Community Detection in Partial Correlation Network Models

Christian Brownlees[†] Guðmundur Stefán Guðmundsson[†] Gábor Lugosi[†,◇]

June 2, 2016

Abstract

Many real-world networks exhibit a community structure: The vertices of the network are partitioned into groups such that the concentration of linkages is high among vertices in the same group and low otherwise. This motivates us to introduce a class of Gaussian graphical models with a community structure that replicates this empirical regularity. A natural question that arises in this framework is how to detect the communities from a random sample of observations. We introduce an algorithm called Blockbuster that recovers the communities using the eigenvectors of the sample covariance matrix. We study the properties of the procedure and establish consistency. The methodology is used to study real activity clustering in the U.S. and Europe.

Keywords: Partial Correlation Networks, Random Graphs, Community Detection, Spectral Clustering, Graphical Models

JEL: C3, C33, C55

[†] Department of Economics and Business, Universitat Pompeu Fabra and Barcelona GSE;
e-mail: christian.brownlees@upf.edu; stefan.gudmundsson@upf.edu; gabor.lugosi@upf.edu.
[◇] ICREA;
Christian Brownlees acknowledges financial support from the Spanish Ministry of Science and Technology (Grant MTM2015-67304) and from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2011-0075).
We have benefited from discussions with Majid Al-Sadoon, Banu Demir Pakel, Jin-Chuan Duan, Geert Mesters, Fabrizio Lillo, Cavit Pakel, Nicolas Verzelen, Michael Wolf and Piotr Zwiernik.

1 Introduction

Network analysis has become an active field of research in statistics and econometrics recently (Meinshausen and Bühlmann, 2006; Peng, Wang, Zhou, and Zhu, 2009; Diebold and Yilmaz, 2013). Broadly speaking, contributions to network analysis in these areas focus on developing methodologies to learn the interdependence structure of large multivariate systems. Network techniques have been applied successfully for dimensionality reduction and regularisation of such systems.

In the analysis of networks it is often found that vertices are partitioned into groups where the concentration of linkages is high among vertices in the same group and low otherwise. This empirical regularity is called *community structure* or *clustering* (Girvan and Newman, 2002). Over the years, an extensive literature has emerged and developed tools to detect communities in large networks, see Fortunato (2010) for a comprehensive overview. In this work we formalise the community detection problem for a class of Gaussian graphical models with a community structure and introduce an algorithm for community detection tailored to this setting.

We begin by introducing a class of graphical models (Lauritzen, 1996) where the network structure of the variables in the system is random. We consider an n-dimensional random vector from a multivariate Gaussian distribution. The covariance matrix of the model is assumed to be a function of a latent random graph defined on n vertices (Chung and Lu, 2006; van der Hofstad, 2015). The model is such that the i-th and j-th components of the random vector are conditionally independent (conditioned on the rest of the components) if and only if vertices i and j in the latent graph are not connected by an edge. Notice that two layers of randomness are present in this framework: The first layer comes from the random graph that determines the covariance structure of the observations and the second from the multivariate Gaussian sampling.

We use this framework to introduce two types of random graphical models with a community structure. In the first case, we assume that the latent graph is a stochastic block model (Holland, Laskey, and Leinhardt, 1983). The stochastic block model is a

random graph in which vertices are partitioned into k communities and edges are randomly determined by a Bernoulli trial, independent of all other edges. In the simplest version of the model, the probability of a link between two vertices is p if they belong to the same community and q otherwise. A limitation of the stochastic block model is that the expected degrees of all vertices in a given community is homogeneous, which is often an unrealistic assumption. In the second case, we assume that the latent graph is a degree-corrected stochastic block model (Karrer and Newman, 2011), which is an extension of the stochastic block model that allows for general heterogeneity in the expected degrees.

A natural question that arises in this setting is how to detect the communities of the model from a random sample of T independent and identically distributed (i.i.d.) observations of the multivariate Gaussian vector. An extensive literature (Fortunato, 2010) deals with the problem of community detection when the network structure of the data is observable. A popular algorithm to carry out community detection is the spectral clustering algorithm, which uses the eigenvectors of the graph Laplacian to detect community structure. In our framework, community detection is more challenging as the network structure of the data is not observed. Motivated by spectral clustering methods (Ng, Jordan, and Weiss, 2001; von Luxburg, 2007), we propose a community detection procedure called Blockbuster in which spectral clustering is applied to the sample covariance matrix of the observations. In particular, the algorithm applies k-means clustering to a matrix whose columns are the eigenvectors corresponding to the k largest eigenvalues of the sample covariance matrix. The k-means partition of the rows of the eigenvector matrix is the estimate of the community partition.

We study the properties of the Blockbuster community detection procedure and show that it consistently detects the communities when the size of the random vector and number of observations are large. More precisely, our key result establishes a bound on the fraction of vertices that the algorithm misclusters in a similar fashion to Rohe, Chatterjee, and Yu (2010), Qin and Rohe (2013) and Sarkar and Bickel (2013), and shows that it tends to zero with high probability as n and T grow large, provided that $n/T \to 0$. We carry out a simulation study which shows that the algorithm performs

adequately in finite samples.

A number of extensions to the baseline methodology are developed. Typically, panels of economic and financial time series exhibit evidence of a factor structure. To this extent, we consider an extension of the baseline model in which the components of the random vector are also influenced by a set of common factors. We then introduce a variant of the Blockbuster algorithm which detects communities in this setting and establish consistency. Last, we introduce a regularised covariance estimator based on Blockbuster motivated by the block covariance structure of the models introduced in this work.

We apply the methodology to analyse two panels of real activity growth measures for the U.S. and Europe. Detecting the communities of the European panel is of interest in light of recent statements of the ECB president Mario Draghi, who warned that a high degree of economic heterogeneity between Euro-zone countries might threaten the monetary union.[1] The methodology developed in this work can detect which areas, or communities, of Europe exhibit different behaviour from the data. For the U.S. we use a dataset constructed by Hamilton and Owyang (2011) comprising quarterly employment growth rates for the states of the U.S. (excluding Alaska and Hawaii) from 1956-Q2 to 2007-Q4. For Europe we use data from Eurostat comprising yearly gross regional product for the European Nomenclature of Territorial Units for Statistics (NUTS 1) regions (excluding Iceland and Turkey) from 2000 to 2013. Blockbuster delivers a meaningful partition of the states/regions in the two panels. In particular, the U.S. communities bear close resemblance with previously published results by Hamilton and Owyang (2011). In the U.S. dataset we also carry out an out-of-sample validation exercise and show that the covariance regularisation procedure based on Blockbuster improves covariance prediction compared to a number of alternative procedures.

This paper is related to several different strands of literature. First of all, the network and graphical modelling literature (Dempster, 1972; Lauritzen, 1996). Recently, estimation of large dimensional network models (typically through LASSO type estimation) has

[1] On May 23rd 2015 at the ECB Forum on Central Banking in Sintra (Portugal) Mario Draghi stated "*In a monetary union you can't afford to have large and increasing structural divergences between countries, they tend to become explosive*".

become an active field of research, see Meinshausen and Bühlmann (2006), Peng *et al.* (2009), Diebold and Yilmaz (2013) and Barigozzi and Brownlees (2013). Second, it is related to the community detection literature, see Donath and Hoffman (1973), Fiedler (1973), von Luxburg (2007), Girvan and Newman (2002), Newman (2006a) and Newman (2006b). Third, the paper builds upon the stochastic block model, which has garnered much interest since the work of Holland *et al.* (1983), see for example Abbe, Bandeira, and Hall (2014), Mossel, Neeman, and Sly (2012), Joseph and Yu (2013), Rohe *et al.* (2010), Qin and Rohe (2013) and Sarkar and Bickel (2013). Finally, an influential paper showing the relevance of networks in economics is, *inter alia*, Acemoglu, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012).

The rest of the paper is organised as follows. Section 2 presents the mathematical framework. Section 3 introduces a number of extensions to the baseline methodology. Section 4 carries out a simulation study to assess the finite-sample properties of the procedure. Section 5 uses our community detection methodology to study real activity clustering in the U.S. and Europe. Concluding remarks follow in Section 6.

2 Methodology

2.1 A Partial Correlation Network Model with Communities with Homogeneous Degrees

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ be an n-dimensional vector of zero-mean Gaussian random variables with covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}']$, where $\boldsymbol{Y}'$ is the transpose of $\boldsymbol{Y}$. The inverse covariance matrix $\mathbf{K} = \boldsymbol{\Sigma}^{-1}$, provided that it exists, plays a key role in this paper and is referred to as the concentration matrix hereafter. It is well known that the conditional dependence structure of the system is embedded in the concentration matrix. To see this, it is useful to introduce the notion of partial correlation, defined as the conditional

correlation between Y_i and Y_j given the remaining variables

$$\rho^{ij} = \mathrm{Corr}\left(Y_i, Y_j | \{Y_s : s \neq i, j\}\right).$$

Partial correlation measures the linear dependence between two variables after partialling out the influence of all other variables. It can be shown (Dempster, 1972; Pourahmadi, 2013) that the elements of the concentration matrix $[\mathbf{K}]_{ij}$ are related to the partial correlations through the identity

$$\rho^{ij} = -\frac{[\mathbf{K}]_{ij}}{\sqrt{[\mathbf{K}]_{ii}[\mathbf{K}]_{jj}}}, \quad i \neq j.$$

Thus, the concentration matrix and the partial correlations of $\boldsymbol{Y}$ share the same sparsity structure: The (i, j)-th element of $\mathbf{K}$ is zero if and only if Y_i and Y_j are conditionally uncorrelated. As $\boldsymbol{Y}$ is Gaussian, a zero partial correlation between variables Y_i and Y_j implies conditional independence given the remaining components of $\boldsymbol{Y}$.

In this work we assume that the conditional dependence structure of $\boldsymbol{Y}$ is determined by a latent graph. We denote an undirected graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, \ldots, n\}$ is the vertex set and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the edge set. It is useful to introduce the adjacency, degree and Laplacian matrices to represent the structure of a graph. The adjacency matrix $\mathbf{A}$ of a graph is defined as an $n \times n$ matrix with the (i, j)-th element $[\mathbf{A}]_{ij}$ equal to one if there is an edge between vertices i and j and zero otherwise. Let d_i denote the degree of vertex i, that is, $d_i = \sum_{j=1}^{n} [\mathbf{A}]_{ij}$. We define the degree matrix $\mathbf{D}$ as an $n \times n$ diagonal matrix with element $[\mathbf{D}]_{ii} = d_i$. Finally, the symmetric degree-normalised Laplacian is defined as $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Note that both the adjacency matrix and the Laplacian are symmetric.

We assume that the concentration matrix $\mathbf{K}$ of the graphical model is a function of the latent graph $\mathcal{G}$ and takes the form

$$\mathbf{K} = \frac{1}{\sigma^2}\mathbf{I}_n + \frac{\phi}{\sigma^2}\mathbf{L}, \tag{1}$$

where $\sigma^2 > 0$ is called the network-variance parameter and $\phi \geq 0$ the network-dependence parameter. Note that this definition guarantees that $[\mathbf{K}]_{ij}$ is zero if and only if i and j are not joined by an edge. The graph $\mathcal{G}$ thus determines the partial correlation network structure of the random vector $\boldsymbol{Y}$. Note that the model also guarantees that the concentration matrix is symmetric and positive definite.

In this work we assume that the latent graph $\mathcal{G}$ is generated randomly from a stochastic block model (Holland *et al.*, 1983). A random graph is defined as a graph in which the vertex set $\mathcal{V}$ is fixed and the existence of an edge in $\mathcal{E}$ is determined by a Bernoulli trial, independent of all other edges. The stochastic block model is an extension of the Erdős-Rényi random graph in which the vertex set $\mathcal{V}$ is partitioned into k subsets $\mathcal{V}_1, \ldots \mathcal{V}_k$, typically referred to as communities. An edge is present between vertices i and j with probability p_s if both vertices belong to $\mathcal{V}_s$ and probability q_{sv} if they belong to $\mathcal{V}_s$ and $\mathcal{V}_v$, respectively, with $s \neq v$. Let $\mathbb{1}_{a=b}$ be equal to one if $a = b$ and zero otherwise. We formally define the stochastic block model as follows, see (Qin and Rohe, 2013).

Definition 1. *Let $z : \mathcal{V} \to \{1, \ldots, k\}$ be the community assignment function. Let $\mathbf{Z}$ be the $n \times k$ community membership matrix, such that $[\mathbf{Z}]_{is} = \mathbb{1}_{z_i = s}$. Let $\mathbf{B}$ be the symmetric $k \times k$ matrix of community-specific edge probabilities such that $[\mathbf{B}]_{ss} = p_s$ and $[\mathbf{B}]_{sv} = q_{sv}$ if $s \neq v$. In a stochastic block model with parameters $\mathbf{Z}$ and $\mathbf{B}$, the probability of an edge between vertices i and j is $[\mathbf{B}]_{z_i z_j}$ and all edges are independent. It is convenient to write*

$$\mathcal{G} \sim \mathrm{SBM}\left(\mathbf{Z}, \mathbf{B}\right) \tag{2}$$

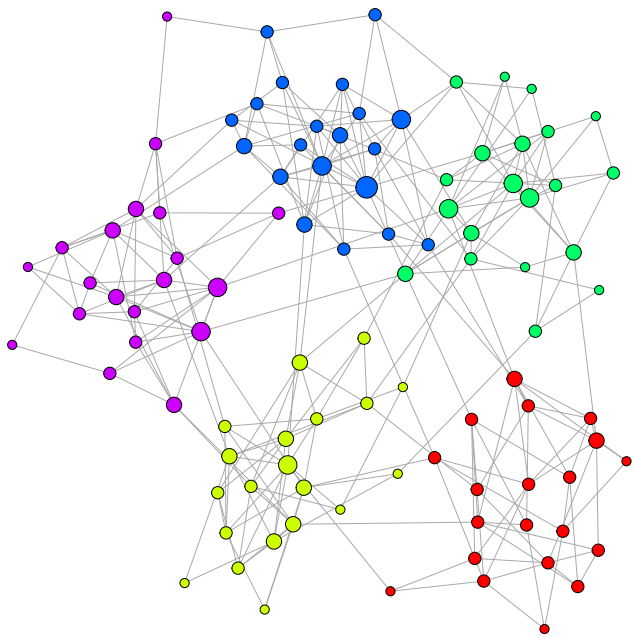
to indicate that a random graph $\mathcal{G}$ is a stochastic block model with parameters $\mathbf{Z}$ and $\mathbf{B}$.

We require $\mathbf{B}$ to be invertible so that all k communities are distinguishable. We furthermore assume that $\mathbf{B}$ is positive definite.[2] We assume there are no empty communities, so that each column of $\mathbf{Z}$ has at least one non zero entry. Notice that $\mathbf{Z}$ defines the community partition $\mathcal{V}^k = \{\mathcal{V}_1, \ldots \mathcal{V}_k\}$ of the vertex set $\mathcal{V}$. We let $n_s = |\mathcal{V}_s|$ denote the size of

[2]In fact, the algorithm we propose in this work can easily handle the case where $\mathbf{B}$ is negative definite. This corresponds to a bipartite graph with a heterophilic community structure, where there are few edges within the communities but many across them (Rohe *et al.*, 2010).

community s. We may assume that the elements of $\mathcal{V}_1$ are the first n_1 elements of $\mathcal{V}$, the elements of $\mathcal{V}_2$ the next n_2, and so on. Figure 1 gives an example of a stochastic block model with $n = 100$, $k = 5$, $p_s = p = 0.25$ and $q_{sv} = q = 0.01$.

Figure 1: The Stochastic Block Model



The figure displays a visual representation of a stochastic block model realisation with $n = 100$, $k = 5$, $p = 0.25$ and $q = 0.01$.

It is natural to characterize the behaviour of a random graph when n grows as a function of the edge probabilities in $\mathbf{B}$. For ease of notation, we introduce a common rate variable ρ_n that all the elements of $\mathbf{B}$ are proportional to, following Sarkar and Bickel (2013). That is, $[\mathbf{B}]_{ij} = b_{ij}\rho_n$ with $b_{ij} \geq 0$ for all $i, j = 1, \ldots, n$. The regime where ρ_n is constant results in a dense graph as n grows large, whereas regimes where the probabilities shrink to zero yield less dense graphs. We make a distinction between the sparse regime where ρ_n is proportional to $1/n$ and regimes where it shrinks to zero at a slower rate, for example $\log n/n$. In this paper we focus exclusively on the regime where $\log n/(n\rho_n)$ tends to zero as $n \to \infty$ which is termed the semi-sparse regime by Sarkar and Bickel (2013). This is among the sparsest regimes where exact recovery of the communities is possible

(Abbe *et al.*, 2014). Furthermore, norm convergence of graph matrices is well established in this regime, see (Oliveira, 2009; Chung and Radcliffe, 2011; Chaudhuri, Chung, and Tsiatas, 2012).

To summarise, we introduce a Gaussian graphical model for the vector $\boldsymbol{Y}$ where the concentration matrix $\mathbf{K}$ is based on a latent random graph $\mathcal{G}$ with a community structure. We call this model the stochastic block partial correlation network model and formally define it as follows.

Definition 2. *Let* $\mathcal{G} \sim \mathrm{SBM}\,(\mathbf{Z}, \mathbf{B})$ *be a stochastic block model as in* Definition 1*. Let* $\mathbf{K}$ *be the* $n \times n$ *concentration matrix corresponding to the random graph* $\mathcal{G}$*, defined as in* (1)*. In a stochastic block partial correlation network model, the* n*-dimensional vector* $\boldsymbol{Y}$ *is drawn from a multivariate normal with mean zero and covariance matrix* $\boldsymbol{\Sigma} = \mathbf{K}^{-1}$*.*

2.2 A Partial Correlation Network Model with Communities with Heterogeneous Degrees

In the stochastic block model, all vertices within a given community have the same expected degree. This is often an unrealistic assumption. In this section we extend our methodology by assuming that $\mathcal{G}$ is generated by a degree-corrected stochastic block model (Karrer and Newman, 2011), an extension of the stochastic block model that allows for a general degree distribution. A formal definition follows naturally from Definition 1.

Definition 3. *Let* $z : \mathcal{V} \to \{1, \ldots, k\}$ *be the community assignment function. Let* $\mathbf{Z}$ *be the* $n \times k$ *community membership matrix, such that* $[\mathbf{Z}]_{is} = \mathbb{1}_{z_i = s}$*. Let* $\mathbf{B}$ *be the symmetric* $k \times k$ *matrix of community-specific edge probabilities and let* $\boldsymbol{\Theta}$ *be the* $n \times n$ *diagonal matrix of non-negative, fixed and unknown vertex-specific probability weights. In a degree-corrected stochastic block model with parameters* $\mathbf{Z}$*,* $\mathbf{B}$ *and* $\boldsymbol{\Theta}$*, the probability of an edge between vertices* i *and* j *is* $[\boldsymbol{\Theta}]_{ii}[\mathbf{B}]_{z_i z_j}[\boldsymbol{\Theta}]_{jj} \leq 1$ *and all edges are independent. It is convenient to write*

$$\mathcal{G} \sim \mathrm{DCSBM}\,(\mathbf{Z}, \mathbf{B}, \boldsymbol{\Theta}) \tag{3}$$

to indicate that a random graph $\mathcal{G}$ is a degree-corrected stochastic block model.

The edge probabilities in the degree-corrected stochastic block model are weighted with vertex-specific weights $\theta_i = [\mathbf{\Theta}]_{ii} > 0$ for all $i = 1, \ldots, n$, so that two vertices that belong to the same community can have different expected degrees. Notice that $\mathbf{\Theta}$ and $\mathbf{B}$ are only unique up to a multiplicative constant. We follow Karrer and Newman (2011) and normalise θ_i such that they sum to one within communities, that is, $\sum_i \theta_i \mathbb{1}_{z_i=s} = 1$ for all $s = 1, \ldots, k$. Then $[\mathbf{B}]_{sv}$ can be interpreted as the expected number of edges between communities s and v if $s \neq v$ and twice the expected number of edges within community s if $s = v$.

We introduce a generalisation of the stochastic block partial correlation network model from Definition 2 based on the degree-corrected stochastic block model that allows for a general degree distribution.

Definition 4. *Let $\mathcal{G} \sim \mathrm{DCSBM}\,(\mathbf{Z}, \mathbf{B}, \mathbf{\Theta})$ be a degree-corrected stochastic block model as in* Definition 3. *Let $\mathbf{K}$ be the $n \times n$ concentration matrix corresponding to the random graph $\mathcal{G}$, defined as in* (1). *In a degree-corrected stochastic block partial correlation network model, the n-dimensional vector $\boldsymbol{Y}$ is drawn from a multivariate normal with mean zero and covariance matrix $\mathbf{\Sigma} = \mathbf{K}^{-1}$.*

2.3 The Blockbuster Algorithm

Suppose we observe a sample $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T$ of i.i.d. observations from a degree-corrected stochastic block partial correlation network model as in Definition 4 arranged into the $T \times n$ matrix $\mathbf{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T]'$, and let k be the number of communities. The community structure of the model is assumed to be unknown, although we assume that k is known. We adapt spectral clustering techniques to detect the communities of the model from the sample in a procedure we call Blockbuster. The proposed algorithm detects the communities using the eigenvectors of the sample covariance matrix.

We provide a description of the Blockbuster procedure in Algorithm 1. Given the sample and the number of communities k, the procedure consists of first constructing the

$n \times k$ matrix $\widehat{\mathbf{U}}$ of the eigenvectors corresponding to the k largest eigenvalues of the sample covariance matrix $\widehat{\mathbf{\Sigma}} = (1/T)\mathbf{Y}'\mathbf{Y}$.[3] We refer to this matrix as the matrix of partitioning eigenvectors. Then form the matrix $\widehat{\mathbf{X}}$ by normalising the rows of $\widehat{\mathbf{U}}$ to unit length, that is, $\widehat{\mathbf{X}} = \widehat{\mathbf{N}}\widehat{\mathbf{U}}$ where $\widehat{\mathbf{N}}$ is an $n \times n$ diagonal matrix with its i-th element $[\widehat{\mathbf{N}}]_{ii} = 1/\left\|[\widehat{\mathbf{U}}]_{i\bullet}\right\|$ and $\left\|[\widehat{\mathbf{U}}]_{i\bullet}\right\|$ is the Euclidean norm of $[\widehat{\mathbf{U}}]_{i\bullet}$, the i-th row of $\widehat{\mathbf{U}}$. The algorithm then applies k-means clustering to the rows of $\widehat{\mathbf{X}}$. The k-means algorithm partitions a set of data points into k clusters by solving the optimisation problem

$$\min_{\{\boldsymbol{m}_1,\ldots,\boldsymbol{m}_k\}} \sum_{i=1}^{n} \min_{s} \|[\widehat{\mathbf{X}}]_{i\bullet} - \boldsymbol{m}_s\|^2. \tag{4}$$

Let $\boldsymbol{m}_s^* \in \mathbb{R}^k$ be the vectors that solve it.[4] These vectors are called the centroids of the clusters returned by k-means. Each row $[\widehat{\mathbf{X}}]_{i\bullet}$ is then assigned to the cluster corresponding to the centroid closest to it. This yields a partition $\widehat{\mathcal{V}}^k = \{\widehat{\mathcal{V}}_1, \ldots, \widehat{\mathcal{V}}_k\}$ of the vertex set, which is our estimator of the community partition $\mathcal{V}^k$.

It is well known that the exact solution of (4) is NP hard, and that the standard iterative algorithm for approximation to its solution is prone to finding local minima, which may yield an approximation that is arbitrarily bad as compared to the optimal solution. We may use a variant of the algorithm called k-means++ (Arthur and Vassilvitskii, 2007), which guarantees an approximation within $O(\log k)$ of the optimum. For the simplicity of discussion, we assume that the minimum is computed. All theoretical findings remain true for k-means++.

2.4 Theory

In this section we show that the fraction of vertices that are incorrectly clustered by the Blockbuster algorithm tends to zero with probability approaching one as the cross-sectional dimension n and the number of observations T grow. We hereafter refer to an event that happens with probability approaching one as an event that happens with high

[3] If the k-th and $k+1$-th are tied, we may take both.

[4] The solution to the problem is not necessarily unique, but this is not a concern. We let $\boldsymbol{m}_s^*$ be some set of vectors that achieve the minimum in (4).

Algorithm 1 The Blockbuster Algorithm

INPUT: Sample $\boldsymbol{Y}_t$ for $t = 1, \ldots, T$, number of communities k.

PROCEDURE:

1. Compute the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$.
2. Construct the $[n \times k]$ eigenvector matrix $\widehat{\mathbf{U}}$ such that its columns are the eigenvectors corresponding to the k largest eigenvalues of $\widehat{\boldsymbol{\Sigma}}$.
3. Standardise each row of $\widehat{\mathbf{U}}$ by its norm and denote the row-normalized eigenvector matrix by $\widehat{\mathbf{X}}$, so that $[\widehat{\mathbf{X}}]_{ij} = [\widehat{\mathbf{U}}]_{ij} / \left\| [\widehat{\mathbf{U}}]_{i\bullet} \right\|$.
4. Apply the k-means algorithm to the rows of $\widehat{\mathbf{X}}$.

OUTPUT: Return the k-means partition $\widehat{\mathcal{V}}^k = \{\widehat{\mathcal{V}}_1, \ldots, \widehat{\mathcal{V}}_k\}$ as the estimate of the community structure.

probability. We work with the degree-corrected model as in Definition 4 in this section, as the results for the homogeneous degrees model of Definition 2 may be recovered by setting $\boldsymbol{\Theta} = \mathbf{I}_n$.

We introduce additional notation that is used throughout this work. Let $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$ denote the spectral and Frobenius norms of the $n \times n$ matrix $\mathbf{A}$, respectively. We denote the i-th smallest eigenvalue of $\mathbf{A}$ as $\lambda_i(\mathbf{A})$ and the set of eigenvalues of $\mathbf{A}$ that are in the interval S as $\lambda_S(\mathbf{A})$. Notice that we use the convention $\lambda_1(\mathbf{A}) \leq \ldots \leq \lambda_n(\mathbf{A})$. We refer to the eigenvectors corresponding to the k largest (smallest) eigenvalues of $\mathbf{A}$ as its k top (bottom) eigenvectors, counting multiplicities.

Define $\mathcal{A} = \mathbb{E}[\mathbf{A}]$ as the population adjacency matrix of the graph $\mathcal{G}$. Notice that if $\mathcal{G}$ is a degree-corrected stochastic block model, we may decompose the population adjacency matrix as $\mathcal{A} = \boldsymbol{\Theta}\mathbf{Z}\mathbf{B}\mathbf{Z}'\boldsymbol{\Theta}$. We also define the population degree matrix $\mathcal{D}$ as the diagonal matrix with $[\mathcal{D}]_{ii} = \sum_{j=1}^{n} [\mathcal{A}]_{ij}$, the population normalised Laplacian $\mathcal{L} = \mathbf{I}_n - \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}$ and the population concentration matrix $\mathcal{K} = (1/\sigma^2)\,\mathbf{I}_n + (\phi/\sigma^2)\,\mathcal{L}$, analogously to $\mathbf{D}$, $\mathbf{L}$ and $\mathbf{K}$. In this section we state results in terms of the sample concentration matrix $\widehat{\mathbf{K}} = \widehat{\boldsymbol{\Sigma}}^{-1}$, with $\widehat{\mathbf{U}}$ being the matrix of its bottom k eigenvectors. Let $\mathbf{U}$ and $\mathcal{U}$ be the $n \times k$ matrices of the bottom k eigenvectors of $\mathbf{K}$ and $\mathcal{K}$, respectively. Define $\mathbf{N}$ and $\mathcal{N}$ as $n \times n$ diagonal matrices with $[\mathbf{N}]_{ii} = \|[\mathbf{U}]_{i\bullet}\|^{-1}$ and $[\mathcal{N}]_{ii} = \|[\mathcal{U}]_{i\bullet}\|^{-1}$,

respectively. Let $\mathbf{X} = \mathbf{N}\mathbf{U}$ and $\mathcal{X} = \mathcal{N}\mathcal{U}$ be the row-normalised counterparts of $\mathbf{U}$ and $\mathcal{U}$.

To bound the fraction of misclustered vertices, we follow the strategy of Rohe *et al.* (2010) and Qin and Rohe (2013). We begin by noting that the k-means objective function of (4) can be written as

$$\min_{\{\boldsymbol{m}_1,\ldots,\boldsymbol{m}_k\}} \sum_{i=1}^{n} \min_{s} \|[\widehat{\mathbf{X}}]_{i\bullet} - \boldsymbol{m}_s\|^2 = \min_{\mathbf{M}\in\mathcal{M}(n,k)} \left\|\widehat{\mathbf{X}} - \mathbf{M}\right\|_F^2,$$

where

$$\mathcal{M}(n,k) = \{\mathbf{M} \in \mathbb{R}^{n\times k} : \mathbf{M} \text{ has no more than } k \text{ different rows}\}.$$

Let the estimated centroid matrix $\widehat{\mathbf{C}}$ be defined as

$$\widehat{\mathbf{C}} = \operatorname*{argmin}_{\mathbf{M}\in\mathcal{M}(n,k)} \left\|\widehat{\mathbf{X}} - \mathbf{M}\right\|_F^2. \tag{5}$$

Its i-th row $[\widehat{\mathbf{C}}]_{i\bullet}$ is equal to the k-means centroid that is closest to row i of the eigenvector matrix $\widehat{\mathbf{X}}$, so that $[\widehat{\mathbf{C}}]_{i\bullet} \in \{\boldsymbol{m}_1^*, \ldots, \boldsymbol{m}_k^*\}$. It is clear that the k-means centroid matrix has no more than k different rows. Let the population centroid matrix $\mathcal{C}$ be defined analogously to (5) with $\widehat{\mathbf{X}}$ replaced with $\mathcal{X}\mathcal{O}$, where $\mathcal{O}$ is a $k \times k$ orthonormal rotation matrix that depends on $\widehat{\mathbf{X}}$ and $\mathcal{X}$. This matrix is discussed in detail in Theorem 2. We adopt the same definition of misclustered vertices as Rohe *et al.* (2010) and Qin and Rohe (2013) and say that vertex i is correctly clustered if $[\widehat{\mathbf{C}}]_{i\bullet}$ is closer to $[\mathcal{C}]_{i\bullet}$ than any other population centroid $[\mathcal{C}]_{j\bullet}$ for $j \neq i$. Define the set

$$\mathscr{M} = \left\{i : \left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathcal{C}]_{i\bullet}\right\| \geq \sqrt{1/2}\right\}. \tag{6}$$

The condition $\left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathcal{C}]_{i\bullet}\right\| \geq \sqrt{1/2}$ is necessary for vertex i to be misclustered, as shown and discussed in more detail in Lemma 2. The size of the set $\mathscr{M}$ is thus an upper bound of the number of misclustered vertices.

Theorem 1 bounds the fraction $|\mathscr{M}|/n$. The theorem is in the spirit of Theorem 3.1

from Rohe *et al.* (2010) and Theorem 4.4 of Qin and Rohe (2013).

Theorem 1 (Misclustered Vertices)**.** *Consider a degree-corrected stochastic block partial correlation network model as in* Definition 4*. Let* $\mathscr{M}$ *be as in* (6)*. Assume that* $\mathbf{B}$ *is positive definite. Furthermore, assume that* $[\mathbf{B}]_{ij} = b_{ij}\rho_n$ *with* $b_{ij} \geq 0$ *for all* $i, j = 1, \ldots, n$ *and that* $\log n/(n\rho_n) \to 0$ *as* $n \to \infty$*.*

Then

$$\frac{|\mathscr{M}|}{n} = O\left(\frac{n}{T} + \frac{\log n}{n\rho_n}\right)$$

with high probability.

The theorem shows that the fraction of misclassified vertices goes to zero with high probability under the assumption that $\log n/(n\rho_n) \to 0$ as $n \to \infty$ and $n/T \to 0$.

The strategy we follow to establish Theorem 1 is the following. Lemma 1 and Corollary 1 abstract from all randomness and show that Blockbuster recovers the partition $\mathcal{V}^k$ trivially from the eigenvectors of the population concentration matrix $\mathcal{K}$. Theorem 2 uses random matrix concentration results from Oliveira (2009) and Vershynin (2010) to show the concentration of the first k eigenvectors of $\widehat{\mathbf{K}}$ around those of $\mathcal{K}$ with high probability as $n, T \to \infty$. Using these results, we can prove Theorem 1. Proofs of the statements in this section can be found in Appendix A.

The following lemma is a slight variation of Lemma 3.3 from Qin and Rohe (2013). The lemma establishes the form of the matrix of the bottom k eigenvectors of $\mathcal{L}$, and its row-normalised counterpart.

Lemma 1. *Let* $\mathcal{G} \sim \mathrm{DCSBM}\left(\mathbf{Z}, \mathbf{B}, \mathbf{\Theta}\right)$ *be a degree-corrected stochastic block model as in* Definition 3*. Let* $\mathcal{A} = \mathbf{\Theta Z B Z' \Theta}$ *be the population adjacency matrix,* $\mathcal{D}$ *the population degree matrix,* $\mathcal{L} = \mathbf{I}_n - \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}$ *the population normalised Laplacian, and* $\lambda_i(\mathcal{L})$ *the* i*-th eigenvalue of* $\mathcal{L}$*. Assume that* $\mathbf{B}$ *is positive definite.*

Then $\lambda_i(\mathcal{L}) = 1$ *for all* $i = k+1, \ldots, n$ *and* $\lambda_i\left(\mathcal{L}\right) \in [0, 1)$ *for* $i = 1, \ldots, k$*. Furthermore, there exists a* $k \times k$ *orthonormal matrix* $\mathbf{V}$ *such that* $\mathbf{\Theta}^{1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Theta}\mathbf{Z}\right)^{-1/2}\mathbf{V}$ *is an* $n \times k$ *matrix with the* k *bottom eigenvectors of* $\mathcal{L}$ *as columns and the row-normalised counterpart*

of this matrix is $\mathbf{ZV}$*, that is,* $[\mathbf{ZV}]_{ij} = [\mathbf{\Theta}^{1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Theta}\mathbf{Z}\right)^{-1/2}\mathbf{V}]_{ij} / \left\| [\mathbf{\Theta}^{1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Theta}\mathbf{Z}\right)^{-1/2}\mathbf{V}]_{i\bullet} \right\|$ *for* $i, j = 1, \ldots, n$*.*

An immediate corollary of Lemma 1 is:

Corollary 1. *Let* $\mathcal{G} \sim \mathrm{DCSBM}\left(\mathbf{Z}, \mathbf{B}, \mathbf{\Theta}\right)$ *be a degree-corrected stochastic block model as in* Lemma 1*. Let the population concentration matrix* $\mathcal{K}$ *be defined as* $\frac{1}{\sigma^2}\mathbf{I}_n + \frac{\phi}{\sigma^2}\mathcal{L}$ *as the population analogue of* (1) *and let* $\mathcal{U}$ *be the matrix of its bottom* k *eigenvectors. Assume that* $\mathbf{B}$ *is positive definite.*

Then $\lambda_i(\mathcal{K}) = (1+\phi)/\sigma^2$ *for all* $i = k+1, \ldots, n$ *and* $\lambda_i\left(\mathcal{K}\right) \in [1/\sigma^2, (1+\phi)/\sigma^2)$ *for* $i = 1, \ldots, k$*. Furthermore, there exists a* $k \times k$ *orthonormal matrix* $\mathbf{V}$ *such that* $\mathcal{U} = \mathbf{\Theta}^{1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Theta}\mathbf{Z}\right)^{-1/2}\mathbf{V}$ *are the bottom* k *eigenvectors of* $\mathcal{K}$ *and* $\mathcal{X} = \mathbf{ZV}$ *is its row-normalised counterpart, that is,* $[\mathcal{X}]_{ij} = [\mathcal{U}]_{ij} / \|[\mathcal{U}]_{i\bullet}\|$ *for* $i, j = 1, \ldots, n$*.*

Corollary 1 establishes the form of $\mathcal{U}$ and its row-normalised counterpart $\mathcal{X}$. In particular, rows i and j of $\mathcal{U}$ have the same direction if $z_i = z_j$, but are orthogonal if the vertices do not belong to the same community. Row i of $\mathcal{U}$ has length

$$\|[\mathcal{U}]_{i\bullet}\| = \left(\frac{\theta_i}{\sum_j \theta_j \mathbb{1}_{z_i = z_j}} \right)^{1/2} \tag{7}$$

and rows corresponding to vertices in the same community may thus have different lengths. Furthermore, $[\mathcal{X}]_{i\bullet} = [\mathcal{X}]_{j\bullet}$ if and only if $[\mathbf{Z}]_{i\bullet} = [\mathbf{Z}]_{j\bullet}$, so the rows corresponding to two vertices that belong to the same community are equal in $\mathcal{X}$. Hence there are only k different rows in $\mathcal{X}$ and k-means can trivially recover the partition $\mathcal{V}^k$ from $\mathcal{X}$ by selecting each of the different rows as centroids.

Theorem 2 establishes the concentration of the eigenvectors of $\widehat{\mathbf{K}}$ around those of $\mathcal{K}$, using concentration results from Oliveira (2009)[5] and Vershynin (2010).

Theorem 2 (Concentration)**.** *Let* $\mathbf{Y}_t$ *for* $t = 1, \ldots, T$ *be i.i.d. observations from a degree-corrected stochastic block partial correlation network model as in* Definition 4*. Let* $\widehat{\mathbf{K}}$ *be*

[5] It should be noted that Chung and Radcliffe (2011) extend the results of Oliveira (2009) by improving the bound and weakening the minimum degree assumption, both by a constant.

the inverse of the sample covariance matrix, $\widehat{\mathbf{U}}$ *the matrix of its bottom* k *eigenvectors and let* $\mathcal{U}$ *be as in* Corollary 1. *Let* $\widehat{\mathbf{N}}$ *and* $\mathcal{N}$ *be diagonal matrices with* $[\widehat{\mathbf{N}}]_{ii} = \left\|[\widehat{\mathbf{U}}]_{i\bullet}\right\|^{-1}$ *and* $[\mathcal{N}]_{ii} = \|[\mathcal{U}]_{i\bullet}\|^{-1}$, *respectively. Let* $\widehat{\mathbf{X}} = \widehat{\mathbf{N}}\widehat{\mathbf{U}}$ *and* $\mathcal{X} = \mathcal{N}\mathcal{U}$ *be the row-normalised counterparts of* $\widehat{\mathbf{U}}$ *and* $\mathcal{U}$. *Assume that the matrix* $\mathbf{B}$ *is positive definite. Furthermore, assume that* $[\mathbf{B}]_{ij} = b_{ij}\rho_n$ *with* $b_{ij} \geq 0$ *for all* $i, j = 1, \ldots, n$ *and that* $\log n/(n\rho_n) \to 0$ *as* $n \to \infty$.

Then there exists a $k \times k$ *orthonormal rotation matrix* $\mathcal{O}$, *that depends on* $\widehat{\mathbf{U}}$ *and* $\mathcal{U}$, *such that*

$$\left\|\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}\right\| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)$$

and

$$\left\|\widehat{\mathbf{X}} - \mathcal{X}\mathcal{O}\right\| = O\left(\frac{n}{\sqrt{T}} + \sqrt{\frac{\log n}{\rho_n}}\right)$$

with high probability.

The theorem shows that the spectral norm of the difference between $\widehat{\mathbf{U}}$ and a rotation of $\mathcal{U}$ is of the order $O\left(\sqrt{n/T} + \sqrt{\log n/(n\rho_n)}\right)$ with high probability as n and T grow. The eigenvectors of $\widehat{\mathbf{K}}$ thus converge to the rotated eigenvectors of $\mathcal{K}$ with high probability under the assumptions that $\log n/(n\rho_n) \to 0$ and $n/T \to 0$.[6] The theorem also establishes that the spectral norm of $\widehat{\mathbf{X}} - \mathcal{X}\mathcal{O}$ is of the order $O\left(n/\sqrt{T} + \sqrt{\log n/\rho_n}\right)$, which is larger by a factor of $\sqrt{n}$ as compared to the unnormalised case. This implies that the row-normalised eigenvectors do not necessarily converge under our assumption on the probabilities, but the next theorem demonstrates that this is not a concern.

The estimated centroid matrix $\widehat{\mathbf{C}}$ was defined in 5. We define

$$\mathcal{C} = \underset{\mathbf{M} \in \mathcal{M}(n,k)}{\operatorname{argmin}} \|\mathcal{X}\mathcal{O} - \mathbf{M}\|_F^2. \tag{8}$$

We include the orthonormal rotation because the sample eigenvector matrix $\widehat{\mathbf{X}}$ concentrates around the rotated population eigenvectors $\mathcal{X}\mathcal{O}$ as shown in Theorem 2. In the

[6]The rotation is required because $\widehat{\mathbf{K}} = \mathcal{K}$ does not necessarily imply $\widehat{\mathbf{U}} = \mathcal{U}$. At the very least, the columns might be permuted.

degree-corrected stochastic block model we have $\mathcal{C} = \mathcal{X}\mathcal{O} = \mathbf{Z}\mathbf{V}\mathcal{O}$. It follows that $[\mathcal{C}]_{i\bullet} = [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O}$. Thus for the degree-corrected stochastic block model we can equivalently define vertex i as correctly clustered if $[\widehat{\mathbf{C}}]_{i\bullet}$ is closer to $[\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O}$ than any other rotated population centroid $[\mathbf{Z}]_{j\bullet}\mathbf{V}\mathcal{O}$, $j \neq i$.

The next lemma provides a condition that implies that vertex i is correctly clustered. The lemma is an extension of Lemma 3.2 of Rohe *et al.* (2010).

Lemma 2. *Let* $\widehat{\mathbf{C}}$ *be the estimated centroid matrix as in* (5)*. Let* $\mathbf{V}$ *and the community membership matrix* $\mathbf{Z}$ *be as in* Corollary 1*, and let the orthonormal rotation* $\mathcal{O}$ *be as in* Theorem 2*.*

Then for all $j \neq i$

$$\left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O}\right\| < \left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{j\bullet}\mathbf{V}\mathcal{O}\right\|$$

whenever $\left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O}\right\| < \sqrt{1/2}$.

The lemma justifies bounding the number of misclustered vertices by the number of nodes that do not satisfy the condition $\left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O}\right\| < \sqrt{1/2}$, which is sufficient for vertex i to be correctly clustered. The set $\mathscr{M}$ from (6) hence contains the set of all misclustered nodes. At this point, all the tools for proving Theorem 1 have been established.

2.5 Discussion

A number of comments on the degree-corrected stochastic block partial correlation network model of Definition 4 and the Blockbuster algorithm are in order.

Blockbuster is essentially an application of the spectral clustering algorithm of Ng *et al.* (2001) to the sample covariance matrix. Spectral clustering is a popular algorithm that detects communities by applying k-means clustering to the bottom k eigenvectors of the normalised Laplacian. For more background on spectral clustering and its various versions, see von Luxburg (2007).

An alternative approach to detect communities consists of estimating the partial correlation network and then applying spectral clustering to the estimated network Laplacian. The estimation of the graphical structure of the data is typically carried out using LASSO techniques, which also requires the selection of a tuning parameter that determines the sparsity of the estimated network. A highlight of our approach is that it allows one to learn the community structure without estimating the structure of the network.

The stochastic block partial correlation network model bears some similarities to a factor model with a block structure, that is, a model in which the variables in the system are generated by a factor model and each variable is only loaded on by exactly one factor. Examples of block factor models in the literature include Goldstein and Browne (2002), Choi, Kim, Kim, and Kwark (2014) and Moench, Ng, and Potter (2009). However, a block factor model is not equivalent to a stochastic block partial correlation network model in general. Notice that a block factor model is associated with a partial correlation network in which each community is a clique, that is, a sub-graph where every possible pair of vertices is joined by an edge. Our model can replicate such network structures with $p_s = 1$ and $q_{vr} = 0$, but it also allows for structures where the communities are less dense.

Our results in Section 2 hold without modifications when $\boldsymbol{Y}$ has a sub-Gaussian distribution. The family of sub-Gaussian probability distributions is characterised by a super-exponential tail decay. A random variable $\boldsymbol{Y}$ is said to be sub-Gaussian when $\mathbb{P}\left(|\boldsymbol{Y}| > t\right) \leq e^{1-t^2/C^2}$ for all $t \geq 0$, where $C > 0$, see Vershynin (2010). Furthermore, the analysis may be extended to heavy-tailed distributions at the cost of a worse bound, where $(n \log n)/T \to 0$ would be required for convergence, see Corollary 5.52 in Vershynin (2010).

The concentration results in Theorem 2 rely on the minimum degree being of order $\Omega\left(\log n\right)$. While this is true under our assumption on the community-specific probabilities in $\mathbf{B}$ for both the stochastic block model and the degree-corrected stochastic block model, low vertex weights in $\boldsymbol{\Theta}$ may be problematic. A common way to deal with this problem is to regularise the Laplacian as suggested in Chen, Amini, Bickel, and Levina (2012) and implemented in amongst others Chaudhuri *et al.* (2012), Joseph and Yu (2013) and Qin

and Rohe (2013). We however do not pursue this for the concentration matrix in this paper.

3 Extensions

3.1 Community Detection in the Presence of Common Factors

In this section we introduce an extension of the model in Definition 4 in which the components of the random vector $\boldsymbol{Y}$ are influenced by a set of common factors, and a modification of the Blockbuster algorithm that consistently detects communities in this setting.

Let $\boldsymbol{F}$ be an R-dimensional Gaussian vector of common factors with mean zero and covariance matrix $\mathbf{I}_R$ and let $\boldsymbol{q}_r$ be n-dimensional fixed vectors of factor loadings for $r = 1, \ldots, R$. We assume that the random vector $\boldsymbol{Y}$ is generated as

$$\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{F} + \boldsymbol{\epsilon}, \tag{9}$$

where $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_R]$ is an $n \times R$ matrix of factor loadings and $\boldsymbol{\epsilon}$ is generated by a degree-corrected stochastic block partial correlation model as in Definition 4 with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. We define the degree-corrected stochastic block partial correlation model with common factors as follows.

Definition 5. *Let $\mathcal{G} \sim \mathrm{DCSBM}\,(\mathbf{Z}, \mathbf{B}, \boldsymbol{\Theta})$ be a degree-corrected stochastic block model as in* Definition 3*. Let $\mathbf{K}_{\boldsymbol{\epsilon}}$ be the $n \times n$ concentration matrix corresponding to the random graph $\mathcal{G}$, defined as in* (1)*. In a degree-corrected stochastic block partial correlation network model with common factors, the n-dimensional vector $\boldsymbol{Y}$, defined as in* (9)*, is drawn from a multivariate normal with mean zero and covariance matrix $\boldsymbol{\Sigma} = \mathbf{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R} \boldsymbol{q}_r \boldsymbol{q}_r'$.*

Consider a sample of i.i.d. observations drawn from the model in Definition 5 arranged into a $T \times n$ matrix $\mathbf{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T]'$. Under appropriate conditions, the top R eigenvectors of $\widehat{\boldsymbol{\Sigma}} = (1/T)\mathbf{Y}'\mathbf{Y}$ correspond to the R factors and the next k eigenvectors are the partitioning eigenvectors. This motivates Algorithm 2: Given $\mathbf{Y}$ and k, first construct

the $n \times k$ matrix $\widehat{\mathbf{U}}$ such that it contains the eigenvectors corresponding to the $R+1$-th largest to the $R+k$-th largest eigenvalues of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$. Then form the row-normalised eigenvector matrix $\widehat{\mathbf{X}}$ as in Algorithm 1 and apply k-means clustering to its rows.

Algorithm 2 The Blockbuster Algorithm in the Presence of Common Factors

INPUT: Sample $\boldsymbol{Y}_t$ for $t = 1, \ldots, T$, number of communities k, number of factors R.

PROCEDURE:

1. Compute the sample covariance matrix $\widehat{\mathbf{\Sigma}}$.
2. Construct the $[n \times k]$ eigenvector matrix $\widehat{\mathbf{U}}$ such that its columns are the eigenvectors corresponding to the $R+1$-th to the $R+k$-th largest eigenvalues of $\widehat{\mathbf{\Sigma}}$.
3. Standardise each row of $\widehat{\mathbf{U}}$ by its norm and denote the row-normalized eigenvector matrix by $\widehat{\mathbf{X}}$, so that $[\widehat{\mathbf{X}}]_{ij} = [\widehat{\mathbf{U}}]_{ij} / \left\| [\widehat{\mathbf{U}}]_{i\bullet} \right\|$.
4. Apply the k-means algorithm to the rows of $\widehat{\mathbf{X}}$.

OUTPUT: Return the k-means partition $\widehat{\mathcal{V}}^k = \{\widehat{\mathcal{V}}_1, \ldots, \widehat{\mathcal{V}}_k\}$ as the estimate of the community structure.

It is convenient to state the results of this section in terms of the concentration matrices rather than covariances. Let $\mathcal{K}$ and $\mathcal{K}_{\boldsymbol{\epsilon}}$ be the population concentration matrices of $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ from (9), respectively, and let $\boldsymbol{u}_i(\mathcal{K})$ and $\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}})$ for $i = 1, \ldots, n$ be their eigenvectors. Let $\mathcal{U}_{\boldsymbol{\epsilon}}$ be the $n \times k$ matrix of the bottom k eigenvectors of $\mathcal{K}_{\boldsymbol{\epsilon}}$ and $\mathcal{U}$ be the matrix of the eigenvectors corresponding to the $R+1$-th smallest to the $R+k$-th smallest eigenvalues of $\mathcal{K}$, so that $[\mathcal{U}]_{\bullet i} = \boldsymbol{u}_{R+i}(\mathcal{K})$ for $i = 1, \ldots, k$, where $[\mathcal{U}]_{\bullet i}$ refers to the i-th column of $\mathcal{U}$.

Throughout this section we assume that the factors are mutually orthogonal so that $\boldsymbol{q}_r' \boldsymbol{q}_v = 0$ for all $r, v = 1, \ldots, R$, $r \neq v$. We furthermore assume that $\boldsymbol{q}_r' [\mathcal{U}]_{\bullet i} = 0$, for all $r = 1, \ldots, R$ and $i = 1, \ldots, k$. This assumption implies that the factor loading vectors carry no information on the community structure nor the degrees of the underlying graph $\mathcal{G}$. We may assume without loss of generality that $\|\boldsymbol{q}_1\| \geq \|\boldsymbol{q}_2\| \geq \ldots \geq \|\boldsymbol{q}_R\|$. Finally, to guarantee that the factors are strong enough to be dominant in the spectrum $\mathcal{K}$, we assume that $\|\boldsymbol{q}_R\|^2$ is larger than a constant which does not depend on n. Notice that this assumption is trivially satisfied when n is large enough, if we assume $\|\boldsymbol{q}_R\|^2 = O(n)$,

which is standard in the factor model literature.

We now show that Algorithm 2 recovers the communities with high probability under suitable conditions. Theorem 3 extends Theorem 1 to bound the fraction $|\mathscr{M}|/n$ under the model in Definition 5.

Theorem 3. *Consider a degree-corrected stochastic block partial correlation network model with common factors as in* Definition 5. *Let $\mathscr{M}$ be defined analogously to* (6) *for this model. Assume that $\mathbf{B}$ is positive definite. Furthermore, assume that $[\mathbf{B}]_{ij} = b_{ij}\rho_n$ with $b_{ij} \geq 0$ for all $i, j = 1, \ldots, n$ and that $\log n/(n\rho_n) \to 0$ as $n \to \infty$. Finally, assume that $\boldsymbol{q}_r'\boldsymbol{q}_v = 0$ $r, v = 1, \ldots, R$, $r \neq v$, $\boldsymbol{q}_r'[\mathcal{U}]_{\bullet i} = 0$ for all $r = 1, \ldots, R$ and $i = 1, \ldots, k$, and $\|\boldsymbol{q}_R\|^2 > \frac{\sigma^2\phi}{1+\phi}$.*

Then

$$\frac{|\mathscr{M}|}{n} = O\left(\frac{n}{T}\|\mathbf{\Sigma}\|^2 + \frac{\log n}{n\rho_n}\right)$$

with high probability.

We follow a similar strategy to Section 2 to establish Theorem 3. Lemma 3 extends the population results in Lemma 1 and Corollary 1 to the model in Definition 5.

Lemma 3. *Let $\mathcal{G} \sim \mathrm{DCSBM}(\mathbf{Z}, \mathbf{B}, \mathbf{\Theta})$ be a degree-corrected stochastic block model as in* Lemma 1. *Let the population covariance matrix in the presence of common factors be $\mathcal{K}^{-1} = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R} \boldsymbol{q}_r\boldsymbol{q}_r'$, where $\mathcal{K}_{\boldsymbol{\epsilon}}$ is defined as $\frac{1}{\sigma^2}\mathbf{I}_n + \frac{\phi}{\sigma^2}\mathcal{L}$ as the population analogue of* (1). *Let $\mathcal{U}$ be the matrix of the R-th to the $R+k$-th bottom eigenvectors of $\mathcal{K}$. Assume that $\mathbf{B}$ is positive definite. Furthermore, assume that $\boldsymbol{q}_r'\boldsymbol{q}_v = 0$ $r, v = 1, \ldots, R$, $r \neq v$, $\boldsymbol{q}_r'[\mathcal{U}]_{\bullet i} = 0$ for all $r = 1, \ldots, R$ and $i = 1, \ldots, k$, and $\|\boldsymbol{q}_R\|^2 > \frac{\sigma^2\phi}{1+\phi}$.*

Then $\lambda_i(\mathcal{K}) = 1/\left(\sigma^2/(1+\phi) + \|\boldsymbol{q}_i\|^2\right)$ for $i = 1, \ldots R$, $\lambda_i(\mathcal{K}) \in [1/\sigma^2, (1+\phi)/\sigma^2)$ for $i = R+1, \ldots, R+k$ and $\lambda_i(\mathcal{K}) = (1+\phi)/\sigma^2$ for all $i = R+k+1, \ldots, n$. Furthermore, there exists a $k \times k$ orthonormal matrix $\mathbf{V}$ such that $\mathcal{U} = \mathbf{\Theta}^{1/2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{\Theta}\mathbf{Z}\right)^{-1/2}\mathbf{V}$ are the $R+1$-th to the $R+k$-th eigenvectors of $\mathcal{K}$ and $\mathcal{X} = \mathbf{Z}\mathbf{V}$ is its row-normalised counterpart, that is, $[\mathcal{X}]_{ij} = [\mathcal{U}]_{ij}/\|[\mathcal{U}]_{i\bullet}\|$ for $i, j = 1, \ldots, n$.

The $R+1$-th to the $R+k$-th bottom eigenvectors of $\mathcal{K}$ contained in $\mathcal{U}$ thus allow

k-means to recover the community partition, by a similar logic to before. Arguments analogous to those in Section 2 then allow us to prove Theorem 3.

3.2 Community Structure and Covariance Estimation

The presence of a community structure in the data suggests a natural covariance estimation strategy. The covariance matrix of the data can be estimated by a block covariance matrix estimator where the diagonal blocks contain the sample covariance matrices of each community, while the off-diagonal blocks are set to zero.[7] We provide a description of an estimator based on this idea in Algorithm 3.

Algorithm 3 Blockbuster Covariance Estimator

INPUT: Sample $\boldsymbol{Y}_t$ for $t = 1, \ldots, T$, number of communities k.

PROCEDURE:

1. Run the Blockbuster algorithm and obtain an estimate of the community partition of the panel $\widehat{\mathcal{V}}^k = \{\widehat{\mathcal{V}}_1, \ldots, \widehat{\mathcal{V}}_k\}$.
2. Reorder the series in the panel so that the first $\hat{n}_1 = |\widehat{\mathcal{V}}_1|$ series are the ones in community $\widehat{\mathcal{V}}_1$, the following $\hat{n}_2$ series are the ones in $\widehat{\mathcal{V}}_2$, and so on.
3. Let $\widehat{\boldsymbol{\Sigma}}_s$ denote the $\hat{n}_s \times \hat{n}_s$ sample covariance matrix of the series in community $\widehat{\mathcal{V}}_s$.
4. The Blockbuster covariance estimator $\widehat{\boldsymbol{\Sigma}}_B$ is defined as

$$\widehat{\boldsymbol{\Sigma}}_{B\,sv} = \begin{cases} \widehat{\boldsymbol{\Sigma}}_s & s = v \\ \mathbf{0}_{\hat{n}_s, \hat{n}_v} & \text{otherwise} \end{cases}$$

   where $\widehat{\boldsymbol{\Sigma}}_{B\,sv}$ denotes the (s, v)-th $\hat{n}_s \times \hat{n}_v$ block of $\widehat{\boldsymbol{\Sigma}}_B$ and $\mathbf{0}_{\hat{n}_s, \hat{n}_v}$ the $\hat{n}_s \times \hat{n}_v$ matrix of zeros.

OUTPUT: Return the Blockbuster covariance $\widehat{\boldsymbol{\Sigma}}_B$.

In case the data has a factor structure with R common factors as in the model in Definition 5, we may employ a regularisation approach similar to POET (Fan, Liao, and Mincheva, 2013). Let $\lambda_i(\widehat{\boldsymbol{\Sigma}})$ and $\boldsymbol{u}_i(\widehat{\boldsymbol{\Sigma}})$ denote the i-th eigenvalue and eigenvector of the

[7] Other possible shrinkage estimation strategies can be considered. For example, one could instead shrink or threshold the off-diagonal blocks of the covariance matrix.

sample covariance matrix. Then we can regularise the sample covariance of the data using

$$\widehat{\mathbf{\Sigma}}_B = \sum_{i=n-R+1}^{n} \lambda_i(\widehat{\mathbf{\Sigma}})\boldsymbol{u}_i(\widehat{\mathbf{\Sigma}})\boldsymbol{u}_i'(\widehat{\mathbf{\Sigma}}) + \widehat{\mathbf{R}}_B,$$

where $\widehat{\mathbf{R}}_B$ is the Blockbuster covariance estimator applied to the residual matrix $\widehat{\mathbf{R}} = \sum_{i=1}^{n-R} \lambda_i(\widehat{\mathbf{\Sigma}})\boldsymbol{u}_i(\widehat{\mathbf{\Sigma}})\boldsymbol{u}_i'(\widehat{\mathbf{\Sigma}})$.

In the empirical application section of this paper we use the estimator described in Algorithm 3 to validate the community partition estimated by Blockbuster out-of-sample. Analysing the theoretical properties of the estimator is not the focus of this work however, so we leave it open for future research.

4 Simulation Study

In this section we carry out a simulation study to investigate the finite-sample properties of the Blockbuster algorithm.

The simulation exercise consists of simulating a sample of size T on n vertices from the stochastic block partial correlation network model in Definition 2, its degree-corrected variant of Definition 4 and the degree corrected variant with common factors as in Definition 5. The number of communities k is set to 5 and each community has size n/k. The edge probabilities are equal to $p_s = p = \frac{c_p(\log n)^{1.01}}{n}$ for all $s = 1, \ldots, k$ and $q_{vr} = q = \frac{c_q(\log n)^{1.01}}{n}$ for all $v, r = 1, \ldots, k$, $v \neq r$. We calibrate c_p and c_q so that when $n = 100$ the (p, q) pair is equal to either $(0.25, 0.01)$, $(0.25, 0.05)$ or $(0.50, 0.01)$. Note that as the edge probabilities are functions of n, varying n changes the probabilities. The network-dependence parameter ϕ is either 5 or 50 while the network variance σ^2 is 1. When simulating from the degree-corrected models, we draw θ_i once from a power law distribution $f(x) = \alpha x_m^{\alpha}/x^{\alpha+1}$ for $x \geq x_m$ with parameters $x_m = 1$ and $\alpha = 2.5$, and keep it constant over replications. In the model with common factors, we draw a single factor $\boldsymbol{q}_r = \boldsymbol{q}$ once from a standard normal and keep it constant over replications.

We then apply the Blockbuster algorithm to the sample to recover the communities. To

measure the quality of the Blockbuster partition $\widehat{\mathcal{V}}^k$ we compare it to $\mathcal{V}^k$ by calculating the fraction of correctly classified vertices, which we call hit ratio. As we previously discussed, $\widehat{\mathcal{V}}^k$ only estimates $\mathcal{V}^k$ up to a permutation. As the number of communities k is low in our simulations, we calculate the hit percentage for every possible permutation and select the maximum as the final hit ratio figure.

We replicate the Monte Carlo experiment 1 000 times for different values of n (50, 100 and 500) and T (50, 100, 1 000 and 10 000). The results of the study are summarised in Table 1 for the stochastic block partial correlation network model, in Table 2 for the degree-corrected model and in Table 3 for the degree-corrected model with common factors.

Table 1: Hit Ratios Stochastic Block Partial Corelation Network Model

	Panel A: $\phi = 5$				Panel B: $\phi = 50$			
	$T = 50$	100	1 000	10 000	50	100	1 000	10 000
p/q	$n = 50$				$n = 50$			
0.25/0.01	74%	87%	97%	97%	90%	95%	97%	98%
0.50/0.01	93%	99%	100%	100%	100%	100%	100%	100%
0.25/0.05	43%	47%	70%	76%	47%	53%	74%	76%
	$n = 100$				$n = 100$			
0.25/0.01	54%	72%	97%	98%	82%	93%	98%	98%
0.50/0.01	72%	92%	100%	100%	98%	100%	100%	100%
0.25/0.05	35%	37%	60%	76%	37%	41%	67%	78%
	$n = 500$				$n = 500$			
0.25/0.01	30%	33%	92%	99%	42%	61%	98%	99%
0.50/0.01	33%	43%	99%	100%	66%	87%	100%	100%
0.25/0.05	25%	26%	31%	76%	26%	27%	38%	82%

The table reports hit ratios for Blockbuster applied to simulated data from the Stochastic Block Partial Correlation Network Model of Definition 2. The probabilities in the rows correspond to $n = 100$.

Table 2: Hit Ratio Degree-Corrected Stochastic Block Partial Corelation Network Model

	Panel A: $\phi = 5$				Panel B: $\phi = 50$			
	$T = 50$	100	1 000	10 000	50	100	1 000	10 000
p/q	$n = 50$				$n = 50$			
0.25/0.01	72%	84%	95%	96%	88%	92%	95%	96%
0.50/0.01	90%	99%	100%	100%	99%	100%	100%	100%
0.25/0.05	43%	47%	64%	69%	47%	52%	67%	69%
	$n = 100$				$n = 100$			
0.25/0.01	52%	69%	95%	96%	79%	90%	96%	97%
0.50/0.01	70%	90%	100%	100%	96%	99%	100%	100%
0.25/0.05	34%	37%	55%	62%	37%	41%	58%	63%
	$n = 500$				$n = 500$			
0.25/0.01	29%	33%	88%	98%	41%	59%	96%	98%
0.50/0.01	33%	42%	98%	100%	65%	85%	100%	100%
0.25/0.05	25%	26%	31%	47%	26%	26%	35%	50%

The table reports hit ratios for Blockbuster applied to simulated data from the Degree-Corrected Stochastic Block Partial Correlation Network Model of Definition 4. The vertex-specific weight $\boldsymbol{\Theta}$ is drawn once from a power law with $x_m = 1$ and $\beta = 2.5$, and then kept constant over replications. The probabilities in the rows correspond to $n = 100$.

Table 3: Hit Ratio Degree-Corrected Stochastic Block Partial Corelation Network Model with Common Factors

	Panel A: $\phi = 5$				Panel B: $\phi = 50$			
	$T = 50$	100	1 000	10 000	50	100	1 000	10 000
p/q	$n = 50$				$n = 50$			
0.25/0.01	71%	85%	95%	95%	88%	91%	94%	95%
0.50/0.01	90%	98%	100%	100%	99%	100%	100%	100%
0.25/0.05	43%	46%	64%	69%	46%	52%	66%	69%
	$n = 100$				$n = 100$			
0.25/0.01	52%	68%	95%	96%	77%	89%	96%	97%
0.50/0.01	69%	90%	100%	100%	96%	99%	100%	100%
0.25/0.05	34%	36%	55%	61%	36%	40%	58%	63%
	$n = 500$				$n = 500$			
0.25/0.01	29%	32%	87%	98%	41%	58%	96%	98%
0.50/0.01	33%	41%	98%	100%	64%	84%	100%	100%
0.25/0.05	25%	26%	31%	46%	26%	26%	35%	50%

The table reports hit ratios for Blockbuster applied to simulated data from the Degree-Corrected Stochastic Block Partial Correlation Network Model with Common Factors of Definition 5. The vertex-specific weight $\boldsymbol{\Theta}$ is drawn once from a power law with $x_m = 1$ and $\beta = 2.5$, and then kept constant over replications. The factor loading vector $\boldsymbol{q}$ is drawn once from a standard normal and kept constant over replications. The probabilities in the rows correspond to $n = 100$.

Panels A of Tables 1 through 3 show the results for the three models with network dependence parameter $\phi = 5$, whereas panels B show the results when $\phi = 50$. The first two rows of panel A in Table 1 show that Blockbuster performs quite well when $n = 50$.

If the probability of within-community edges is high as in the second row, the algorithm has a 93% hit ratio even with a small sample of $T = 50$. However, the results are worse if q is large relative to p as in the third row, even with a larger sample of $T = 10\,000$. It is clear that the results when $n = 100$ and $n = 500$ are worse than those with $n = 50$, except when the sample size is very large. For larger n, more samples are required to consistently estimate the covariance matrix and thus the community partition as the results of Theorem 1 suggest. The first, fourth and seventh rows of the last column of panel A in Table 1 show that with sufficient observations of $T = 10\,000$, higher n improves hit ratio. Comparing the results in panels B with A shows that the results improve across the board with stronger network dependence. Comparing Table 1 with Tables 2 and 3 shows that the performance of Blockbuster is quite similar for all three models, with performance slightly worse for the general degree and common factor models.

5 Real Activity Clustering

We use the methodology developed in the paper to study two panels of real activity growth measures for the U.S. and Europe. Our objective is to partition the series in the panels into communities characterised by a high degree of interdependence in real activity growth.

5.1 Community Detection in the U.S.

For our first application we consider a dataset constructed by Hamilton and Owyang (2011). The data consists of quarterly growth rates of payroll employment for the U.S. states (excluding Alaska and Hawaii) from the second quarter of 1956 to the fourth of 2007, which results in a panel of $n = 48$ time series over $T = 207$ periods. The data are seasonally adjusted and annualised. See Hamilton and Owyang (2011) for further details on the dataset.

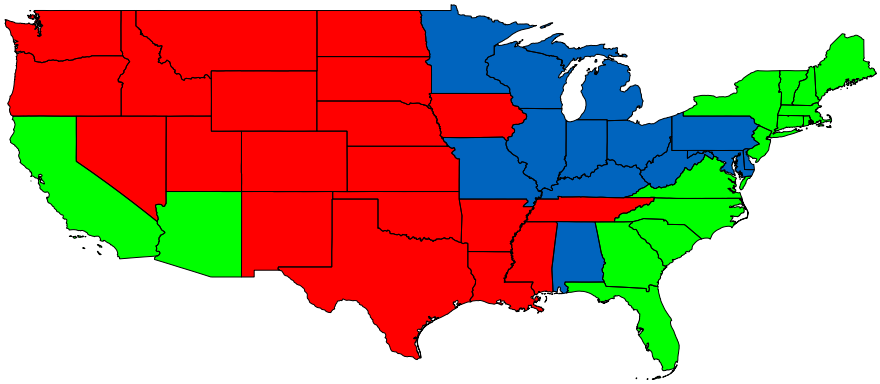
Hamilton and Owyang (2011) use their dataset to study business cycle synchronicity. We summarise their findings for the purpose of comparison, even though they deal with a

research question different from the one considered here. The authors introduce a Bayesian methodology based on a Markov-switching model to cluster states into communities with similar business cycle timing. They also propose a cross-validation procedure to estimate the number of communities and find evidence of three clusters. We focus on the results where they use a set of exogenous state level characteristics in addition to the employment growth rate. They find (see right column of figure 3 in Hamilton and Owyang, 2011) that the states are partitioned, roughly speaking, into a cluster of oil-producing and agricultural states, a cluster containing several East Coast states together with California, Arizona and Colorado, and a cluster containing the remaining states. Note that our results are not directly comparable with theirs, as we provide point estimates while they provide community membership posterior probability. Moreover, they use more information than us with the exogenous explanatory variables. It is also worth pointing out that in their results some states have a low posterior probability of belonging to any cluster, whereas in our algorithm every state is assigned to a community.

5.1.1 In-Sample Community Detection

We show the results of applying Blockbuster to the entire sample in Figure 2. A scree plot suggests that the data has one common factor, so we apply Algorithm 2 with $R = 1$. The number of communities is set to three as in Hamilton and Owyang (2011). Our results bear interesting similarities to their findings. The red community roughly matches the first cluster of Hamilton and Owyang (2011) and contains oil-producing and agricultural states. The green cluster contains East Coast states, California and Arizona, which roughly corresponds to their third cluster. Finally, the remaining states are Mid West states together with Alabama. Notice that the communities estimated by Blockbuster mostly form geographically contiguous areas even though no spatial information is used in the algorithm.

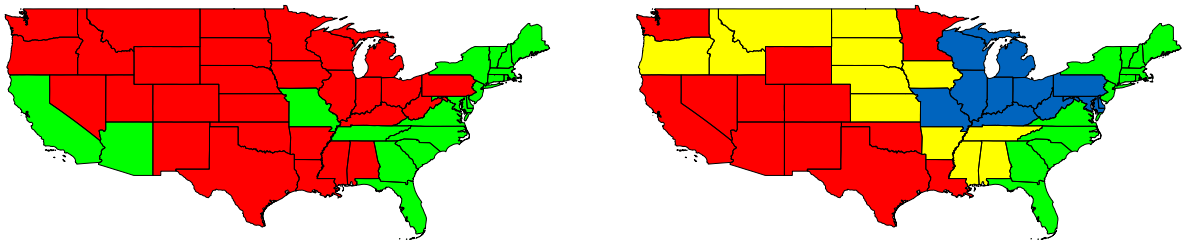
Figure 2: U.S. Real Activity Clustering ($k = 3$)



The figure displays the U.S. communities detected by Blockbuster when the number of communities is set to three.

We also run the Blockbuster algorithm with the number of communities set to two and four. The results are reported in Figure 3. When the number of communities is set to two, the algorithm partitions the U.S. into East Coast states together with California, Arizona, Missouri and Tennessee, and a residual cluster containing all remaining states. When the number of communities is set to four, in comparison to the baseline case, the oil-producing and agricultural states community gets split into into two separate clusters and California and Arizona are absorbed into the cluster containing oil-producing states. Note that the community corresponding to the East Coast is relatively stable across different choices of the number of clusters.

Figure 3: U.S. Real Activity Clustering ($k = 2$ and 4)



The figure displays the U.S. communities detected by Blockbuster when the number of communities is set to two (left) and four (right).

We report in-sample statistics on the Blockbuster partition in the left panel of Table 4. The table reports the average correlation between states in the same community and the average correlation between states in different communities. For k equal to 3 and 4, the intra-community average correlation is around 0.15 while the inter-community correlation has a much smaller magnitude. In the case of $k = 2$ the intra-community correlation is smaller. Detailed inspection of the results shows that this is driven by the low number of clusters. The average correlation between states in the green community is 0.13, but the correlation in the residual cluster is close to zero. We interpret this as the consequence of the residual cluster being the union of smaller communities. The table also reports the proportion of variance explained by the principal components ranging from 2 to $k + 1$, for k ranging from 2 to 4. The first principal component in the panel explains 50% of the total variance. The proportion of the variance explained by the principal components associated with the communities is sizeable and lies between 12% and 19% of the total variation of the panel.

Table 4: U.S. Real Activity Clustering

In-Sample Analysis				Out-of-Sample Analysis			
k	2	3	4	k	2	3	4
Avg Cor Within Blocks	0.079	0.134	0.164	KL Loss	285.0	282.0	286.4
Avg Cor Between Blocks	-0.098	-0.086	-0.074	Rel Gain Sample Cov	2.046	3.077	1.590
Prop of Explained Var	12.294	16.272	19.453	Rel Gain POET($k+1$)	4.346	-0.631	-4.465
				Rel Gain LW	17.689	18.556	17.307
k-means Loss	23.511	20.151	17.348	k-means Loss	23.497	29.461	28.018

The table reports summary in-sample and out-of-sample statistics for the Blockbuster results for different choices of the number of communities k. The in-sample statistics are the average correlation between states in the same community, the average correlation between states in different communities, the proportion of variance explained by the principal components ranging from 2 to $k + 1$ and the loss of the k-means algorithm. The out-of-sample statistics are the Kullback-Leibler loss of the Blockbuster covariance estimator, the relative gain of the Blockbuster covariance estimator with respect to the sample covariance estimator, the relative gain with respect to the POET covariance estimator, the relative gain with respect to the Ledoit and Wolf shrinkage covariance estimator and the out-of-sample k–means loss (defined in the text).

5.1.2 Out-of-Sample Community Validation

We carry out an out-of-sample validation exercise to assess the performance of the Blockbuster community detection algorithm on the basis of the covariance regularisation procedure described in Algorithm 3. The validation exercise is designed as follows. We split the

entire sample into an in-sample and an out-of-sample period, each containing 50% of the entire sample. We estimate the Blockbuster covariance matrix in-sample and then evaluate the estimator out-of-sample by predicting the out-of-sample sample covariance matrix of the panel. Since there is evidence of one factor in the data, we apply the Blockbuster covariance estimator which accounts for one factor. We measure the precision of the forecast on the basis of the Kullback-Leibler loss proposed by Stein (Stein, 1956; Pourahmadi, 2013). We run the procedure with the number of communities ranging from two to four. We compare the performance of the Blockbuster covariance estimator with the (standard) sample covariance as well as the POET covariance estimator (Fan *et al.*, 2013), which is a regularisation procedure based on an approximate factor model representation of the panel, and the Ledoit and Wolf shrinkage covariance estimator (Ledoit and Wolf, 2004).

We report the out-of-sample results in the right panel of Table 4. The table reports the Kullback-Leibler loss of the Blockbuster covariance estimator as well as its relative gain over the sample covariance, POET and shrinkage covariance estimators. The table shows that when the number of communities is set to two the Blockbuster estimator achieves the best out-of-sample performance, and that it performs favourably relative to the set of alternative estimators considered.

As an additional out-of-sample diagnostic, Table 4 also reports what we call the out-of-sample k-means loss. This is defined as the loss of the k-means algorithm in the out-of-sample period of the k-means partition estimated in the in-sample period. It is interesting to notice that k equal to two is the only instance in which the out-of-sample k-means loss does not inflate considerably with respect to in-sample k-means loss.

5.2 Community Detection in Europe

For our second application we construct a dataset using data from Eurostat. The data consists of yearly growth rates of the gross regional product (GRP) for the first-level European Nomenclature of Territorial Units for Statistics (NUTS 1) regions of Europe (excluding Iceland and Turkey) from 2000 to 2013, which results in a panel of $n = 99$ time

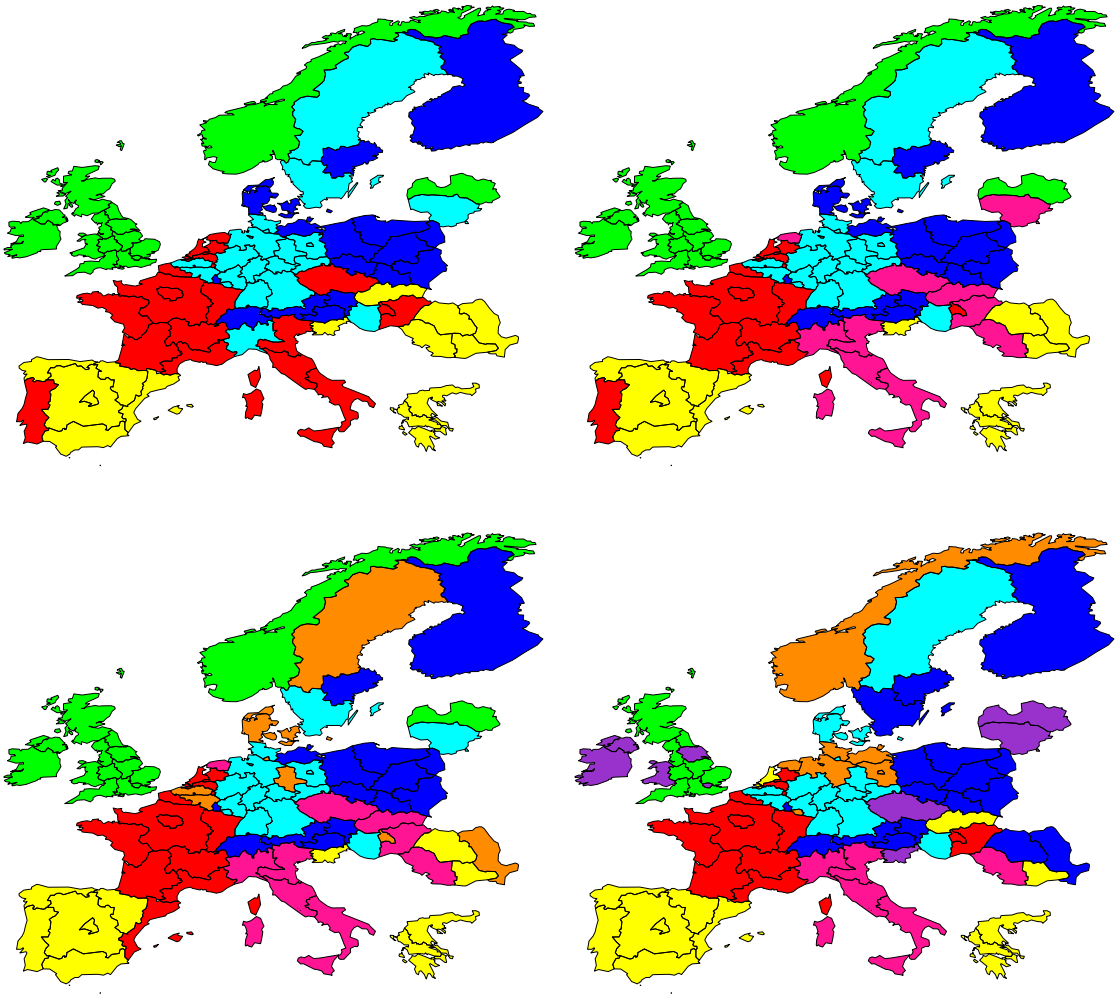
series over $T = 13$ periods. The GRP data is in purchasing power standard units. The analysis of this panel is clearly more challenging, as the large sample theory developed in the paper is not applicable in this setting. However, simulations suggest that even when T is much smaller than n the Blockbuster algorithm performs satisfactorily provided that the probability of the intra-community linkages is sufficiently strong relative to the probability of inter-community linkages. Due to size of the sample, we only carry out an in-sample analysis.

5.2.1 In-Sample Community Detection

We show the results for Europe in Figure 4. A scree plot suggests that the data has one common factor as for the U.S., so we use Algorithm 2 setting the number of factors to one. As we have little guidance from the literature on how to choose the number of clusters, and due to the fact that the dataset exhibits a higher degree of heterogeneity than the U.S. dataset, we let the number of communities range from five to eight. In the following discussion we focus on the clustering results when then number of communities is set to seven. Blockbuster finds mostly geographically contiguous areas. The four largest European economies (France, Germany, Italy and the UK) essentially form clusters of their own. In some cases, these clusters also include smaller neighbouring economies that have strong economic ties with these large economies. For instance, Ireland is in the UK cluster and Albania is in the Italian cluster.[8] On the other hand, smaller economies form larger clusters made up of more regions. For instance, there is evidence of a Southern European cluster, containing Greece, Portugal and Spain (with the exception of Catalonia), as well as a Central-Eastern European cluster containing (among others) Austria, Finland, Poland, Romania, and parts of Sweden and Switzerland.

[8]Italy is the main trade partner of Albania for most of the years in the sample.

Figure 4: Europe Real Activity Clustering ($k = 5, 6, 7$ and 8)



The figure displays the Europe communities detected by Blockbuster when the number of communities ranges from five to eight.

We report statistics on the Blockbuster partition in Table 5. The table reports the average correlation between regions in the same community and the average correlation between regions in different communities. In comparison to the U.S. results, the European panel exhibits a higher degree of intra-community correlation. Overall, the clustering results provide interesting insight into the community structure of Europe, despite the small sample size.

Table 5: Europe Real Activity Clustering

In-Sample Analysis

k	5	6	7	8
Avg Cor Within Blocks	0.476	0.495	0.491	0.511
Avg Cor Between Blocks	-0.109	-0.086	-0.065	-0.064
Prop of Explained Var	27.900	30.035	31.492	32.766
k-means Loss	36.660	38.547	39.423	39.365

The table reports the average correlation between regions in the same community, the average correlation between regions in different communities for different choices of the number communities k, the proportion of variance explained by the principal components ranging form 2 to k and the loss of k–means algorithm.

6 Conclusion

In this paper we consider the problem of community detection in partial correlation network models. We begin by introducing a class of Gaussian graphical models in which the underlying network structure is random and determined by a latent random graph (Chung and Lu, 2006; van der Hofstad, 2015). We use this framework to introduce a class of graphical models with a community structure by assuming that the latent random graph is a degree-corrected stochastic block model (Holland *et al.*, 1983; Karrer and Newman, 2011). The degree-corrected stochastic block model is a random graph in which vertices are partitioned into different communities and the probability of an edge between two vertices depends on the communities the vertices belong to and a vertex-specific weight.

A natural question that arises in this framework is how to detect communities from a random sample of observations. We propose an algorithm called Blockbuster to tackle this task. The algorithm uses the k-means clustering procedure on the top eigenvectors of the estimated covariance matrix to detect communities. We study the large sample properties of the algorithm and establish consistency of the procedure when the number of variables n and observations T is large. A simulation study shows that the method performs well in finite samples.

A number of extensions of the baseline methodology are introduced. First we extend the baseline methodology to the case where pervasive factors affect the dependence structure of the data. Second, we introduce a covariance regularisation procedure based on

Blockbuster which is motivated by the special block covariance structure of the model.

We apply the methodology to study real activity clustering in the U.S. and Europe. The objective of the application is to partition the two panels into communities characterised by a high degree of interdependence in real activity growth. For the U.S. we employ a dataset of quarterly employment growth rates for the states from 1956 to 2007 constructed by Hamilton and Owyang (2011), whereas for Europe we construct a dataset of yearly GRP growth rates for the NUTS 1 regions from 2000 until 2013 using data from Eurostat. Results show that Blockbuster detects a meaningful partition of the series in the panels. Interestingly, the procedure performs well in the European panel despite the short sample size. In the U.S. panel we design an out-of-sample validation exercise and show that the covariance regularisation procedure based on Blockbuster improves covariance prediction compared to a number of alternative procedures.

References

Abbe, E., Bandeira, A. S., and Hall, G. (2014). Exact Recovery in the Stochastic Block Model. *ArXiv e-prints*.

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

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Barigozzi, M. and Brownlees, C. (2013). Nets: Network estimation for time series. Working Papers 723, Barcelona Graduate School of Economics.

Bhatia, R. (1997). *Matrix Analysis*. Graduate Texts in Mathematics. Springer New York.

Chaudhuri, K., Chung, F., and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *JMLR Workshop and Conference Proceedings*, **23**.

Chen, A., Amini, A. A., Bickel, P. J., and Levina, E. (2012). Fitting community models to large sparse networks. *CoRR*, **abs/1207.2340**.

Choi, I., Kim, D., Kim, Y. J., and Kwark, N.-S. (2014). A multilevel factor model: Identification, asymptotic theory and applications. Technical report.

Chung, F. and Lu, L. (2006). *Complex Graphs and Networks*. American Mathematical Society, Providence.

Chung, F. and Radcliffe, M. (2011). On the spectra of general random graphs. *Electr. J. Comb.*, **18**.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, **7**(1), 1–46.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

Diebold, F. and Yilmaz, K. (2013). On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms. *Journal of Econometrics*, page forthcoming.

Donath, W. E. and Hoffman, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.*, **17**(5), 420–425.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. *Journal of the Royal Statistical Society, Series B*.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, **23**(2), 298–305.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, **486**, 75–174.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Science*, **99**, 7821–7826.

Goldstein, H. and Browne, W. (2002). *Multilevel factor analysis modelling using Markov Chain Monte Carlo (MCMC) estimation*, pages 225 – 243. Lawrence Erlbaum, New Jersey.

Hamilton, J. D. and Owyang, M. T. (2011). The propagation of regional recessions. Working Paper 16657, National Bureau of Economic Research.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, **5**(2), 109 – 137.

Joseph, A. and Yu, B. (2013). Impact of regularization on Spectral Clustering. *ArXiv e-prints*.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *ArXiv e-prints*, **83**(1), 016107.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Ledoit, O. and Wolf, M. (2004). A Well-Conditioned Estimator for Large–Dimensional Covariance Matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *ArXiv Mathematics e-prints*.

Moench, E., Ng, S., and Potter, S. (2009). Dynamic hierarchical factor models. Technical report, Federal Reserve Bank of New York Staff Reports.

Mossel, E., Neeman, J., and Sly, A. (2012). Stochastic Block Models and Reconstruction. *ArXiv e-prints*.

Newman, M. E. J. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev.*, **74**(3), 036104.

Newman, M. E. J. (2006b). From the Cover: Modularity and community structure in networks. *Proceedings of the National Academy of Science*, **103**, 8577–8582.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press.

Oliveira, I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *ArXiv e-prints*.

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, **104**(486), 735–746. PMID: 19881892.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Inc.

Qin, T. and Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. *CoRR*, **abs/1309.4111**.

Rohe, K., Chatterjee, S., and Yu, B. (2010). Spectral clustering and the high-dimensional stochastic blockmodel. *ArXiv e-prints*.

Sarkar, P. and Bickel, P. (2013). Role of Normalization in Spectral Clustering for Stochastic Blockmodels. *ArXiv e-prints*.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, pages 157–164. University of California Press.

Stewart, G. and Sun, J. (1990). *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press.

van der Hofstad, R. (2015). *Random Graphs and Complex Networks*.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *ArXiv e-prints*.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, **17**(4), 395–416.

A Proofs

Proof of Lemma 1. See Lemmas 3.2 and 3.3 of Qin and Rohe (2013) and let $\tau = 0$. Their proof is for the matrix $-\mathcal{L} + \mathbf{I}_n$ which is easily translatable to our case. In the case of $\Theta = \mathbf{I}_n$, see Lemma 3.1 in Rohe *et al.* (2010), as the proof of Qin and Rohe (2013) is based on the normalisation of θ_i such that they sum to unity within communities. □

Proof of Corollary 1. The results follow from the fact that the spectra of $\mathcal{K}$ and $\mathcal{L}$ are closely related, as $\mathcal{K} = (1/\sigma^2)\mathbf{I}_n + (\phi/\sigma^2)\mathcal{L}$. □

Proof of Theorem 2. We first consider the normalised Laplacian $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, where $\mathbf{A}$ and $\mathbf{D}$ are the adjacency and degree matrices, and its concentration around its population counterpart $\mathcal{L}$. Let $\bar{d}_{min} = [\mathcal{D}]_{ii}$ be the minimum expected degree of $\mathcal{G}$ and apply Theorem 3.1 from Oliveira (2009), which is reproduced in Appendix B for convenience. For any constant $c > 0$ there exits another constant $C(c) > 0$ such that if $\bar{d}_{min} \geq C(c) \log n$, then for all $n^{-c} \leq \eta \leq 1/2$

$$\mathbb{P}\left(\|\mathbf{L} - \mathcal{L}\| \leq 14\sqrt{\frac{\log(4n/\eta)}{\bar{d}_{min}}} \right) \geq 1 - \eta.$$

In our case $\bar{d}_{min} = \Omega(n\rho_n)$, since all the elements of $\mathbf{B}$ are of the same asymptotic order ρ_n. We hence have

$$\|\mathbf{L} - \mathcal{L}\| = O\left(\sqrt{\frac{\log n}{\bar{d}_{min}}}\right) = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right)$$

with high probability. By the definition of $\mathbf{K}$ and $\mathcal{K}$, this implies

$$\|\mathbf{K} - \mathcal{K}\| = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right). \tag{10}$$

We next bound the spectral norm of the difference between $\mathbf{\Sigma}$ and $\widehat{\mathbf{\Sigma}}$ with Theorem 5.39 and Remark 5.40 from Vershynin (2010) for sub-Gaussian random variables. A random variable $\boldsymbol{Y}$ is said to be sub-Gaussian when $\mathbb{P}(|\boldsymbol{Y}| > t) \leq e^{1-t^2/C^2}$ for all $t \geq 0$, where $C > 0$ is a constant, and the sub-Gaussian norm of $\boldsymbol{Y}$ is defined as $\|\boldsymbol{Y}\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2} \left(\mathbb{E}|\boldsymbol{Y}|^p\right)^{1/p}$, see Vershynin (2010) for more details. In particular, Gaussian random variables are sub-Gaussian. As the $\boldsymbol{Y}_t$ are centred, independent Gaussian random vectors in $\mathbb{R}^n$ with second moment matrix $\mathbf{\Sigma}$, we have, for every $s \geq 0$

$$\mathbb{P}\left(\left\| \mathbf{\Sigma} - \widehat{\mathbf{\Sigma}} \right\| \leq \max(\delta, \delta^2) \|\mathbf{\Sigma}\| \right) \geq 1 - \exp\left(-cs^2\right)$$

with $\delta = C\sqrt{\frac{n}{T}} + \frac{s}{\sqrt{T}}$, and $C > 0$ and $c > 0$ depend only on the maximum of the sub-Gaussian norms of the isotropic random vectors $\mathbf{\Sigma}^{-1/2}\boldsymbol{Y}_t$, which are standard Gaussian

and hence bounded by absolute constants. Now let $s = \sqrt{C'n}$ where C' is a constant. Then for C' sufficiently large

$$\mathbb{P}\left(\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\| \leq \max(\delta, \delta^2)\|\mathbf{\Sigma}\|\right) \geq 1 - \exp\left(-cC'n\right) \geq 1 - \exp\left(-n\right),$$

where $\delta = \left(C + \sqrt{C'}\right)\sqrt{\frac{n}{T}}$. Hence

$$\mathbb{P}\left(\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\| \leq C''\sqrt{\frac{n}{T}}\right) \geq 1 - \exp\left(-n\right)$$

with $C'' = \left(C + \sqrt{C'}\right)\|\mathbf{\Sigma}\| > 0$ a constant. Note that $\|\mathbf{\Sigma}\|$ is bounded due to the sparsity of the underlying graph $\mathcal{G}$. To see this, note that the top eigenvalue of $\mathbf{\Sigma}$ is the reciprocal of the smallest eigenvalue of $\mathbf{K}$. Weyl's inequality and (10) establish that the eigenvalues of $\mathbf{K}$ converge with high probability to those of $\mathcal{K}$ as n grows large, under our assumption on p. Corollary 4 shows that $\lambda_i(\mathcal{K}) \in [1/\sigma^2, (1+\phi)/\sigma^2]$. In particular, as $\mathcal{L}$ is non-negative definite and its rows sum to zero, $\lambda_1(\mathcal{L}) = 0$ so that $\lambda_1(\mathcal{K}) = 1/\sigma^2$. Hence with high probability

$$\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\| = O\left(\sqrt{\frac{n}{T}}\right). \tag{11}$$

To turn this into a statement about the concentration matrix, note that we may write

$$\left\|\widehat{\mathbf{K}} - \mathbf{K}\right\| = \left\|\widehat{\mathbf{K}}\left(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right)\mathbf{K}\right\| \leq \left\|\widehat{\mathbf{K}}\right\|\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\|\left\|\mathbf{K}\right\|$$

by the Cauchy-Schwarz inequality. Applying Weyl's inequality with (11) implies that the eigenvalues of $\widehat{\mathbf{\Sigma}}$ converge to those of $\mathbf{\Sigma}$ with high probability, provided that $n/T \to 0$. As $\widehat{\mathbf{\Sigma}}$ is invertible, this implies $\left\|\widehat{\mathbf{K}}\right\| \to \left\|\mathbf{K}\right\|$ with high probability under the same conditions. We showed in the previous paragraph that the eigenvalues of $\mathbf{K}$ converge to those of $\mathcal{K}$. Taken together with the results of Corollary 1, this implies $\|\mathbf{K}\| \to (1+\phi)/\sigma^2$ with high probability. Therefore

$$\left\|\widehat{\mathbf{K}} - \mathbf{K}\right\| = O\left(\left\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\right\|\right) = O\left(\sqrt{\frac{n}{T}}\right).$$

Then we have

$$\left\|\widehat{\mathbf{K}} - \mathcal{K}\right\| \leq \left\|\widehat{\mathbf{K}} - \mathbf{K}\right\| + \left\|\mathbf{K} - \mathcal{K}\right\| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right) \tag{12}$$

by the triangle inequality and (10).

We next apply the Davis-Kahan theorem (Davis and Kahan (1970), Stewart and Sun (1990), Bhatia (1997)) to bound the angles between the subspaces of the bottom k eigenvectors of $\widehat{\mathbf{K}}$ and $\mathcal{K}$, see Appendix B for a statement of the theorem. We need to choose the interval S such that it contains the k smallest eigenvalues of both $\widehat{\mathbf{K}}$ and $\mathcal{K}$ to ensure

that the subspaces in question are comparable. Weyl's inequality yields

$$\left|\lambda_j(\widehat{\mathbf{K}}) - \lambda_j(\mathcal{K})\right| \leq \left\|\widehat{\mathbf{K}} - \mathcal{K}\right\| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right), \quad \forall j = 1, \ldots, n$$

so that the eigenvalues converge with high probability under our assumption on p. We follow Sarkar and Bickel (2013) and define $\gamma = (\lambda_{k+1}(\mathcal{K}) - \lambda_k(\mathcal{K}))/4 > 0$, $a = \lambda_1(\mathcal{K}) - 2\gamma$, $b = \lambda_k(\mathcal{K}) + 2\gamma$ and let $S = [a, b]$. The interval S contains only the k first eigenvalues of $\mathcal{K}$ by construction. Notice that for n large enough, $\gamma > \left\|\widehat{\mathbf{K}} - \mathcal{K}\right\|$. Hence S also contains only the first k eigenvalues of $\widehat{\mathbf{K}}$ with high probability. Finally, notice that the eigenvalues of interest are isolated from the rest of the spectrum in the sense that the set $(a - \gamma, a + \gamma) \cup (b - \gamma, b + \gamma)$ contains none of the eigenvalues of $\mathcal{K}$ or $\widehat{\mathbf{K}}$ with high probability. We define

$$\delta = \min\{|\lambda - s| : \lambda \in \lambda_{\overline{S}}(\mathcal{K}), s \in S\}.$$

It follows that $\delta = \lambda_{k+1}(\mathcal{K}) - b = (\lambda_{k+1}(\mathcal{K}) - \lambda_k(\mathcal{K}))/2$. We have $\delta \in (0, \phi/(2\sigma^2))]$ by Corollary 1. Let $\sin\Phi\left(\widehat{\mathbf{U}}, \mathcal{U}\right)$ be the $k \times k$ diagonal matrix of the sines of the principal angles between the column spaces of $\widehat{\mathbf{U}}$ and $\mathcal{U}$. The Davis-Kahan theorem gives

$$\left\|\sin\Phi\left(\widehat{\mathbf{U}}, \mathcal{U}\right)\right\| \leq \frac{\left\|\widehat{\mathbf{K}} - \mathcal{K}\right\|}{\delta} = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right).$$

To convert this into a direct statement about the eigenvector matrices, note that

$$\left\|\sin\Phi\left(\widehat{\mathbf{U}}, \mathcal{U}\right)\right\|_F = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)$$

as $\sin\Phi\left(\widehat{\mathbf{U}}, \mathcal{U}\right)$ is a $k \times k$ diagonal matrix. Apply the singular value decomposition to $\mathcal{U}'\widehat{\mathbf{U}}$ and decompose it as $\mathbf{F}\mathbf{\Psi}\mathbf{G}'$, where $\mathbf{F}$ and $\mathbf{G}$ orthonormal $k \times k$ matrices and $\mathbf{\Psi}$ a diagonal matrix of the singular values. We can then directly apply the arguments made in Appendix B of Rohe *et al.* (2010) to obtain

$$\left\|\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}\right\|_F = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right) \tag{13}$$

with high probability, where $\mathcal{O} \equiv \mathbf{F}\mathbf{G}'$ is a $k \times k$ orthonormal rotation matrix that depends on $\widehat{\mathbf{U}}$ and $\mathcal{U}$. Note that we furthermore require $n/T \to 0$ to guarantee that the eigenvalues converge, to ensure that $\widehat{\mathbf{U}}$ and $\mathcal{U}$ contain comparable eigenvectors.

We next bound $\left\|\widehat{\mathbf{X}} - \mathcal{X}\mathcal{O}\right\| = \left\|\widehat{\mathbf{N}}\widehat{\mathbf{U}} - \mathcal{N}\mathcal{U}\mathcal{O}\right\|$. For this, we first consider the convergence of $\widehat{\mathbf{N}}$ to $\mathcal{N}$. Notice that (13) implies

$$\left\|[\widehat{\mathbf{U}}]_{\bullet i} - \mathcal{U}[\mathcal{O}]_{\bullet i}\right\| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)$$

for all $i = 1, \ldots, k$, by the definition of the operator norm and the fact that the Frobenius norm bounds it. It then follows that

$$\left|[\widehat{\mathbf{U}}]_{ji} - [\mathcal{U}]_{j\bullet}[\mathcal{O}]_{\bullet i}\right| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right).$$

As the rows of $\widehat{\mathbf{U}}$ and $\mathcal{U}$ have only k elements

$$\sqrt{\sum_{i=1}^{k}\left([\widehat{\mathbf{U}}]_{ji} - [\mathcal{U}]_{j\bullet}[\mathcal{O}]_{\bullet i}\right)^2} = \left\|[\widehat{\mathbf{U}}]_{j\bullet} - [\mathcal{U}]_{j\bullet}\mathcal{O}\right\| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)$$

which bounds the row norms of $\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}$. Notice that $[\widehat{\mathbf{N}}^{-1} - \mathcal{N}^{-1}]_{jj} = \left\|[\widehat{\mathbf{U}}]_{j\bullet}\right\| - \left\|[\mathcal{U}]_{j\bullet}\mathcal{O}\right\|$ by the definition of $\widehat{\mathbf{N}}$ and $\mathcal{N}$, and the fact that the Euclidean norm is invariant under orthonormal transformations. Then by the triangle inequality, we have $\left|\left\|[\widehat{\mathbf{U}}]_{j\bullet}\right\| - \left\|[\mathcal{U}]_{j\bullet}\mathcal{O}\right\|\right| \leq \left\|[\widehat{\mathbf{U}}]_{j\bullet} - [\mathcal{U}]_{j\bullet}\mathcal{O}\right\|$ which implies

$$\left\|\widehat{\mathbf{N}}^{-1} - \mathcal{N}^{-1}\right\| = O\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)$$

as $\widehat{\mathbf{N}}^{-1} - \mathcal{N}^{-1}$ is a diagonal matrix. Finally, note that $\left\|\widehat{\mathbf{N}} - \mathcal{N}\right\| = \left\|\widehat{\mathbf{N}}(\mathcal{N}^{-1} - \widehat{\mathbf{N}}^{-1})\mathcal{N}\right\| \leq \left\|\widehat{\mathbf{N}}\right\|\left\|(\mathcal{N}^{-1} - \widehat{\mathbf{N}}^{-1})\right\|\|\mathcal{N}\|$. Furthermore, $\left\|\widehat{\mathbf{N}}\right\|$ approaches $\|\mathcal{N}\|$ as $n \to \infty$. Hence we have

$$\left\|\widehat{\mathbf{N}} - \mathcal{N}\right\| = O\left(\|\mathcal{N}\|\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)\right)$$

with high probability. Notice that $\|\mathcal{N}\| = 1/\min_j\|[\mathcal{U}]_{j\bullet}\|$, the reciprocal of the shortest row length of $\mathcal{U}$. Let $l = \operatorname{argmin}_j\|[\mathcal{U}]_{j\bullet}\|$. It then follows that

$$\|\mathcal{N}\| = \left(\frac{\sum_j \theta_j \mathbf{1}_{z_l, z_j}}{\theta_l}\right)^{1/2} = O\left(\sqrt{n_{z_l}}\right) = O\left(\sqrt{n}\right)$$

from (7) as θ_i are constants. Hence, we have

$$\|\mathbf{N} - \mathcal{N}\| = O\left(\sqrt{n}\left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}}\right)\right) = O\left(\frac{n}{\sqrt{T}} + \sqrt{\frac{\log n}{\rho_n}}\right). \tag{14}$$

It remains to put the pieces together. We may write

$$\left\|\widehat{\mathbf{N}}\widehat{\mathbf{U}} - \mathcal{N}\mathcal{U}\mathcal{O}\right\| = \left\|(\widehat{\mathbf{N}} - \mathcal{N})\widehat{\mathbf{U}} + \mathcal{N}\widehat{\mathbf{U}} - \mathcal{N}\mathcal{U}\mathcal{O}\right\|$$

$$= \left\|(\widehat{\mathbf{N}} - \mathcal{N})(\widehat{\mathbf{U}} + \mathcal{U}\mathcal{O} - \mathcal{U}\mathcal{O}) + \mathcal{N}(\widehat{\mathbf{U}} + \mathcal{U}\mathcal{O} - \mathcal{U}\mathcal{O}) - \mathcal{N}\mathcal{U}\mathcal{O}\right\|$$

$$= \left\| (\widehat{\mathbf{N}} - \mathcal{N})(\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}) + (\widehat{\mathbf{N}} - \mathcal{N})\mathcal{U}\mathcal{O} + \mathcal{N}(\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}) \right\|.$$

The triangle inequality then implies

$$\left\| \widehat{\mathbf{N}}\widehat{\mathbf{U}} - \mathcal{N}\mathcal{U}\mathcal{O} \right\| \leq \left\| (\widehat{\mathbf{N}} - \mathcal{N})(\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}) \right\| + \left\| (\widehat{\mathbf{N}} - \mathcal{N})\mathcal{U}\mathcal{O} \right\| + \left\| \mathcal{N}(\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}) \right\|.$$

Notice that $\|\mathcal{U}\mathcal{O}\|$ is a constant, as the matrix $\mathcal{U}\mathcal{O}$ has k columns of unit length. This fact, taken together with (13) and (14) yields

$$\left\| \widehat{\mathbf{X}} - \mathcal{X}\mathcal{O} \right\| = O\left(\frac{n}{\sqrt{T}} + \sqrt{\frac{\log n}{\rho_n}} \right) + O\left(\sqrt{n} \left(\sqrt{\frac{n}{T}} + \sqrt{\frac{\log n}{n\rho_n}} \right) \right) = O\left(\frac{n}{\sqrt{T}} + \sqrt{\frac{\log n}{\rho_n}} \right)$$

with high probability. Note, as before, that we require that n/T tends to zero to guarantee that the eigenvalues converge, to ensure that $\widehat{\mathbf{X}}$ and $\mathcal{X}$ contain comparable eigenvectors. □

Proof of Lemma 2. This follows from a straightforward extension of the proof of Lemma 3.2 from Rohe *et al.* (2010). Furthermore, see the proof of Theorem 4.4 of Qin and Rohe (2013). □

Proof of Theorem 1. The proof follows the approach in Rohe *et al.* (2010), Sarkar and Bickel (2013) and Qin and Rohe (2013) closely. Note that as $\mathcal{C} \in \mathcal{M}(n, k)$

$$\left\| \widehat{\mathbf{X}} - \widehat{\mathbf{C}} \right\|_F \leq \left\| \widehat{\mathbf{X}} - \mathcal{C} \right\|_F \tag{15}$$

and by the triangle inequality

$$\left\| \widehat{\mathbf{C}} - \mathcal{C} \right\|_F \leq \left\| \widehat{\mathbf{C}} - \widehat{\mathbf{X}} \right\|_F + \left\| \widehat{\mathbf{X}} - \mathcal{C} \right\|_F \leq 2\left\| \widehat{\mathbf{X}} - \mathcal{C} \right\|_F, \tag{16}$$

where the last inequality follows from (15).

We now bound the cardinality of the set of misclustered vertices. To this end, it suffices to bound the size of the set $\mathscr{M}$. Notice that

$$|\mathscr{M}| = \sum_{i \in \mathscr{M}} 1 \leq 2 \sum_{i \in \mathscr{M}} \left\| [\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O} \right\|^2$$

$$\leq 2\left\| \widehat{\mathbf{C}} - \mathcal{C} \right\|_F^2 \leq 8\left\| \widehat{\mathbf{X}} - \mathcal{C} \right\|_F^2 = 8\left\| \widehat{\mathbf{X}} - \mathcal{X}\mathcal{O} \right\|_F^2.$$

The first inequality follows from the definition of $\mathscr{M}$ in (6). The second follows from the fact that $\sum_{i=1}^n \left\| [\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O} \right\|^2 = \left\| \widehat{\mathbf{C}} - \mathcal{C} \right\|_F^2$ and that the sum only includes a subset of the vertices. The final inequality follows from (16). Theorem 2 implies

$$\left\| \widehat{\mathbf{X}} - \mathcal{X}\mathcal{O} \right\|_F = O\left(\frac{n}{\sqrt{T}} + \sqrt{\frac{\log n}{\rho_n}} \right)$$

as $\left\|\widehat{\mathbf{X}} - \mathcal{X}\mathcal{O}\right\|_F \leq \sqrt{k}\left\|\widehat{\mathbf{X}} - \mathcal{X}\mathcal{O}\right\|$ for a rank k matrix. It follows that

$$|\mathscr{M}| = O\left(\frac{n^2}{T} + \frac{\log n}{\rho_n} + 2n\sqrt{\frac{\log n}{T\rho_n}}\right) = O\left(\frac{n^2}{T} + \frac{\log n}{\rho_n}\right)$$

and

$$\frac{|\mathscr{M}|}{n} = O\left(\frac{n}{T} + \frac{\log n}{n\rho_n}\right).$$

It should be noted that the same result is obtained if the unnormalised eigenvectors $\widehat{\mathbf{U}}$ are used in (15). The definition of the set $\mathscr{M}$ would change to $\mathscr{M} = \{i : \left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\boldsymbol{\mu}\mathcal{O}\right\| \geq \sqrt{1/(2n_*)}\}$ where $n_* = O(n)$ is the size of the largest community (see Lemma 3.2 of Rohe *et al.* (2010)). The final result would thus be

$$|\mathscr{M}| \leq 2n_* \sum_{i\in\mathscr{M}} \left\|[\widehat{\mathbf{C}}]_{i\bullet} - [\mathbf{Z}]_{i\bullet}\mathbf{V}\mathcal{O}\right\|^2 \leq 8n_* \left\|\widehat{\mathbf{U}} - \mathcal{U}\mathcal{O}\right\|_F^2 = O\left(\frac{n^2}{T} + \frac{\log n}{\rho_n}\right)$$

as before. □

Proof of Lemma 3. We consider a perturbation of the matrix $\mathcal{K}_{\boldsymbol{\epsilon}}^{-1}$, the spectrum of which we know from Corollary 1, and wish to find the eigenvalues and eigenvectors of the matrix $\mathcal{K}^{-1} = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R} \boldsymbol{q}_r\boldsymbol{q}_r'$. We proceed by induction and consider $\mathcal{K}_1^{-1} = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{q}_1\boldsymbol{q}_1'$ first. We then find the eigenvectors of $\mathcal{K}^{-1} \equiv \mathcal{K}_R^{-1} = \mathcal{K}_{R-1}^{-1} + \boldsymbol{q}_R\boldsymbol{q}_R'$ given those of $\mathcal{K}_{R-1}^{-1}$.

Notice that the eigenvectors of $\mathcal{K}_{\boldsymbol{\epsilon}}$, $\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}})$, form a basis in $\mathbb{R}^n$. We may then write $\boldsymbol{q}_1$ as a linear combination

$$\boldsymbol{q}_1 = \sum_{i=1}^{n} \gamma_{i,1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}),$$

where $\gamma_{i,1}$ are scalars. The assumption that $\boldsymbol{q}_1'[\mathcal{U}]_{\bullet i} = 0$ then amounts to $\gamma_{i,1} = 0$ for $i = 1, \ldots, k$, so that

$$\boldsymbol{q}_1 = \gamma_{k+1,1}\boldsymbol{u}_{k+1}(\mathcal{K}_{\boldsymbol{\epsilon}}) + \ldots + \gamma_{n,1}\boldsymbol{u}_n(\mathcal{K}_{\boldsymbol{\epsilon}}).$$

We guess and verify the eigenvectors and eigenvalues of $\mathcal{K}_1$ using the eigenvalue equation

$$\mathcal{K}_1^{-1}\boldsymbol{u}_i(\mathcal{K}_1) = \lambda_i(\mathcal{K}_1)^{-1}\boldsymbol{u}_i(\mathcal{K}_1)$$

and ensure they are mutually orthogonal. We begin with the bottom eigenvalue and eigenvector of $\mathcal{K}_1$ and guess $\lambda_1(\mathcal{K}_1)^{-1} = \lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1} + \|\boldsymbol{q}_1\|^2$ and $\boldsymbol{u}_1(\mathcal{K}_1) = \boldsymbol{q}_1/\|\boldsymbol{q}_1\|$. We have

$$\mathcal{K}_1^{-1}\boldsymbol{q}_1 = \left(\mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{q}_1\boldsymbol{q}_1'\right)\boldsymbol{q}_1 = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{q}_1 + \|\boldsymbol{q}_1\|^2\boldsymbol{q}_1 = \left(\lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1} + \|\boldsymbol{q}_1\|^2\right)\boldsymbol{q}_1,$$

where the last equality follows from

$$\mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{q}_1 = \sum_{i=k+1}^{n} \gamma_{i,1}\mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \sum_{i=k+1}^{n} \gamma_{i,1}\lambda_i(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{q}_1$$

as $\lambda_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})$ for $i = k+1, \ldots, n$ from Corollary 1. Dividing through by $\|\boldsymbol{q}_1\|$ we have the eigenvector.

Next we consider the second to the $k+1$-th bottom eigenvalues and eigenvectors of $\mathcal{K}_1$. We postulate $\lambda_{1+i}(\mathcal{K}_1)^{-1} = \lambda_i(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}$ and $\boldsymbol{u}_{1+i}(\mathcal{K}_1) = \boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}})$ for $i = 1, \ldots, k$. Consider

$$\mathcal{K}_1^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) + \boldsymbol{q}_1\boldsymbol{q}_1'\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \lambda_i(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}})$$

for $i = 1, \ldots, k$, which follows from the fact that the bottom k eigenvectors of $\mathcal{K}_{\boldsymbol{\epsilon}}^{-1}$ are orthogonal to $\boldsymbol{q}_1$.

It remains to find the last $n-k-1$ eigenvectors of $\mathcal{K}_1$. Let

$$\boldsymbol{u}_n(\mathcal{K}_1) = \gamma_{n-1,1}\boldsymbol{u}_n(\mathcal{K}_{\boldsymbol{\epsilon}}) - \gamma_{n,1}\boldsymbol{u}_{n-1}(\mathcal{K}_{\boldsymbol{\epsilon}})$$

and

$$\boldsymbol{u}_{n-i}(\mathcal{K}_1) = \gamma_{n,1}\boldsymbol{u}_n(\mathcal{K}_{\boldsymbol{\epsilon}}) + \gamma_{n-1,1}\boldsymbol{u}_{n-1}(\mathcal{K}_{\boldsymbol{\epsilon}}) + \ldots + \gamma_{n-i,1}\boldsymbol{u}_{n-i}(\mathcal{K}_{\boldsymbol{\epsilon}}) - \left(\frac{\sum_{j=n-i}^{n}\gamma_{j,1}^2}{\gamma_{n-i-1,1}}\right)\boldsymbol{u}_{n-i-1}(\mathcal{K}_{\boldsymbol{\epsilon}})$$

for all $i = 1, \ldots n-k-2$. These vectors are orthogonal to $\boldsymbol{q}_1$, as the last term always cancels out all the others. They are also orthogonal to each other by a similar argument. Returning to the eigenvalue equation, we have

$$\mathcal{K}_1^{-1}\boldsymbol{u}_i(\mathcal{K}_1) = \left(\mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{q}_1\boldsymbol{q}_1'\right)\boldsymbol{u}_i(\mathcal{K}_1) = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{u}_i(\mathcal{K}_1) = \lambda_n\left(\mathcal{K}_{\boldsymbol{\epsilon}}\right)^{-1}\boldsymbol{u}_i(\mathcal{K}_1)$$

for $i = k+2, \ldots, n$. Dividing through by the norm delivers the remaining eigenvectors of $\mathcal{K}_1$, all with eigenvalue $\lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}$.

Now assume that we know the eigenvectors $\boldsymbol{u}_i(\mathcal{K}_{R-1})$ of $\mathcal{K}_{R-1}$. We look for the eigenvectors of $\mathcal{K}^{-1} \equiv \mathcal{K}_R^{-1} = \mathcal{K}_{R-1}^{-1} + \boldsymbol{q}_R\boldsymbol{q}_R'$ and note that $\mathcal{K}_{R-1}^{-1} = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R-1}\boldsymbol{q}_r\boldsymbol{q}_r'$. Proceeding as before, we begin with the bottom R eigenvalues and eigenvectors of $\mathcal{K}$ and guess $\lambda_i(\mathcal{K})^{-1} = \lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1} + \|\boldsymbol{q}_i\|^2$ and $\boldsymbol{u}_1(\mathcal{K}) = \boldsymbol{q}_i/\|\boldsymbol{q}_i\|$ for $i = 1, \ldots R$. We have

$$\mathcal{K}^{-1}\boldsymbol{q}_i = \left(\mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R}\boldsymbol{q}_r\boldsymbol{q}_r'\right)\boldsymbol{q}_i = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{q}_i + \|\boldsymbol{q}_i\|^2\boldsymbol{q}_i = \left(\lambda_n(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1} + \|\boldsymbol{q}_i\|^2\right)\boldsymbol{q}_i$$

for all $i = 1, \ldots, R$ as $\boldsymbol{q}_i'\boldsymbol{q}_v = 0$ for all $v = 1, \ldots, R$, $v \neq i$. We next postulate $\lambda_{R+i}(\mathcal{K})^{-1} = \lambda_i(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}$ and $\boldsymbol{u}_{R+i}(\mathcal{K}) = \boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}})$ for $i = 1, \ldots, k$. As before

$$\mathcal{K}^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) + \sum_{r=1}^{R}\boldsymbol{q}_r\boldsymbol{q}_r'\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}}) = \lambda_i(\mathcal{K}_{\boldsymbol{\epsilon}})^{-1}\boldsymbol{u}_i(\mathcal{K}_{\boldsymbol{\epsilon}})$$

for $i = 1, \ldots, k$. Finally, notice that $\mathcal{K}$ is a rank-one update of $\mathcal{K}_{R-1}$, so we may apply a similar logic to before: Notice that $\boldsymbol{u}_i(\mathcal{K}_{R-1})$ form a basis for $\mathbb{R}^n$ and write

$$\boldsymbol{q}_R = \sum_{i=R+k}^{n}\gamma_{i,R}\boldsymbol{u}_i(\mathcal{K}_{R-1}).$$

Then let

$$\boldsymbol{u}_n(\mathcal{K}) = \gamma_{n-1,R}\boldsymbol{u}_n(\mathcal{K}_{R-1}) - \gamma_{n,R}\boldsymbol{u}_{n-1}(\mathcal{K}_{R-1})$$

and

$$\boldsymbol{u}_{n-i}(\mathcal{K}) = \gamma_{n,R}\boldsymbol{u}_n(\mathcal{K}_{R-1})+\gamma_{n-1,R}\boldsymbol{u}_{n-1}(\mathcal{K}_{R-1})+\ldots+\gamma_{n-i,R}\boldsymbol{u}_{n-i}(\mathcal{K}_{R-1})-\left(\frac{\sum_{j=n-i}^{n}\gamma_{j,R}^2}{\gamma_{n-i-1,R}}\right)\boldsymbol{u}_{n-i-1}(\mathcal{K}_{R-1})$$

for all $i = 1, \ldots n-k-R-1$. From the eigenvalue equation

$$\mathcal{K}^{-1}\boldsymbol{u}_i(\mathcal{K}) = \left(\mathcal{K}_{R-1}^{-1} + \boldsymbol{q}_R\boldsymbol{q}_R'\right)\boldsymbol{u}_i(\mathcal{K}) = \mathcal{K}_{R-1}^{-1}\boldsymbol{u}_i(\mathcal{K}) = \lambda_n\left(\mathcal{K}_{R-1}\right)^{-1}\boldsymbol{u}_i(\mathcal{K})$$

for all $i = 1, \ldots n-k-R-1$. Notice that by induction $\lambda_n\left(\mathcal{K}_{R-1}\right)^{-1} = \lambda_n\left(\mathcal{K}_{\boldsymbol{\epsilon}}\right)^{-1}$. After normalising, we have the last $n-k-R$ eigenvectors, all with eigenvalue $\lambda_n\left(\mathcal{K}_{\boldsymbol{\epsilon}}\right)^{-1}$. □

Proof of Theorem 3. We have

$$\left\|\mathbf{K}^{-1} - \mathcal{K}^{-1}\right\| = \left\|\mathbf{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R}\boldsymbol{q}_r\boldsymbol{q}_r' - \left(\mathcal{K}_{\boldsymbol{\epsilon}}^{-1} + \sum_{r=1}^{R}\boldsymbol{q}_r\boldsymbol{q}_r'\right)\right\| = \left\|\mathbf{K}_{\boldsymbol{\epsilon}}^{-1} - \mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\right\| = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right),$$

where the last step follows from the fact that $\|\mathbf{K}_{\boldsymbol{\epsilon}} - \mathcal{K}_{\boldsymbol{\epsilon}}\| = O\left(\sqrt{\log n/\left(n\rho_n\right)}\right)$ from Theorem 2 and that $\|\mathcal{K}_{\boldsymbol{\epsilon}}^{-1}\|$ is bounded by Lemma 1. This implies

$$\|\mathbf{K} - \mathcal{K}\| = O\left(\sqrt{\frac{\log n}{n\rho_n}}\right)$$

as $\|\mathcal{K}\|$ is bounded by Lemma 3.

We next bound the spectral norm of the difference between $\boldsymbol{\Sigma}$ and $\widehat{\boldsymbol{\Sigma}}$. We have, for every $s \geq 0$

$$\mathbb{P}\left(\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\| \leq \max(\delta, \delta^2)\|\boldsymbol{\Sigma}\|\right) \geq 1 - \exp\left(-cs^2\right)$$

with $\delta = C\sqrt{\frac{n}{T}} + \frac{s}{\sqrt{T}}$, and $C > 0$ and $c > 0$ are bounded by absolute constants. Now let $s = \sqrt{C'n}$ where C' is a constant. Then for C' sufficiently large

$$\mathbb{P}\left(\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\| \leq \max(\delta, \delta^2)\|\boldsymbol{\Sigma}\|\right) \geq 1 - \exp\left(-cC'n\right) \geq 1 - \exp\left(-n\right),$$

where $\delta = \left(C + \sqrt{C'}\right)\sqrt{\frac{n}{T}}$. Hence

$$\mathbb{P}\left(\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\| \leq C''\|\boldsymbol{\Sigma}\|\sqrt{\frac{n}{T}}\right) \geq 1 - \exp\left(-n\right)$$

with $C'' = \left(C + \sqrt{C'}\right) > 0$ a constant. Hence with high probability

$$\left\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\right\| = O\left(\sqrt{\frac{n}{T}}\|\boldsymbol{\Sigma}\|\right)$$

which implies

$$\left\|\widehat{\mathbf{K}} - \mathbf{K}\right\| = O\left(\sqrt{\frac{n}{T}}\|\boldsymbol{\Sigma}\|\right)$$

as $\|\mathbf{K}\|$ is bounded with high probability. Then we have

$$\left\|\widehat{\mathbf{K}} - \mathcal{K}\right\| \leq \left\|\widehat{\mathbf{K}} - \mathbf{K}\right\| + \left\|\mathbf{K} - \mathcal{K}\right\| = O\left(\sqrt{\frac{n}{T}}\|\mathbf{\Sigma}\| + \sqrt{\frac{\log n}{n\rho_n}}\right)$$

by the triangle inequality. We may then apply the Davis-Kahan theorem in a similar way to before in the proof of theorem 2. Finally, applying similar arguments as in Theorem 1

$$\frac{|\mathscr{M}|}{n} = O\left(\frac{n}{T}\|\mathbf{\Sigma}\|^2 + \frac{\log n}{n\rho_n}\right).$$

□

B Technical Tools

This appendix provides some results that are used in Sections 2 and 3. These have been adapted to our notation and purposes.

The following result of Oliveira (2009) bounds the spectral norm of the difference between the Laplacian and the population Laplacian.

Theorem B.1 (Theorem 3.1 of Oliveira (2009))**.** *Consider a random graph on n vertices $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $(i,j) \in \mathcal{E}$ with probability $p_{i,j} = p_{j,i}$ independently. Let $\mathbf{L}$ and $\mathcal{L}$ the normalised Laplacian and population normalised Laplacian corresponding to $\mathcal{G}$. Let $\bar{d}_{min}$ be the minimum expected degree of $\mathcal{G}$. For any constant $c > 0$ there exists another constant $C(c) > 0$ independent of n and the $p_{i,j}$, such that if $\bar{d}_{min} \geq C(c) \log n$, then for all $n^{-c} \leq \eta \leq 1/2$*

$$\mathbb{P}\left(\|\mathbf{L} - \mathcal{L}\| \leq 14 \sqrt{\frac{\log(4n/\eta)}{\bar{d}_{min}}} \right) \geq 1 - \eta.$$

Proof. See Oliveira (2009). □

We follow von Luxburg (2007) and Rohe *et al.* (2010) in the discussion of principal angles and the following theorem. Principal angles, also called canonical angles, are a common way to measure distances between subspaces. If $\mathbf{W}$ and $\mathcal{W}$ are $n \times k$ matrices with orthonormal columns, the singular values of $\mathbf{W}'\mathcal{W}$ are the cosines of the k principal angles between the column spaces of $\mathbf{W}$ and $\mathcal{W}$. Let $\sin \Phi(\mathbf{W}, \mathcal{W})$ be a $k \times k$ diagonal matrix of the sines of the principal angles of $\mathbf{W}$ and $\mathcal{W}$. The following theorem, known as the Davis-Kahan theorem (Davis and Kahan (1970), Stewart and Sun (1990), Bhatia (1997)), bounds the principal angles between the eigenspaces of two symmetric matrices. We only need to consider the case where the eigenspaces are of the same dimension.

Lemma B.1 (Theorem 7 of von Luxburg (2007))**.** *Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric $n \times n$ matrices. Let S be an interval and $\lambda_S(\mathbf{A})$ the set of eigenvalues of $\mathbf{A}$ that are in the interval S and $\lambda_S(\mathbf{B})$ the same for $\mathbf{B}$. Let $\lambda_{\overline{S}}(\mathbf{A})$ denote the eigenvalues of $\mathbf{A}$ that are not in S. Define $\mathbf{W}$ as the image of the spectral projection induced by $\lambda_S(\mathbf{A})$ and $\mathcal{W}$ as the analogous for $\lambda_S(\mathbf{B})$. Assume that both $\mathbf{W}$ and $\mathcal{W}$ are $n \times k$. Finally, let*

$$\delta = \min\{|\lambda - s| : \lambda \in \lambda_{\overline{S}}(\mathbf{A}), s \in S\}$$

and $\sin \Phi(\mathbf{W}, \mathcal{W})$ be the $k \times k$ diagonal matrix of the sines of the principal angles of $\mathbf{W}$ and $\mathcal{W}$. Then

$$\|\sin \Phi(\mathbf{W}, \mathcal{W})\| \leq \frac{\|\mathbf{A} - \mathbf{B}\|}{\delta}.$$

Proof. See Bhatia (1997) and Stewart and Sun (1990). □

The next theorem bounds the spectral norm of the difference between the sample covariance matrix and its population analogue.

Theorem B.2 (Theorem 5.39 and Remark 5.40 of Vershynin (2010))**.** *Assume $\mathbf{A}$ is a $T \times n$ matrix the rows of which $\boldsymbol{a}_{i\bullet}$ are centred, independent sub-Gaussian random vectors with second moment matrix $\mathbf{\Sigma}$. Then for every $s \geq 0$*

$$\mathbb{P}\left(\left\| \frac{1}{T}\mathbf{A}'\mathbf{A} - \mathbf{\Sigma} \right\| \leq \max(\delta, \delta^2) \|\mathbf{\Sigma}\| \right) \geq 1 - \exp\left(-cs^2\right)$$

with $\delta = C\sqrt{\frac{n}{T}} + \frac{s}{\sqrt{T}}$. $C > 0$ *and* $c > 0$ *are absolute constants that only depend on the maximum of the sub-Gaussian norms of the isotropic vectors* $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}_{i\bullet}$.

Proof. See Vershynin (2010). □