2$ into rectangles, but this is not true in general for networks.

For any integer $k > 1$, define the *kth-order Cheeger constant* of $\boldsymbol{A}$

$$h_k(\boldsymbol{A}) = \min \left\{ \max_{1 \leqslant \ell \leqslant k} \phi_{\boldsymbol{A}}(S_\ell) \colon S_1, \ldots, S_k \text{ partitions } \mathcal{N}_n \right\}. \tag{4}$$

This is a well-known graph invariant that measures network segregation. If $\boldsymbol{A}$ has

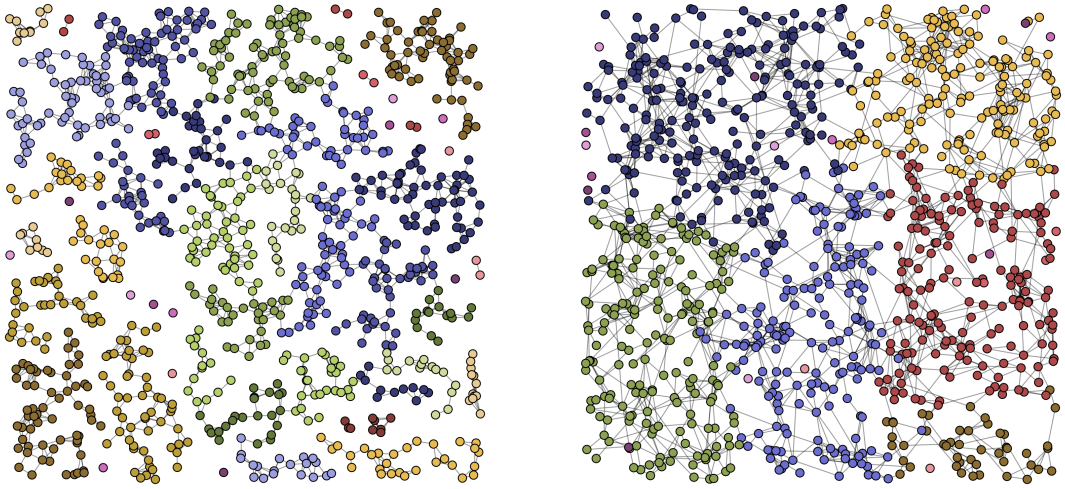

Figure 1: Low-conductance clusters of a random geometric graph (left panel) and a draw from the random connections model (right panel), obtained by spectral clustering. See §5 for a description of these models.

k components, then it is maximally segregated, and $h_k(\boldsymbol{A}) = 0$, whereas if $\boldsymbol{A}$ is completely connected, then $h_k(\boldsymbol{A}) = 1$ for any k. The outer minimization problem corresponds to solving an *adjusted* "mincut" problem (von Luxburg, 2007). By contrast, the goal of *unadjusted* mincut is to divide the network into k subnetworks with a minimal number of links between the subnetworks. The problem with unadjusted mincut is that it does not produce a satisfactory partition in practice because it does not account for subnetwork size; it often generates singleton clusters obtained by cutting a single link. The adjusted mincut problem in (4) accounts for this by dividing by volume in the definition of conductance. Thus if we set the numerator of $\phi_{\boldsymbol{A}}(S)$ to be one for all S, for instance, then the minimum is achieved by a partition such that all of its elements have equal volume.

Clearly, a necessary condition for (3) is

$$h_L(\boldsymbol{A}) \to 0. \tag{5}$$

Thus, if $h_L(\boldsymbol{A})$ could be computed, it would provide a simple way to assess whether low-conductance clusters exist. Furthermore, the argmin would be the best possible partition for cluster-robust inference. Unfortunately, this optimization problem is

NP-complete (Šíma and Schaeffer, 2006).

Fortunately, the spectrum of the graph Laplacian, which can be efficiently computed, is highly informative about the magnitude of the Cheeger constant. Let $\boldsymbol{D}$ be the $n \times n$ diagonal matrix with iith entry equal to i's degree $\sum_{j=1}^{n} A_{ij}$. The *(normalized) Laplacian* of $\boldsymbol{A}$ is

$$\boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2},$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix. Let us order the eigenvalues of the Laplacian as

$$\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n.$$

The following facts are well known: $\lambda_k \in [0, 2]$ for all k, and $\lambda_k = 0$ if and only if $\boldsymbol{A}$ has at least k components (hence $\lambda_1 = 0$) (Chung, 1997). The latter property suggests that if λ_k is close to zero, then $\boldsymbol{A}$ should contain a set of k clusters with low conductance since having k components is the "ideal" case of k clusters with zero maximal conductance.

The (higher-order) Cheeger inequality formalizes this intuition by relating Cheeger constants to the spectrum of the Laplacian as follows:

$$\frac{\lambda_k}{2} \leqslant h_k(\boldsymbol{A}) \leqslant C\lambda_k^{1/2}, \tag{6}$$

where C is a constant that does not depend on n and is $O(k^2)$ (Lee et al., 2014, Theorem 1.1).[3] This yields the following simple corollary.

Proposition 1. *$h_L(\boldsymbol{A}) \to 0$ if and only if $\lambda_L \to 0$.*

Proof. The "if" direction is immediate from both inequalities in (6). The "only if" direction follows from the second inequality and the fact that $\lambda_k \in [0, 2]$ for all k. ∎

The proposition gives us a feasible way of assessing (5), which is to examine the magnitude of λ_L. Based on (5) and the proposition, we make the following definition.

[3]See Chung (1997) and Trevisan (2016) for proofs for $k = 2$. The lower bound holds because, from the variational characterization of λ_k, one can rewrite λ_k as the optimum of an objective that corresponds to a continuous relaxation of the optimization problem corresponding to $h_k(\boldsymbol{A})$.

Definition 2. A sequence of networks is *well clustered at* L if $\lambda_L \to 0$ as $n \to \infty$. It is *well clustered* if it is well clustered at L for some $L > 1$ and *poorly clustered* otherwise.

That is, well-clustered networks can be partitioned into L low-conductance clusters, which can be used for cluster-robust inference by our asymptotic theory. Note that this is a very minimal notion of being well clustered, and in practice, we should aim for $L \geqslant 5$ clusters that are not unbalanced, as discussed in §2.3 below.

The next natural question is what kinds of networks are well clustered, meaning what models of network formation generate well-clustered networks. Our discussion in §5 indicates that a variety of networks satisfy this condition, but there are important classes of networks that apparently do not. It is therefore important to assess whether a given network is well clustered in practice.

2.3 Practical Recommendations

Based on these results, we make several recommendations for empirical practice.

Conductance. For any candidate set of clusters $\{\mathcal{C}_\ell\}_{\ell=1}^L$, however it is obtained, one should compute its maximal conductance $\max_{1\leqslant\ell\leqslant L} \phi_{\boldsymbol{A}}(\mathcal{C}_\ell) \in [0,1]$. By (3), the goal is to obtain a value close to zero, and our simulations in §6.1 indicate that values as high as about 0.1 can still ensure adequate size control. The remaining two recommendations concern whether we can find and how to find such clusters.

Laplacian. The ideal set of clusters constitutes the partition that minimizes conductance. As previously discussed, an exact solution is computationally infeasible, which motivates the use of spectral methods. Specifically, we would like to choose L such that λ_L is small and λ_{L+1} is large. If such an L exists, then the Cheeger inequality implies L low-conductance clusters exist but $L+1$ such clusters do not. Of course, there are no universal thresholds for "small" and "large." Nonetheless, this heuristic is widely used in practice to determine the number of clusters for spectral clustering and principal components analysis (von Luxburg, 2007).

Any heuristic is necessarily subjective, as is quite clear when one inspects the spectra of various graphs (see e.g. Figure 2). Different definitions of "small" or "large" can potentially generate rather different clusters. Fortunately for us, we are not interested in interpreting the clusters themselves, as is usually the case when applying

community detection algorithms, but rather in finding a set of low-conductance clusters. One can therefore compute as many partitions as desired using any number of algorithms in order to find one with the smallest conductance.[4]

Still, we would like to make a more specific recommendation for how to use the spectrum to choose L. The discussion above involves a combination of two heuristics. What might be called the "cutoff heuristic" specifies a desired cutoff c and chooses L satisfying $\lambda_L \leqslant c < \lambda_{L+1}$. On the other hand, the "gap heuristic" simply solves $\text{argmax}_L(\lambda_{L+1} - \lambda_L)$. Let $L_1(c)$ be the result of the cutoff heuristic under some chosen cutoff c and L_2 that of the gap heuristic. In §5.2, we illustrate the need to combine both heuristics and present simulation results using the following rule that combines the two, which may be a reasonable starting point in practice:

$$L = \mathbf{1}\{L_2 < 5\} \max\{L_1(c), L_2\} + \mathbf{1}\{L_2 \geqslant 5\} \min\{L_1(c), L_2\}. \tag{7}$$

To understand the idea, first note that if L is chosen very small, cluster-robust methods have little power, and an alternative procedure may be preferable. The simulation results of Cameron et al. (2008) and Cai et al. (2021) show good performance of their respective methods for clusters of size as few as five, so $L \geqslant 5$ seems to be a good rule of thumb. Now, as we will discuss in §5.2, some graphs have a large gap very early in the spectrum, in which case L_2 is potentially too small. In this case, the heuristic errs on the side of potentially obtaining more clusters by taking the max of the two values. In other graphs, the location of the gap is quite random, potentially yielding huge values of L_2. In this case, the heuristic errs on the side of obtaining fewer, better clusters with lower conductance by taking the minimum. Our simulation results in §5.2 indicate that using this rule with cutoffs $c < 0.05$ seems to produce values of L such that, for well-clustered graphs, spectral clustering (described in §4) delivers clusters with maximal conductance near 0.1, so these are reasonable starting points.

Computing clusters. Many algorithms are available for this purpose. We focus on spectral clustering in our simulation results. In §4, we define the algorithm and discuss why it computes low-conductance clusters when they exist.

Unbalanced clusters. For some networks, community detection algorithms may

[4]Provided these algorithms only use the network data $\boldsymbol{A}$, as is the case for most community detection algorithms, our asymptotic theory remains valid since it treats $\boldsymbol{A}$ as fixed.

usually return an *unbalanced partition* consisting of one large cluster and several very small clusters; see Example 4 below. Such networks may be thought of as being close to poorly clustered. Result (2) implies that only large clusters (of order n size) contribute to the limit distribution, so this situation is little better than having only the large cluster, a setting where cluster-robust methods have trivial power. Consequently, an alternative inference procedure should be used.

It may be tempting to rectify an unbalanced partition by changing the tuning parameters of various network clustering algorithms to mechanically increase the number of clusters beyond what the spectral gap suggests, essentially by producing "deeper cuts" in the large cluster. We strongly recommend against this practice because some of the resulting clusters will have high conductance, violating (3). For instance, if the spectral gap identifies the number of clusters as L, so that λ_L is close to zero but λ_{L+1} is far from zero, then using a deeper cut to obtain $L+1$ clusters means some cluster has high conductance by the Cheeger inequality since λ_{L+1} is large.

Remark 1 (Weighted, directed graphs)**.** Our asymptotic results rely on a CLT that only pertains to binary, undirected networks, but as discussed in Kojevnikov (2021), this can be extended to weighted networks. The definitions of conductance, the Cheeger constant, and the Laplacian immediately apply to weighted ($A_{ij} \in \mathbb{R}$) networks and have been generalized to directed ($A_{ij} \neq A_{ji}$) networks. It is likely possible to formalize a notion of weak network dependence for such networks under which cluster-robust inference is valid when the conductance is asymptotically negligible. Furthermore, the Cheeger inequality applies directly to weighted graphs (Lee et al., 2014) and has been extended to directed graphs (Chung, 2005). Consequently, we believe that our recommendations are also relevant for these types of networks.

3 Asymptotic Theory

We consider a sequence of networks, each network associated with a partition, with both implicitly indexed by the network size n. Recall that $n_\ell = |\mathcal{C}_\ell|$, the size of the ℓth cluster.

Assumption 1 (Limit Sequence)**.** *(a) The number of clusters L in each partition is fixed as $n \to \infty$. (b) For any $\ell = 1, \dots, L$, $n_\ell/n \to \rho_\ell \in [0, \infty)$.*

We consider a small number of clusters in part (a) because the partitions generated by spectral clustering in our simulations are quite small in size. Part (b) defines ρ_ℓ as the asymptotic fraction of units in $\mathcal{C}_\ell$, allowing for the possibility that the cluster has trivial size ($\rho_\ell = 0$).

To formalize weak network dependence, one approach is to adapt a notion of temporal or spatial weak dependence by replacing temporal or spatial distance with path distance. Kojevnikov et al. (2020) take this approach, adapting the concept of ψ-weak dependence, which we employ in what follows. They and Leung (2020) verify ψ-weak dependence for a number of network applications.

Weak dependence simply means the correlation between two sets of observations decays as the network distance between the sets grows. Formalizing this requires some notational overhead. For any $H, H' \subseteq \mathcal{N}_n$, define $\ell_{\boldsymbol{A}}(H, H') = \min\{\ell_{\boldsymbol{A}}(i,j) \colon i \in H, j \in H'\}$, the distance between two sets. Let $\mathcal{L}_d$ be the set of bounded $\mathbb{R}$-valued Lipschitz functions on $\mathbb{R}^d$, $\|f\|_\infty = \sup_x |f(x)|$, $\text{Lip}(f)$ the Lipschitz constant of $f \in \mathcal{L}_d$, and

$$\mathcal{P}_n(h, h'; s) = \{(H, H') \colon H, H' \subseteq \mathcal{N}_n, |H| = h, |H'| = h', \ell_{\boldsymbol{A}}(H, H') \geqslant s\},$$

the set of pairs of sets H, H', with respective sizes h, h', that are at least distance s apart in the network. Define $G_H = (g(W_i, \theta_0))_{i \in H}$, the vector of moments for units in H, and $M_n(s,k) = n^{-1} \sum_{i=1}^n |\mathcal{N}_{\boldsymbol{A}}(i,s)|^k$, the kth moment of the s-neighborhood size. Finally, let

$$\mathcal{H}_n(s,m) = \left\{(i,j,k,\ell) \in \mathcal{N}_n^4 \colon k \in \mathcal{N}_{\boldsymbol{A}}(i,m), \ell \in \mathcal{N}_{\boldsymbol{A}}(j,m), \ell_{\boldsymbol{A}}(\{i,k\},\{j,\ell\}) = s\right\}.$$

This is the set of paired couples (i,j) and (k,ℓ) with the property that the two pairs are exactly path distance s apart, i,k are at most m apart from one another, and likewise with j,ℓ.

Assumption 2 (Weak Network Dependence I)**.**

(a) There exist a constant $C > 0$ and uniformly bounded constants $\{\psi_n(s)\}_{s,n \in \mathbb{N}}$ with $\psi_n(0) = 1$ for all n such that $\sup_n \psi_n(s) \to 0$ as $s \to \infty$ and

$$|Cov(f(G_H), f'(G_{H'}))| \leqslant Chh'(\|f\|_\infty + Lip(f))(\|f'\|_\infty + Lip(f'))\psi_n(s)$$

for all $n, h, h' \in \mathbb{N}$; $s > 0$; $f \in \mathcal{L}_{d_g h}$; $f' \in \mathcal{L}_{d_g h'}$; and $(H, H') \in \mathcal{P}_n(h, h'; s)$.

(b) There exist $p > 4$ and a sequence of positive constants $\{m_n\}_{n\in\mathbb{N}}$ such that $m_n \to \infty$ and

$$\max\left\{\frac{1}{n^2}\sum_{s=0}^{n}|\mathcal{H}_n(s,m_n)|\psi_n(s)^{1-4/p},\quad n^{-1/2}M_n(m_n,2),\quad n^{3/2}\psi_n(m_n)^{1-1/p}\right\}\to 0. \tag{8}$$

This imposes weak network dependence on the set of moments.[5] Part (a) encodes the definition of ψ-weak dependence and Assumption 2.1 of Kojevnikov et al. (2019). Part (b) is Assumption 3.4 of the same reference.[6] The key quantity is $\psi_n(s)$, which essentially bounds the correlation between sets of observations at distance s and is required to decay to zero as s diverges. Leung (2020) shows that $\psi_n(s)$ is uniformly $O(e^{-c\,s})$ for some $c > 0$ in well-known social interactions models.

Part (b) is analogous to mixing conditions for spatial data that require the mixing coefficient to decay sufficiently quickly. In the network setting, this is necessarily more complicated to state because the metric space is non-Euclidean. As discussed in Leung (2020), whereas the number of units in a ball of radius K grows polynomially in Euclidean space, it can grow exponentially in a network. Part (b) requires $\psi_n(s)$ to decay sufficiently quickly relative to the growth rates of s-neighborhoods, the equivalent of balls in Euclidean space, and this is conceptually the same requirement underlying spatial mixing conditions.

Appendix A of Leung (2020) verifies part (b) for different classes of graphs. The second term in (8) restricts s-neighborhood growth rates. For instance, if $|\mathcal{N}_{\boldsymbol{A}}(i,s)|$ is uniformly bounded by an exponential function of s, then choosing m_n to grow logarithmically with n ensures that $M_n(m_n,2) = O(1)$. The third term in (8) requires sufficiently fast decay of $\psi_n(s)$. In the case of exponential decay in s, the assumption is satisfied for m_n diverging at a logarithmic rate. Finally, the first term of (8) essentially requires $\psi_n(s)$ to decay to zero fast enough relative to s-neighborhood sizes. See §A.2 in the appendix for further discussion.

Assumption 3 (Regularity)**.** *(a) $Cov(\sqrt{n_\ell}\hat{G}_\ell(\theta_0), \sqrt{n_m}\hat{G}_m(\theta_0)) \to \boldsymbol{\Sigma}_{\ell m}$ finite for any $\ell, m = 1,\ldots,L$, with $\boldsymbol{\Sigma}_{\ell\ell}$ positive definite for any ℓ. (b) For p in Assumption 2(b),*

[5] It can be verified given an analogous weak network dependence condition imposed on the data $\{W_i\}_{i=1}^n$ and smoothness conditions on $g(\cdot)$ (Kojevnikov et al., 2020, Appendix A.1).

[6] This can be replaced with Assumption 3.4 of Kojevnikov et al. (2020), but we find the 2019 version easier to use.

$\sup_{n\in\mathbb{N}} \max_{i\in\mathcal{N}_n} \mathbf{E}[\|g(W_i,\theta_0)\|^p]^{1/p} < \infty$.

Proposition 2. *Under Assumptions 1–3,* (2) *holds.*

Assumption 3 is standard. The proof of the proposition and all other results in this section are given in §A.1. Under additional standard regularity conditions, we can establish joint asymptotic normality of $(\sqrt{n}(\hat{\theta}_\ell - \theta_0))_{\ell=1}^L$. Since this type of result is well known, we omit these conditions and the corresponding result.

We next state conditions required for our first main result, which shows that the off-diagonal blocks of $\mathbf{\Sigma}^*$ in (2) are zero. Combined with Proposition 2, this verifies the key high-level condition required by cluster-robust inference methods for a small number of clusters. Let $\delta(\boldsymbol{A}) = n^{-1}\sum_{i=1}^n \sum_{j=1}^n A_{ij}$, the average degree.

Assumption 4 (Conductance)**.** $\max_{1\leqslant\ell\leqslant L} \phi_{\boldsymbol{A}}(\mathcal{C}_\ell)\cdot\delta(\boldsymbol{A}) \to 0$.

This requires the largest conductance over elements of the partition to vanish as the network grows since $\delta(\boldsymbol{A})$ generally is bounded away from zero (otherwise, the network would be empty in the limit). In dense networks, $\delta(\boldsymbol{A})$ diverges, in which case the assumption requires the maximal conductance to shrink to zero faster. However, for settings with nontrivial network dependence, Assumption 2 requires $\boldsymbol{A}$ to be sparse in the sense that $\delta(\boldsymbol{A}) = O(1)$.

Assumption 5 (Weak Network Dependence II)**.** $\sum_{s=1}^n sM_n(s,2(1+\epsilon))^{1/(1+\epsilon)}\psi_n(s) = O(1)$ *for some* $\epsilon > 0$.

This is conceptually the same as Assumption 2(b), requiring dependence to decay quickly enough relative to the growth rate of s-neighborhood sizes. We verify the condition for some examples in §A.2. Note that validity of the HAC estimator (Kojevnikov et al., 2020, Proposition 4.1) does not require Assumptions 4 or 5; instead, it requires conditions relating the bandwidth and kernel to the network topology.

Theorem 1 (Sufficiency)**.** *Under Assumptions 1–5,* $\sqrt{\rho_\ell\rho_m}\,\mathbf{\Sigma}_{\ell m} = \mathbf{0}$ *for all* $\ell \neq m$.

As previously discussed, this justifies the use of cluster-robust methods for network data. Lemma 1 of Bester et al. (2011) is the analogous result for spatial data. Our

last result establishes the necessity of Assumption 4 for obtaining zero off-diagonals.

Theorem 2 (Necessity)**.** *Consider any sequence of networks and associated clusters such that Assumption 1 holds,* $\min_\ell \rho_\ell > 0$ *(only nontrivial clusters are used), each cluster is a connected subnetwork, and* $\delta(\boldsymbol{A}) = O(1)$ *(network is sparse), but clusters fail to satisfy Assumption 4. There exists a data-generating process for* $\{g(W_i, \theta_0)\}_{i=1}^n$ *satisfying Assumptions 2, 3, and 5 such that, for some* $\ell, m = 1, \ldots, L$ *with* $\ell \neq m$*,* $\sqrt{\rho_\ell \rho_m}\, \boldsymbol{\Sigma}_{\ell m} \neq \boldsymbol{0}$*.*

4 Constructing Clusters

Ideally we would like to find the set of clusters solving (4), but as discussed in §2.2, this is not computationally feasible. This partly motivates a large, multi-disciplinary literature on network clustering algorithms. Zacchia (2020) uses a popular modularity-based algorithm due to Blondel et al. (2008) to construct clusters. Such algorithms seek to find a partition that approximately minimizes a "modularity" criteria, which is a measure of community structure related to, but not quite the same as, conductance. Modularity-based algorithms are the subject of a large literature in computer science and physics (e.g. Barabási, 2015, Ch. 9).

Spectral clustering algorithms are another popular method (von Luxburg, 2007), which has been more directly shown to deliver low-conductance clusters. Given a desired number of clusters L, these algorithms apply k-means or some other clustering method to L eigenvectors of the Laplacian. One common version of the algorithm is the following.

1. Given a graph $\boldsymbol{A}$ and desired number of clusters L, compute the Laplacian and its eigenvalues $\lambda_1 \leqslant \ldots \leqslant \lambda_n$.

2. Let V_ℓ be the eigenvector associated with λ_ℓ and $V_{\ell i}$ its ith component. Embed the n units in $\mathbb{R}^L$ by associating each unit i with a position

$$\rho_i = \left(\frac{V_{1i}}{(\sum_{\ell=1}^L V_{\ell i}^2)^{1/2}}, \ldots, \frac{V_{Li}}{(\sum_{\ell=1}^L V_{\ell i}^2)^{1/2}} \right).$$

3. Cluster the positions $(\rho_i)_{i=1}^n$ using k-means with $k = L$ to obtain $\mathcal{C}_1, \ldots, \mathcal{C}_L$.

We use this in the simulations that follow. As discussed in von Luxburg (2007), it can be interpreted as a continuous relaxation of the ideal program (4).

There are well-known results justifying why spectral clustering produces low-conductance clusters, provided they exist, and our simulation results in the next sections support these results. Recall that if $\lambda_L = 0$, the network consists of L components, which are "ideal" clusters with exactly zero conductance. This intuitively suggests that if λ_L is close to zero, the network has L low-conductance clusters.

Consider an "ideal" network $\boldsymbol{A}^*$ consisting of L components. The eigenvector V_L associated with λ_L then almost perfectly identifies the clusters because it can be written as $V_L = \boldsymbol{D}^{1/2} V_L^*$, where $V_{Li}^* = V_{Lj}^*$ if and only if i, j are in the same component (Peng et al., 2017). That is, up to degree scaling due to $\boldsymbol{D}^{1/2}$, units in the same component are assigned the same value by V_L, whereas units in different components are assigned different values. Recovering the clusters is then a simple task for k-means (the normalization in the definition of ρ_i adjusts for degree heterogeneity).

Now suppose more realistically that the observed network $\boldsymbol{A}$ has Lth-order Cheeger constant that is small relative to λ_{L+1} (their ratio tends to zero). This implies $\boldsymbol{A}$ has L low-conductance clusters by the Cheeger inequality. It also implies a spectral gap in that $\lambda_L/\lambda_{L+1} \to 0$.[7] By Theorem 1.1 of Peng et al. (2017), the span of the L eigenvectors of the Laplacian (corresponding to the smallest L eigenvalues) is close to that of L vectors of normalized indicators that identify the infeasible optimal partition that minimizes conductance. Consequently, the output of k-means should be close to the optimum.

5 Spectra of Geometric and Random Graphs

By Theorem 2, (3) is necessary for cluster-robust inference to be valid. As discussed in §2.1, a necessary condition for (3) is that the network must be well clustered, meaning for some L, $h_L(\boldsymbol{A}) \to 0$, or equivalently, $\lambda_L \to 0$ by Proposition 1. This section shows that, unfortunately, not all networks are well clustered, which motivates the recommendations in §2.3. We first survey results from geometry and random graph theory on the spectra and Cheeger constants of various graphs. We then provide

[7] In §5.2, we apply spectral clustering to networks with small spectral gaps that are nonetheless well clustered, yet find the algorithm still delivers low-conductance clusters. Hence, having a large spectral gap is sufficient but apparently not necessary.

simulation evidence supporting the theory and clarifying aspects that are, to our knowledge, incomplete.

5.1 Theoretical Results

The first two examples are of well-clustered graphs.

Example 1. A planar graph is a graph that can be drawn on the plane such that links do not cross. Kelner et al. (2011) show that planar graphs with uniformly bounded degree satisfy $\lambda_L = O(L/n)$. More generally, they show that for graphs embedded in orientable surfaces of genus g, the same result holds if g does not depend on L or n.

Example 2. Random geometric graphs are defined by associating each unit i with a position $X_i \in \mathbb{R}^d$, i.i.d. across units with density f, and setting $A_{ij} = \mathbf{1}\{\|X_i - X_j\| \leqslant r_n\}$ for some $r_n > 0$. For the graph to be sparse, r_n must tend to zero. Several papers characterize the limiting behavior of Cheeger constants. Müller and Penrose (2020) show that the second-order Cheeger constant is $o(1)$ a.s. when $r_n \to 0$ and $nr_n^d \gg \log n$.[8] Trillos et al. (2016) (Theorem 12) provide similar results for kth-order Cheeger constants, albeit defined slightly differently than ours.

A perhaps more realistic model is the random connections model

$$A_{ij} = \mathbf{1}\{\alpha_i + \alpha_j + r_n^{-1}\|X_i - X_j\| > \varepsilon_{ij}\}, \tag{9}$$

which allows units further than distance r_n to form links, albeit with probability vanishing with distance $\|X_i - X_j\|$. Leung and Moon (2020) study generalizations of this model with strategic interactions. To our knowledge, there are no available results on the spectrum, but we provide simulation evidence below showing that this graph appears to be well clustered.

We next discuss examples of graphs that are not well clustered.

Example 3. A k-regular graph is one such that $\sum_j A_{ij} = k$ for all i. Bollobás

[8] This is their Theorem 2.1 for $v = 2$, $b = 1$. Their notion of conductance corresponds to our definition multiplied by the link count $\sum_{i,j} A_{ij}$, what they label $\mathrm{Vol}_{n,2}(\mathcal{X}_n)$. By their equation (2.12), that term is of exact order $n^2 r_n^d$. Furthermore, the limit in (2.11) is finite. Hence, the normalization in their Theorem 2.1 implies the Cheeger constant is $O(r_n) = o(1)$ a.s.

(1988) proves that for $k \geqslant 3$, the isoperimetric number of a randomly drawn k-regular graph is at least a certain positive constant with probability approaching one. The isoperimetric number for k-regular graphs equals $kh_2(\boldsymbol{A})$, the 2nd-order Cheeger constant times k, so the latter is bounded away from zero.

Expander graphs (or rather sequences of them) are those, which, by construction, have Cheeger constants uniformly bounded away from zero, yet may still be sparse (Hoory et al., 2006).

Expander and k-regular graphs are extremely stylized models. The final examples concern more realistic models that have been applied to real-world networks.

Example 4. Inhomogeneous random graphs (Bollobás et al., 2007) satisfy

$$\mathbf{P}(A_{ij} = 1 \mid \alpha_i, \alpha_j) = \frac{\kappa(\alpha_i, \alpha_j)}{n},$$

where the types α_i are usually independent and $\kappa(\alpha_i, \alpha_j)$ is often assumed to have bounded support. Stochastic block models correspond to the special case in which types are finitely supported. These are widely used in the statistics literature for studying community detection.

One can easily choose $\kappa(\cdot)$ to generate homophily in types, where units with similar types have a higher probability of linking. Given this type of structure, it may seem that these graphs can be well clustered under reasonable conditions, say with clusters roughly corresponding to sets of units with the same type. However, Hoffman et al. (2019) write that a body of literature studying Erdős-Rényi graphs (a special case with $\kappa(\cdot)$ constant) "show that the giant component can be partitioned into a well connected expanding core together with small (logarithmic size) graphs attached to the core," where the core is a subgraph of the giant that is order n in size.[9] This suggests that even if λ_L were small for these graphs, any low-conductance partition would be extremely unbalanced, so the graph is close to poorly clustered. Our simulations below support this.

[9]See e.g. Coja-Oghlan (2007) (Theorem 1.2) for formal results for Erdős-Rényi graphs and Zhang and Rohe (2018) for related results for stochastic block models.

5.2 Simulation Evidence

We simulate the random geometric graph (RGG) (Example 2), random connections model (RCM) (9), Erdős-Rényi graph (ER), and stochastic block model (SBM) (Example 4) for $n = 1000$ units. We calibrate parameters to obtain an average degree of about 5 for all graphs. For the RGG, $\{X_i\} \overset{iid}{\sim} \mathcal{U}([0,1]^2)$, and $r_n = (5/(\pi n))^{1/2}$. For the RCM, $\{\alpha_i\} \overset{iid}{\sim} \mathcal{U}([0,1])$, $\{\varepsilon_{ij}\} \overset{iid}{\sim}$ logistic, $\{X_i\} \perp\!\!\!\perp \{\alpha_i\} \perp\!\!\!\perp \{\varepsilon_{ij}\}$, and $r_n = (5/(3.5\pi n))^{1/2}$. For ER, $\rho(\alpha_i, \alpha_j) = \bar{\rho} = 5$. Finally, for the SBM, we construct 10 blocks of 100 units each, where units in the same block have probability $10/n$ of linking and units in different blocks have probability $(5 \cdot 8/9)/n$ of linking.

We analyze the spectrum of and apply spectral clustering to the subnetwork on the giant component. Figure 2 plots histograms and scatterplots of the spectra for a typical draw from each model. We see that both the RGG and RCM have a sizeable mass of eigenvalues near zero but no obvious spectral gap. In contrast, both ER and the SBM have only one zero eigenvalue (λ_1 is necessarily zero) and a large gap between λ_1 and λ_2. Consequently, only the RGG and RCM appear to be well clustered.

The scatterplots also illustrate the potential pitfalls with using only either the cutoff heuristic or the gap heuristic described in §2.3. The gap heuristic identifies one clear cluster for ER and the SBM, whereas for the RGG and RCM, the location of the largest gap looks essentially random given how continuous the spectrum is. For the cutoff heuristic, using this alone could potentially pick out a rather large number of clusters for the RGG and RCM, given the mass of eigenvalues near zero, whereas using our combination of heuristics (7) could improve inference by producing lower-conductance clusters at the relatively small cost of having fewer of them.

We next present results on the conductances of clusters generated from spectral clustering, choosing L according to (7) with $c = 0.05$. If $L_2 < 5$, meaning the gap heuristic finds very few clusters, this gives ER and SBM a chance to find more clusters by taking the larger of the two potential values of L. Table 1 displays the result of 10k simulations. The first column of the table is the maximal conductance, the second the number of clusters, the third the size of the spectral gap, the fourth the Lth smallest eigenvalue, the fifth the median cluster size, the sixth the size of the giant component, and the last the average degree. We find that the heuristic (7) produces a larger number of clusters for the RGG and a moderate number for the RCM, both with conductances around 0.15. In contrast, for ER and the SBM, only one cluster is typically found on average, and this cluster comprises nearly all units in the network.

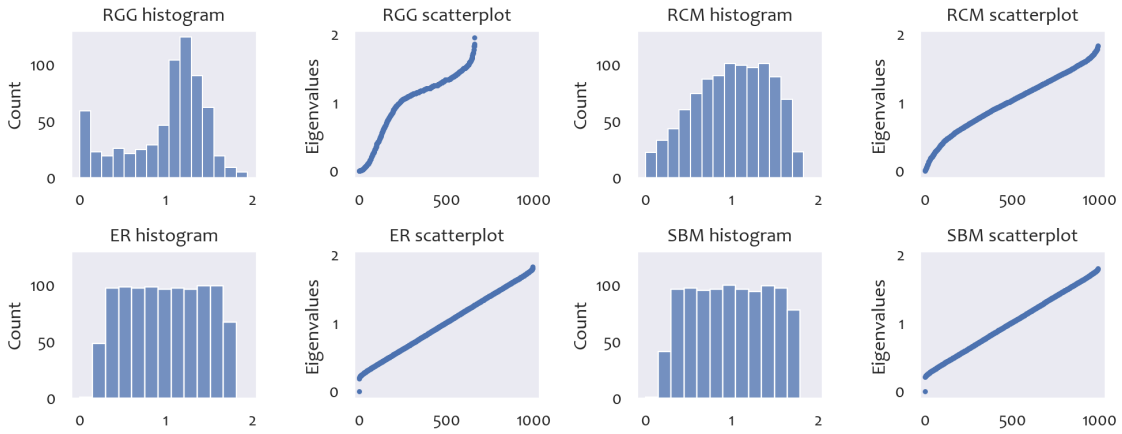

Figure 2: Histograms and scatterplots of eigenvalues.

We obtain similar results for $c = 0.02$ and $c = 0.1$ (not reported in a table). For 0.02, the RGG (RCM) has on average 24 (6) clusters with maximal conductance 0.085 (0.071). For 0.1, the average number of clusters for ER and the SBM is still only 1.01. Figure 1 plots an RGG and a draw from the RCM with clusters obtained using a cutoff of 0.02. In the figure, units are plotted according to their position in $[0, 1]^2$ and colored according to their clusters obtained from spectral clustering. The RCM contains longer-range links, producing a denser-looking figure, so spectral clustering generates fewer clusters compared to the RGG.

Table 1: Spectra and Clusters

	$\max_S \phi(S)$	# Clus.	Gap	λ_L	Med. Clus.	Giant	Degree
RGG	0.158	40.0	0.004	0.048	18.3	758.4	4.83
RCM	0.137	12.7	0.006	0.047	77.3	983.9	4.98
ER	0.002	1.0	0.180	0.000	990.7	993.1	5.00
SBM	0.003	1.0	0.180	0.000	990.5	993.0	5.00

n = 1k. Averages over 10k simulations. "Gap" = size of spectral gap, "Med. Clus." = median cluster size, "Giant" = size of giant component, "Degree" = average degree.

6 Numerical Illustrations

6.1 Monte Carlo

We present simulation results on the finite-sample properties of the randomization test for clusters computed using spectral clustering and t-test using a HAC estimator.

Design 1. We simulate the RGG, RCM, and SBM using the same parameters as the design in §5.2. We then draw $\{\varepsilon_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0,1)$ independently of the network and define $W_i = \varepsilon_i + \sum_j A_{ij}\varepsilon_j / \sum_j A_{ij}$, which generates a simple form of network dependence. We let $\theta_0 = \mathbf{E}[W_i] = 0$ and $g(W_i, \theta_0) = W_i$, so the goal is inference on the mean of W_i, whose true value is zero. This design is deliberately simplistic to show that, even here, cluster-robust methods can break down for poorly clustered networks.

We use the randomization test (1) to test the null that $\theta_0 = 0$ at the 5 percent level. For each simulation draw, we compute $L = 8$ clusters in the giant component. We treat all other components as individual clusters and discard all clusters with size less than 20. We choose $L = 8$ because, based on the results in §5.2, we expect clusters of the RGG to have low conductance, so the randomization test should perform well. Clusters of the RCM will have higher conductance, and it is unclear whether this will translate to substantial size distortion. Finally, clusters of the SBM should have exceedingly high conductance, so we expect the test to perform poorly.

We report rejection rates for the randomization test and two different t-tests. One uses the leading alternative to cluster-robust inference, which is a HAC variance estimator. We use a uniform kernel with the bandwidth chosen according to the rule in Leung (2020), equation (12). The other t-test uses i.i.d. standard errors, which serves to quantify the degree of dependence in the data.

Table 2 reports the results of 10k simulations. We see that the randomization test control size well for the RGG, outperforming the HAC estimator in smaller samples. This is a result of the low maximal conductance of the clusters. More surprising is that the test has good performance for the RCM, despite the conductance being as high as 0.22, with the test again outperforming the HAC estimator in smaller samples. Finally, for the SBM, we see that the randomization test exhibits substantial size distortion due to the high maximal conductance of the clusters, around 0.5. Here the HAC estimator outperforms for all sample sizes.

Table 2: Rejection Rates for Design 1

	RGG			RCM			SBM		
n	250	500	1000	250	500	1000	250	500	1000
Rand	0.052	0.049	0.051	0.058	0.057	0.052	0.098	0.098	0.101
HAC	0.070	0.061	0.058	0.076	0.065	0.056	0.082	0.066	0.057
IID	0.274	0.272	0.279	0.275	0.276	0.282	0.289	0.288	0.287
# Clusters	8.886	9.483	10.275	7.994	8.000	8.000	7.814	7.966	7.999
$\max_S \phi(S)$	0.104	0.046	0.028	0.219	0.141	0.094	0.521	0.516	0.523
1st Clus.	62.3	110.2	197.2	55.0	110.0	222.7	56.0	117.9	240.9
2nd Clus.	40.3	75.4	143.1	44.2	87.1	171.7	41.7	84.1	168.7
Last Clus.	23.0	27.9	45.1	24.3	40.0	69.5	24.2	40.9	78.9

Averages over 10k simulations. The first three rows give rejection rates for level-5% tests. The last three rows are the sizes of the indicated clusters in descending order of size. The number of clusters may be less than the target of 8 because we discard clusters of size less than 20.

Design 2. The next design considers the more realistic problem of estimating network spillovers. We exactly replicate the designs in §5.2 of Leung (2020), which involve two outcome models: a linear-in-means model and a binary game on a network. For the former, $Y_i = V_i(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\varepsilon})$, and for the latter, $Y_i = \mathbf{1}\{V_i(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\varepsilon}) > 0\}$, where

$$V_i(\boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\varepsilon}) = \alpha + \beta \frac{\sum_j A_{ij} Y_j}{\sum_j A_{ij}} + \delta \frac{\sum_j A_{ij} D_j}{\sum_j A_{ij}} + D_i \gamma + \varepsilon_i,$$

where Y_i is unit i's outcome, ε_i a structural error, and D_i is i's binary treatment assignment which is i.i.d. and independent of all other primitives. For details on parameters and distributions of primitives, see Leung (2020). For both models, we estimate a spillover effect using the inverse-probability weighting estimator

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} Y_i \left(\frac{T_i}{\mathbf{P}(T_i = 1)} - \frac{1 - T_i}{\mathbf{P}(T_i = 0)} \right), \quad T_i = \mathbf{1}\Big\{ \max_j A_{ij} D_j > 0 \Big\}.$$

That is, T_i is an indicator for having a treated neighbor. We simulate the outcome models on three different networks. Two follow the ones used in Leung (2020), the configuration model and RGG, which are calibrated to the data on school friendship networks in his empirical application. The configuration model generates a network approximately uniformly at random from the set of all networks with a given degree sequence; that sequence is chosen to be the empirical degree sequence of the network

in his application, which has average degree of about 8. This plays the role of the SBM in Table 2 as the network that turns out to lack low-conductance clusters.

We additionally simulate the RCM using the same design as Table 2. Like the RGG, the average degree is calibrated to the empirical application by setting $r_n = (\kappa/(3\pi n))^{1/2}$, where κ is the average degree in the data. Additionally, as with the RGG design, we set the error terms ε_i in both outcome models equal to $\nu_i + (\rho_{i1} - 0.5)$, where ρ_{i1} is the first component of i's position ρ_i in the linear-in-means model, and ν_i is a normal error term; this generates unobserved homophily in the network. For additional details on the design, see Leung (2020).

Table 3 reports the results of 5k simulations, choosing L as in design 1. Each network formation model is associated with three columns with n corresponding to the population size in the largest, two largest, and four largest schools in the empirical application of Leung (2020). We find that the randomization test performs poorly for the configuration model, which produces poorly clustered networks, but controls size well for the other networks, which are well clustered. The test also exhibits some size distortion for the RCM when conductance exceeds 0.1.

Table 3: Rejection Rates for Design 2

	RGG			RCM			Configuration		
n	365	716	1408	365	722	1427	350	692	1375
LIM Rand	0.053	0.056	0.051	0.066	0.063	0.054	0.251	0.254	0.252
LIM HAC	0.066	0.071	0.063	0.078	0.069	0.058	0.076	0.065	0.063
BG Rand	0.049	0.052	0.049	0.063	0.048	0.056	0.161	0.167	0.163
BG HAC	0.066	0.062	0.059	0.069	0.055	0.054	0.075	0.067	0.058
$\max_S \phi(S)$	0.052	0.037	0.027	0.165	0.119	0.084	0.605	0.604	0.608
# Clusters	7.87	7.95	8.01	7.84	7.93	7.99	8.00	8.00	8.00
1st Clus.	175.5	322.0	616.3	178.1	333.0	639.1	197.0	358.5	671.2
2nd Clus.	141.1	252.8	466.8	134.7	243.3	456.7	119.2	210.4	390.6
Last Clus.	61.0	113.7	213.0	68.8	128.4	243.8	79.8	147.8	21.0

Averages over 5k simulations. The first four rows give rejection rates for a level-5% test, with LIM = linear-in-means, BG = binary game, Rand = randomization test, HAC = t-test with HAC estimator. The last three rows are the sizes of the indicated clusters in descending order of size. The number of clusters may be less than the target of 8 because we discard clusters of size less than 20.

6.2 Empirical Applications

Aral and Nicolaides (2017). This paper finds evidence of peer effects in exercise activity using data from an online social network of 1.1 million runners. The authors partition their network into 15144 clusters (average size 7.7, SD 41) using a modularity-based method and use clustered standard errors. Their network data is not publicly available. However, on p. 35 of the supplementary appendix, they write, "on average 8 out of 10 friends are within cluster while 2 of 10 are across clusters." This is reported based on their belief that this measures independence of clusters.

Our results provide some formal justification for this belief. Their statistic is similar to conductance – it is an average over a measure of conductance defined at the unit level – so we can ballpark $\max_{1 \leqslant \ell \leqslant L} \phi_{\boldsymbol{A}}(\mathcal{C}_\ell)$ at around 0.2 in their application. Our simulations indicate that cluster-robust methods would perform better if this were closer to 0.1. It is likely that the authors could halve the number of clusters, which still leaves a very sizeable number, for a substantial decrease in conductance.

Note that the large number of clusters is not because their network has many small components (in which case finding clusters is trivial). They report that the giant component of their network contains 90 percent of units. A possible explanation for the low conductance of their clusters is that the online social network they study spans many countries, with US users comprising 20 percent of the data. It is likely that link formation is strongly geographically determined, and our discussion in §5 indicates that spatial graphs typically have low conductance. Nonetheless, it is not *a priori* clear how to construct clusters since it is possible that many pairs of users across states or countries are linked in the network. Then it is not obvious whether, say, using states as clusters may produce low-conductance clusters. This illustrates the usefulness of community detection algorithms in providing a more principled way of constructing clusters based on minimizing conductance.

Zacchia (2020). This paper studies knowledge spillovers across firms. The author constructs a weighted, undirected network of 707 firms for each year t. The weighted link $A_{ij,t}$ between firms i and j at time t measures co-patenting between firm inventors. In order to apply a community detection algorithm, which requires a static network, the author sums the networks across time, defining $A_{ij} = \sum_t A_{ij,t}$ for each i, j.

To compute the clusters, Zacchia (2020) applies a variant of the Louvain algorithm to the giant component, which is a modularity-based method (Blondel et al., 2008).

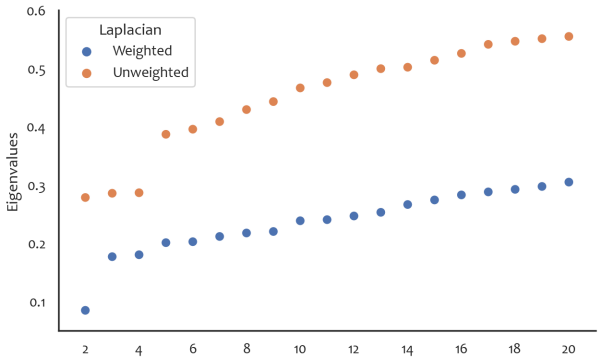

Figure 3: Scatterplot of eigenvalues $\lambda_2 \leqslant \ldots \leqslant \lambda_{20}$.

The variant adds a tuning parameter φ that can be increased to obtain more clusters. The choice of $\varphi = 0.6$ used in the paper yields 20 clusters in the giant. The author treats all units outside of the giant as a single cluster and clusters standard errors.

Our theoretical results in §3 only pertain to unweighted graphs. However, the graph invariants in §2.2 are all defined for weighted graphs, as discussed in Remark 1. We compute these quantities both for the original weighted network and the unweighted version where A_{ij} is set to 1 if and only if the weight is positive. In the unweighted graph, there are 3451 links, so the network is sparse. The analysis that follows focuses on the giant component, which consists of 439 units.

Figure 3 plots the spectra of the Laplacians for both the weighted and unweighted networks, starting at λ_2 (since $\lambda_1 = 0$). The networks have spectral gaps at 2 and 4. However, the former has only one eigenvalue below 0.1, while the latter has none. This indicates that the networks appear to be poorly clustered.

This is further confirmed in Figure 4. The first two columns plot the conductances and sizes of each cluster used in Zacchia (2020) for the weighted network, and the remaining columns plot the same quantities obtained from spectral clustering, for different values of L. The corresponding figure for the unweighted network is essentially the same and therefore omitted. The figure shows that both the Louvain algorithm for $\varphi = 0.6$ and the spectral clustering algorithm for $L = 20$ yield clusters all with high conductance. Choosing smaller values of L for spectral clustering does not appear to improve matters, as only a single cluster with low conductance emerges, but this contains the vast majority of units in the network. Thus, the partition is highly unbalanced, which, as discussed in §2.3, means the power of the test would be little

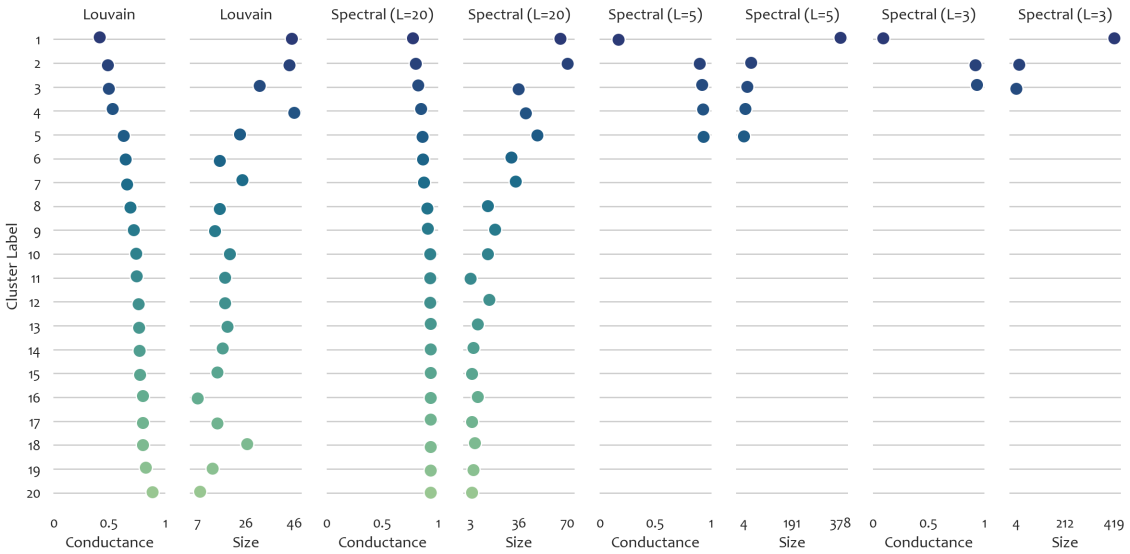

Figure 4: Conductances and cluster sizes for the weighted network.

better than one with one cluster. This network apparently reflects the problem discussed in Example 4 that inhomogeneous random graphs may consist of a large "core" that is not well clustered and small attachments, resulting in an unbalanced partition. Thus, for this dataset, it may be preferable to use an alternative to cluster-robust inference methods.

7 Conclusion

This paper studies the practice of partitioning a network, either manually or using an unsupervised learning algorithm, in order to apply cluster-robust inference methods. We isolate a key condition that, under some assumptions, is necessary and sufficient for this practice to be valid: the clusters must all have low conductance, that is, low boundary-to-volume ratios. We call graphs "well clustered" if a partition with this property exists and provide theoretical and simulation evidence showing that important classes of graphs are not well clustered. Our simulation study shows that cluster-robust inference methods applied to such graphs can exhibit severe size distortion. For graphs that are well clustered, however, they outperform HAC estimators in terms of size control. Our results on conductance connect the literature on cluster-robust inference to spectral clustering, allowing us to use tools from the latter

to construct clusters in a more principled way, namely to minimize conductance.

We provide three recommendations for empirical practice. First, for any candidate set of clusters, one should compute the maximal conductance to assess its quality and aim for a value below 0.1. Second, given a network, one should first compute the spectrum of the Laplacian to assess whether the network is well clustered. Third, given a well-clustered network, one can compute candidate clusters using spectral clustering or any number of community detection algorithms. We note that some networks may be well clustered but only have unbalanced partitions consisting of one large cluster and several small clusters. In this case, cluster-robust methods have poor power and an alternative may be preferred, such as a HAC variance estimator.

A Appendix

A.1 Proofs

Proof of Proposition 2. Let $\mathcal{S}$ be the subset of clusters $\mathcal{C}_\ell$ for which $\rho_\ell > 0$. Then $n^{-1/2}(n_\ell \hat{G}_\ell(\theta_0))_{\ell \in \mathcal{S}}$ is asymptotically normal with the desired limit variance by the CLT of Kojevnikov et al. (2019) (Theorem 3.2) and Cramér-Wold device. Note that to verify their version of Assumption 2(b) (their Assumption 3.4), we need to divide the quantities in (8) by powers of the standard deviation $\mathrm{Var}(\sum_{\ell \in \mathcal{S}} c_\ell n^{-1/2} n_\ell \hat{G}_\ell(\theta_0))^{1/2}$, where $(c_\ell)_{\ell \in \mathcal{S}}$ is any nonzero vector (for the Cramér-Wold device). However, the standard deviation has a strictly positive limit by Assumption 3, hence why we ignore it our formulation of Assumption 2(b). Finally, for all ℓ such that $\rho_\ell = 0$, their CLT implies $n^{-1/2} n_\ell \hat{G}_\ell(\theta_0) \xrightarrow{p} \mathbf{0}$, so we can extend joint convergence for only clusters in $\mathcal{S}$ to joint convergence for the full vector, as in (2). ∎

Proof of Theorem 1. We show $n^{-1}\mathrm{Cov}(n_\ell \hat{G}_\ell(\theta_0), n_m \hat{G}_m(\theta_0)) = o(1)$ for any $\ell \neq m$. Let $\|\cdot\|$ be the matrix sup norm. The covariance is bounded in norm by

$$\frac{1}{n} \sum_{i \in \mathcal{C}_\ell} \sum_{j \in \mathcal{C}_m} \|\mathbf{E}[g(W_i, \theta_0) g(W_j, \theta_0)']\| \leqslant C' \sum_{s=1}^{n} \psi_n(s) \frac{1}{n} \sum_{i \in \mathcal{C}_\ell} \sum_{j \in \mathcal{C}_m} \mathbf{1}\{\ell_{\boldsymbol{A}}(i,j) = s\} \tag{A.1}$$

for some constant $C' > 0$ by Assumption 2 (take f, f' to be the identity function). The sum over s terminates at n because there are only n units in the network, and disconnected units are uncorrelated under Assumption 2(a).

We next bound the term $n^{-1}\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m}\mathbf{1}\{\ell_{\boldsymbol{A}}(i,j)=s\}$ in (A.1). Let

$$\mathcal{B}_{\boldsymbol{A}}(\mathcal{C}_\ell)=\big\{i\in\mathcal{C}_\ell\colon \max_{j\in\mathcal{N}_n\setminus\mathcal{C}_\ell}A_{ij}=1\big\},$$

the boundary of $\mathcal{C}_\ell$. Given $i\in\mathcal{C}_\ell$ and $j\in\mathcal{C}_m$ with $\ell_{\boldsymbol{A}}(i,j)=s$, there must exist some $k\in\mathcal{B}(\mathcal{C}_\ell)$ such that $\ell_{\boldsymbol{A}}(i,k)=d$ and $\ell_{\boldsymbol{A}}(k,j)=d'$, for some d,d' satisfying $d+d'+1=s$. Hence,

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m}\mathbf{1}\{\ell_{\boldsymbol{A}}(i,j)=s\}\\
&\quad\leqslant\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{d=0}^{s-1}\sum_{k=1}^{n}\mathbf{1}\{\ell_{\boldsymbol{A}}(i,k)=d\}\mathbf{1}\{k\in\mathcal{B}_{\boldsymbol{A}}(\mathcal{C}_\ell)\}\mathbf{1}\{\ell_{\boldsymbol{A}}(k,j)=s-1-d\}\\
&\quad\leqslant\sum_{d=0}^{s-1}\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}\{k\in\mathcal{B}_{\boldsymbol{A}}(\mathcal{C}_\ell)\}|\mathcal{N}_{\boldsymbol{A}}(k,d)|\,|\mathcal{N}_{\boldsymbol{A}}(k,s-1-d)|\\
&\leqslant\sum_{d=0}^{s-1}\left(\frac{1}{n}\sum_{k=1}^{n}|\mathcal{N}_{\boldsymbol{A}}(k,d)|^{1+\epsilon}|\mathcal{N}_{\boldsymbol{A}}(k,s-1-d)|^{1+\epsilon}\right)^{1/(1+\epsilon)}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}\{k\in\mathcal{B}_{\boldsymbol{A}}(\mathcal{C}_\ell)\}\right)^{\epsilon/(1+\epsilon)}
\end{aligned}\tag{A.2}$$

for any $\epsilon>0$ by Hölder's inequality. Since

$$\frac{1}{n}\sum_{k=1}^{n}|\mathcal{N}_{\boldsymbol{A}}(k,d)|^{1+\epsilon}|\mathcal{N}_{\boldsymbol{A}}(k,s-1-d)|^{1+\epsilon}\leqslant\frac{1}{n}\sum_{k=1}^{n}|\mathcal{N}_{\boldsymbol{A}}(k,s)|^{2(1+\epsilon)},$$

we have

$$\text{(A.2)}\leqslant s\underbrace{\left(\frac{1}{n}\sum_{k=1}^{n}|\mathcal{N}_{\boldsymbol{A}}(k,s)|^{2(1+\epsilon)}\right)}_{M_n(s,2(1+\epsilon))}^{1/(1+\epsilon)}\left(\underbrace{\frac{|\mathcal{B}_{\boldsymbol{A}}(\mathcal{C}_\ell)|}{\mathrm{vol}_{\boldsymbol{A}}(\mathcal{C}_\ell)}}_{\leqslant\phi_{\boldsymbol{A}}(\mathcal{C}_\ell)}\underbrace{\frac{\mathrm{vol}_{\boldsymbol{A}}(\mathcal{C}_\ell)}{n}}_{\leqslant\delta(\boldsymbol{A})}\right)^{\epsilon/(1+\epsilon)},$$

noting that $|\mathcal{B}_{\boldsymbol{A}}(\mathcal{C}_\ell)|\leqslant|\partial_{\boldsymbol{A}}(\mathcal{C}_\ell)|$. Therefore,

$$\text{(A.1)}\leqslant C'\sum_{s=1}^{n}sM_n(s,2(1+\epsilon))^{1/(1+\epsilon)}\psi_n(s)\left(\max_{1\leqslant\ell\leqslant L}\phi_{\boldsymbol{A}}(\mathcal{C}_\ell)\cdot\delta(\boldsymbol{A})\right)^{\epsilon/(1+\epsilon)}.$$

Choosing ϵ according to Assumption 5, this is $o(1)$ by Assumptions 4 and 5. ∎

PROOF OF THEOREM 2. The assumptions imply $\liminf_{n\to\infty} \max_{1\leqslant\ell\leqslant L} \phi_{\boldsymbol{A}}(\mathcal{C}_\ell) > 0$. Then, for some ℓ, the following has positive limit infimum:

$$\frac{\sum_{i\in\mathcal{C}_\ell}\sum_{j\notin\mathcal{C}_\ell} A_{ij}}{\mathrm{vol}_{\boldsymbol{A}}(\mathcal{C}_\ell)} = \sum_{m\neq\ell} \frac{\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m} A_{ij}}{\mathrm{vol}_{\boldsymbol{A}}(\mathcal{C}_\ell)}.$$

This implies that for some cluster $m \neq \ell$,

$$\liminf_{n\to\infty} \frac{\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m} A_{ij}}{\mathrm{vol}_{\boldsymbol{A}}(\mathcal{C}_\ell)} > 0. \tag{A.3}$$

Consider a data-generating process such that for some universal constants $\gamma > 0$ and $\gamma' \geqslant 0$ and all n, $\mathbf{E}[g(W_i,\theta_0)g(W_j,\theta_0)] = \gamma A_{ij} + \gamma'(1 - A_{ij})$. Then for ℓ, m satisfying (A.3),

$$\begin{aligned}|n^{-1}\mathrm{Cov}(n_\ell\hat{G}_\ell(\theta_0), n_m\hat{G}_m(\theta_0))| &= \left|\frac{1}{n}\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m} \mathbf{E}[g(W_i,\theta_0)g(W_j,\theta_0)]\right| \\ &\geqslant \gamma\,\frac{1}{n}\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m} A_{ij} = \gamma\,\frac{\sum_{i\in\mathcal{C}_\ell}\sum_{j\in\mathcal{C}_m} A_{ij}}{\mathrm{vol}_{\boldsymbol{A}}(\mathcal{C}_\ell)}\frac{n_\ell}{n}\frac{1}{n_\ell}\sum_{i\in\mathcal{C}_\ell}\sum_{j=1}^{n} A_{ij}.\end{aligned}$$

By assumption, $n_\ell/n \to \rho_\ell > 0$. Furthermore, $n_\ell^{-1}\sum_{i\in\mathcal{C}_\ell}\sum_{j=1}^{n} A_{ij}$ is the average degree of units in cluster ℓ, so since clusters are connected subnetworks, this is always at least one. Therefore, the right-hand side of the above display is asymptotically bounded away from zero. ∎

A.2 Verifying Weak Network Dependence

Leung (2020) shows that $\psi_n(s)$ is uniformly $O(e^{-cs})$ for some $c > 0$ in certain models of social interactions. His Appendix A verifies Assumption 2(b) under the assumption that $\psi_n(s) = e^{-cs}$ for graphs with polynomial and exponential neighborhood growth rates, meaning

$$\max_{i\in\mathcal{N}_n}|\mathcal{N}_{\boldsymbol{A}}(i,s)| = Cs^d \quad \text{and} \quad \max_{i\in\mathcal{N}_n}|\mathcal{N}_{\boldsymbol{A}}(i,s)| = Ce^{\beta s},$$

respectively, for $C, d, \beta > 0$. In the polynomial case, no additional conditions are needed. In the exponential case, we need $c > 3\beta$, meaning that $\psi_n(s)$ decays suffi-

ciently fast enough relative to the rate at which neighborhood sizes expand.

We verify Assumption 5 under this setup. In the polynomial case,

$$\sum_{s=1}^{n} sM_n(s, 2(1+\epsilon))^{1/(1+\epsilon)}\psi_n(s) \leqslant C^2 \sum_{s=1}^{n} s^{2d+1}e^{-c\,s} = O(1).$$

In the exponential case, since $c > 3\beta$,

$$\sum_{s=1}^{n} sM_n(s, 2(1+\epsilon))^{1/(1+\epsilon)}\psi_n(s) \leqslant C^2 \sum_{s=1}^{n} e^{(2\beta-c)s} = O(1).$$

Note that it is enough to have $c > 2\beta$, so at least for these classes of graphs, Assumption 5 is weaker than Assumption 2(b).

References

Aral, S. and C. Nicolaides, "Exercise Contagion in a Global Social Network," *Nature Communications*, 2017, *8* (1), 1–8.

_ **and M. Zhao**, "Social Media Sharing and Online News Consumption," *SSRN working paper No. 3328864*, 2019.

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. Jackson, "The Diffusion of Microfinance," *Science*, 2013, *341* (6144).

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

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

Blondel, V., J. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2008, *2008* (10), P10008.

Bollobás, B., "The Isoperimetric Number of Random Regular Graphs," *European Journal of Combinatorics*, 1988, *9* (3), 241–244.

Bollobás, B., S. Janson, and O. Riordan, "The Phase Transition in Inhomogeneous Random Graphs," *Random Structures and Algorithms*, 2007, *31* (1), 3–122.

Cai, Y., I. Canay, D. Kim, and A. Shaikh, "A User's Guide to Approximate Randomization Tests with a Small Number of Clusters," *Northwestern University working paper*, 2021.

Cameron, A. 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., A. Santos, and A. Shaikh, "The Wild Bootstrap with a "Small" Number of "Large" Clusters," *Review of Economics and Statistics (forthcoming)*, 2020.

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

Chung, F., *Spectral Graph Theory* number 92, American Mathematical Soc., 1997.

_ , "Laplacians and the Cheeger Inequality for Directed Graphs," *Annals of Combinatorics*, 2005, *9* (1), 1–19.

Coja-Oghlan, A., "On the Laplacian Eigenvalues of $G(n, p)$," *Combinatorics, Probability & Computing*, 2007, *16* (6), 923.

Conley, T., S. Gonçalves, and C. Hansen, "Inference with Dependent Data in Accounting and Finance Applications," *Journal of Accounting Research*, 2018, *56* (4), 1139–1203.

Hansen, B. and S. Lee, "Asymptotic Theory for Clustered Samples," *Journal of Econometrics*, 2019, *210* (2), 268–290.

Hoffman, C., M. Kahle, and E. Paquette, "Spectral Gaps of Random Graphs and Applications," *International Mathematics Research Notices*, 2019.

Hoory, S., N. Linial, and A. Wigderson, "Expander Graphs and Their Applications," *Bulletin of the American Mathematical Society*, 2006, *43* (4), 439–561.

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

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

Jochmans, K. and M. Weidner, "Fixed-Effect Regressions on Network Data," *Econometrica*, 2019, *87* (5), 1543–1560.

Kelner, J., J. Lee, G. Price, and S. Teng, "Metric Uniformization and Spectral Bounds for Graphs," *Geometric and Functional Analysis*, 2011, *21* (5), 1117.

Kojevnikov, D., "The Bootstrap for Network Dependent Processes," *University of British Columbia working paper*, 2021.

_ **, V. Marmer, and K. Song**, "Limit Theorems for Network Dependent Random Variables," *arXiv preprint arXiv:1903.01059*, 2019.

_ **,** _ **, and** _ , "Limit Theorems for Network Dependent Random Variables," *Journal of Econometrics (forthcoming)*, 2020.

Lee, J., S. Gharan, and L. Trevisan, "Multiway Spectral Partitioning and Higher-Order Cheeger Inequalities," *Journal of the ACM*, 2014, *61* (6), 1–30.

Lei, J. and A. Rinaldo, "Consistency of Spectral Clustering in Stochastic Block Models," *Annals of Statistics*, 2015, *43* (1), 215–237.

Leung, M., "Causal Inference Under Approximate Neighborhood Interference," *arXiv preprint arXiv:1911.07085*, 2020.

_ **and R. Moon**, "Normal Approximation in Large Network Models," *arXiv preprint arXiv:1904.11060*, 2020.

Miguel, E. and M. Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72* (1), 159–217.

Müller, T. and M. Penrose, "Optimal Cheeger Cuts and Bisections of Random Geometric Graphs," *Annals of Applied Probability*, 2020, *30* (3), 1458–1483.

Müller, U. and M. Watson, "Spatial Correlation Robust Inference," *arXiv preprint arXiv:2102.09353*, 2021.

Peng, R., H. Sun, and L. Zanetti, "Partitioning Well-Clustered Graphs: Spectral Clustering Works!," *SIAM Journal on Computing*, 2017, *46* (2), 710–743.

Rohe, K., S. Chatterjee, and B. Yu, "Spectral Clustering and the High-Dimensional Stochastic Blockmodel," *Annals of Statistics*, 2011, *39* (4), 1878–1915.

Šíma, J. and Satu E. Schaeffer, "On the NP-Completeness of some Graph Cluster Measures," in "International Conference on Current Trends in Theory and Practice of Computer Science" Springer 2006, pp. 530–537.

Trevisan, L., "Lecture Notes on Graph Partitioning, Expanders and Spectral Methods," 2016. URL: https://people.eecs.berkeley.edu/~luca/books/expanders-2016.pdf. Last visited on 2020/12/05.

Trillos, N., D. Slepčev, J. Von Brecht, T. Laurent, and X. Bresson, "Consistency of Cheeger and Ratio Graph Cuts," *Journal of Machine Learning Research*, 2016, *17* (1), 6268–6313.

Ugander, J., B. Karrer, L. Backstrom, and C. Marlow, "The Anatomy of the Facebook Social Graph," *arXiv preprint arXiv:1111.4503*, 2011.

von Luxburg, U., "A Tutorial on Spectral Clustering," *Statistics and Computing*, 2007, *17* (4), 395–416.

Zacchia, P., "Knowledge Spillovers Through Networks of Scientists," *Review of Economic Studies*, 2020, *87* (4), 1989–2018.

Zhang, Y. and K. Rohe, "Understanding Regularized Spectral Clustering via Graph Conductance," in "Advances in Neural Information Processing Systems" 2018, pp. 10631–10640.