Mixed Membership Estimation for Social Networks

Jiashun Jin*, Zheng Tracy Ke† and Shengming Luo*

Carnegie Mellon University*
and Harvard University†

Dec. 2, 2022

Abstract

In economics and social science, network data are regularly observed, and a thorough understanding of the network community structure facilitates the comprehension of economic patterns and activities. Consider an undirected network with n nodes and K communities. We model the network using the Degree-Corrected Mixed-Membership (DCMM) model, where for each node $i = 1, 2, \ldots, n$, there exists a membership vector $\pi_i = (\pi_i(1), \pi_i(2), \ldots, \pi_i(K))'$, where $\pi_i(k)$ is the weight that node i puts in community k, $1 \leq k \leq K$. In comparison to the well-known stochastic block model (SBM), the DCMM permits both severe degree heterogeneity and mixed memberships, making it considerably more realistic and general. We present an efficient approach, Mixed-SCORE, for estimating the mixed membership vectors of all nodes and the other DCMM parameters. This approach is inspired by the discovery of a delicate simplex structure in the spectral domain. We derive explicit error rates for the Mixed-SCORE algorithm and demonstrate that it is rate-optimal over a broad parameter space. Our findings provide a novel statistical tool for network community analysis, which can be used to understand network formations, extract nodal features, identify unobserved covariates in dyadic regressions, and estimate peer effects. We applied Mixed-SCORE to a political blog network, two trade networks, a co-authorship network, and a citee network, and obtained interpretable results.

Keywords. Citee network, coauthorship network, communities, node embedding, political blogs, SCORE, simplex, spectral clustering, trade network.

AMS 2000 classification. Primary 62H30, 91C20; Secondary 62P25.

1 Introduction

Many economic activities happen on networks. Some examples of economic networks are the international trade networks, high-school friendship networks, stock co-jump networks, and job information networks. We denote a network with n nodes by its adjacency matrix $A \in \mathbb{R}^{n \times n}$, with $A_{ij} = 1$ if there is an edge between nodes i and j and $A_{ij} = 0$ otherwise.

In network econometrics, there is a surge of interests in understanding the interplay between network topology and economic activities (Graham, 2020). The literature can be divided into two categories, *formation* and *consequence*. Research in *formation* treats the network itself as the object of interest and studies the mechanism of forming the network.

One popular model is the dyadic regression model, including the famous gravity model for bilateral trade (Tinbergen, 1962) as a special example. In this model, $\mathbb{E}[A_{ij}]$ is a function of the dyadic covariates $\boldsymbol{X}_{ij}$ and nodal covariates $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$, and the main goal is estimation and inference of parameters of this function. Another popular model is the strategic model of network formation (Jackson and Wolinsky, 1996). In this model, each node has a utility function $u_i(A)$ that depends on the whole network, so deletion/addition of an edge affects the utilities of all nodes. Given these utility functions $\{u_i\}_{i=1}^n$, the network is in equilibrium if no node wishes to delete an edge and no pair of nodes wish to add an edge. The problems of interest include estimation and inference of these utility functions, e.g., by using network moment statistics (Miyauchi, 2016). Research in *consequence* treats the network as given information and aims to study influence of network structure on economic outcomes. There is a line of literature on estimation of the linear-in-means models (Manski, 1993; Bramoullé et al., 2009). In the simplest case of no covariates, let y_i be the response of node i and d_i be the degree of node i; the linear-in-means model assumes $y_i = \alpha + \beta \sum_j (d_i^{-1} A_{ij}) y_j + \epsilon_i$, with ϵ_i's being i.i.d. noise. The parameter β captures the 'peer effect' and is of main interest.

Independent of the econometric literature, there is also a body of statistical literature on network data analysis, where the main interest is fitting a probabilistic, easy-to-interpret model for an observed network. Pioneered by Bickel and Chen (2009), the stochastic block model (SBM) has attracted much attention. SBM assumes that nodes are divided into a few communities, and $\mathbb{E}[A_{ij}]$ is determined by community memberships of two nodes. Different from the *formation* literature of network econometrics, there are usually no observed covariates and the adjacency matrix A is the only available data. Many methods have been proposed for estimating the underlying community structure from A.

Recently, the two lines of literature have crossed. There are many interests in applying statistical network models in econometrics. Auerbach (2022) proposed a joint regression and network formation model, where the goal is learning latent nodal features from the network and using these features in the regression. Chen et al. (2020) used network modeling to estimate the Bernoulli probability matrix $\mathbb{E}[A]$. They replaced A by $\widehat{\mathbb{E}[A]}$ in fitting a network auto-regression model, in hopes of improving the estimation of peer effects. Graham (2015) combined the dyadic regression in econometrics and the latent space model in statistics to account for both observed and unobserved covariates in network formation.

Unfortunately, despite these encouraging progresses, we note two problems. First, both the statistical literature and the econometric literature have been largely focused on some classical and idealized network models, such as the stochastic block model (SBM) and the

graphon (Lovász and Szegedy, 2006). Second, recent developments in statistical network analysis have suggested new ideas in network modeling, but such ideas are largely unknown in the area of network econometrics. The SBM and graphon models are often too idealized for real networks. Many real networks have the so-called *severe degree heterogeneity*, meaning that the degree of one node is higher than another by 10 or even 100 times (Jin et al., 2021b, Table 1). Also, many networks have the so-called *mixed-membership*, meaning that different network communities overlap with each, and a node may belong to multiple communities (Airoldi et al., 2008); for such networks, the SBM is too idealized, which does not model either mixed-membership or severe degree heterogeneity. The graphon model is also too idealized. It does not model severe degree heterogeneity and requires that the nodes are exchangeable (an assumption that is hard to check and is too strong for many real networks). It is therefore desirable to (a) develop more realistic network models and new algorithms, and (b) introduce the most recent developments in statistical network analysis to the area of network econometrics.

We propose the Degree-Corrected Mixed-Membership (DCMM) model as a more suitable network model. Compared with SBM, DCMM allows for both severe degree heterogeneity and mixed membership and it is much broader. Compared with graphon, DCMM accommodates severe degree heterogeneity and does not require node exchangeability. Since many real networks have strong mixed-membership, an interesting problem is how to estimate the mixed-memberships of nodes. We propose a fast spectral method, Mixed-SCORE, for estimating network mixed-memberships, and show that it is rate-optimal in a decision theory framework. Given the interesting connections between the two areas (statistical network analysis and network econometrics) we discuss above, our model and method not only provide new contributions to the former but also provide new opportunities to the latter. For example, for many existing works in network econometrics that used SBM or graphon as the network model, we may improve the results by using the more realistic DCMM model. Also, our method is useful in several problems of network econometrics. For example, one can use the output of our method to understand network formation, create nodal features, estimate the Bernoulli probability matrix, and learn the unobserved dyadic covariates.

In what follows, we first present a motivating example. In this example, DCMM has a relatively simple form. We use this example to illustrate why DCMM is a reasonable model and how to use the output of our method to answer real questions of interest.

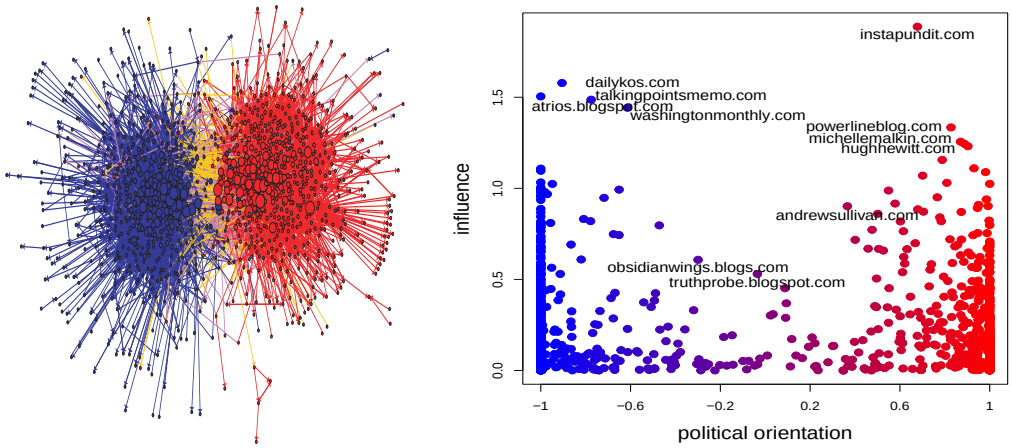


Figure 1: The political blog network and the output of Mixed-SCORE. Left: A visualization of the network (figure source: Adamic and Glance (2005)), where blue/red colors indicate the manually assigned community labels by Adamic and Glance (2005), and yellow/purple colors indicate the edges between two commmunities. Right: The estimated p_i (x-axis) and θ_i (y-axis) by the Mixed-SCORE algorithm.

1.1 A motivating example: Political blog network

The 2004 U.S. Presidential Election was the first presidential election in the United States in which blogging played an important role. Adamic and Glance (2005) recorded the linkages of political blogs in a single day snapshot before the election. We use the data to construct an undirected network, where each node is a blog and two blogs are connected by an edge if they have links between them (one-way or reciprocal). The giant component of the network has $n = 1222$ nodes. We assume each blog has a political orientation parameter $p_i \in [-1, 1]$, where $p_i > 0$, $p_i = 0$ and $p_i < 0$ corresponds to liberal, neutral and conservative. A node with $p_i = 1$ is extremely conservative, while a node with $p_i = 0.2$ is only mildly conservative. We also assume each blog has a popularity score $\theta_i > 0$. The larger θ_i, the more influence of the blog. Suppose the edges are independently generated. We model the edge probability between two nodes as a function of their political orientations and popularities: [1]

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \cdot (\alpha + \beta p_i p_j), \qquad 1 \le i < j \le n. \tag{1.1}$$

Here, $\alpha > 0$ is the baseline effect, and $\beta > 0$ captures the effect of political orientations on linkage probabilities. When two blogs are both liberal or both conservative, $\beta p_i p_j > 0$, so they are more likely to be linked. When one blog is liberal and the other is conservative, $\beta p_i p_j < 0$, so they are less likely to be linked. The more extreme of political orientations of

[1] Model (1.1) is not identifiable, as we can multiple (α, β) by a scalar c and divide each θ_i by $\sqrt{c}$ to make the edge probabilities invariant. For identifiability, we let $\alpha + \beta = 1$. This is the same as the identifiability condition we use for a general DCMM model (see Section 2).

two nodes, the larger $|\beta p_i p_j|$ and the stronger effect on linkage probability. Besides political orientations, the linkage probability is also affected by the popularity of nodes. Suppose two blogs i and j have exactly the same political orientation, but blog i has a larger influence in the internet. It is more likely for other blogs to link to blog i than blog j.

We propose a fast spectral method, Mixed-SCORE, for estimating (p_i, θ_i) of each node and the global parameters (α, β). The details of this method will be deferred to Section 2. Figure 1 plots $(\hat{p}_i, \hat{\theta}_i)$ of political blogs. The points in the top left regions correspond to *influential and liberal* blogs, and those in the top right region are *influential and conservative* blogs. Some of these influential blogs are more 'extreme' than others in political orientation, such as the liberal blog `atrios.blogspot.com` and the conservative blog `hughhewitt.com`. Blogs with large $\hat{\theta}_i$ typically have clear political orientations and are far away from being neutral, with some exceptions like `truthprobe.blogspot.com`.

When Adamic and Glance (2005) collected this data set, they assigned a manual label $\ell_i \in \{\text{liberal}, \text{conservative}\}$ to each blog i by checking the host website directory or reading blog posts. Our method does not need any manual efforts to label the blogs; using the sign of $\hat{p}_i$, we can recover their manual labels with an accuracy of 95.5%. Meanwhile, people are interested in not only the label of a blog but also the extremity of its political orientation, as an extremely conservative blog and a mildly conservative blog can have different opinions on issues such as abortion, gun control, and death penalties (Hindman et al., 2003). The $\hat{p}_i$'s from our method help reveal such information that is not seen in manual labels.

We can use the output of Mixed-SCORE in several different ways. First, it is useful for understanding the formation of links between blogs. Our method obtains $\hat{\beta} = 1 - \hat{\alpha} = 0.471$. It captures the effect of political orientation on link formation.[2] Second, our method creates two covariates, $\hat{p}_i$ ('political orientation') and $\hat{\theta}_i$ ('influence'), for each blog. These covariates will be useful in other tasks such as predicting the opinion of a blogger on a given topic. Third, we obtain $\widehat{\mathbb{E}[A_{ij}]} = \hat{\theta}_i \hat{\theta}_j (\hat{\alpha} + \hat{\beta} \hat{p}_i \hat{p}_j)$, which can be plugged into the linear-in-means model to improve the estimation of peer effect. Let y_i be an outcome of interest (e.g., the frequency of a key word in blog posts). We fit a model $y_i = \gamma + \delta \hat{q}_i^{-1} \sum_j \hat{\theta}_j (\hat{\alpha} + \hat{\beta} \hat{p}_i \hat{p}_j) y_j + \epsilon_i$, where $\hat{q}_i = \sum_{k=1}^n \hat{\theta}_k (\hat{\alpha} + \hat{\beta} \hat{p}_i \hat{p}_k)$. Compared with the standard linear-in-means model, this one better deals with measurement errors on the network itself.

[2] We focus on estimation in this paper. In a companion paper Jin et al. (2021a), we also provide a test for testing against the null hypothesis $\beta = 0$. The p-value is $< 10^{-7}$ for this political blog network, suggesting a significant effect of political orientation on link formation.

1.2 Main results and contributions

Model (1.1) is a special case of the Degree-Corrected Mixed Membership (DCMM) model to be introduced in Section 2. In the DCMM model, the network has K perceivable communities. Each node has a mixed membership vector $\pi_i \in \mathbb{R}^K$, where $\pi_i(k) \geq 0$ is the weight that node i puts on community k, satisfying $\sum_{k=1}^K \pi_i(k) = 1$. When π_i is degenerate (i.e., π_i has only one nonzero entry which is equal to 1, and the other entries are zero), we call node i a *pure node*; otherwise, we call it a *mixed node*. In Model (1.1) for the political blog network, $K = 2$, $\pi_i = (\frac{1-p_i}{2}, \frac{1+p_i}{2})'$, and a node is pure if and only if $p_i \in \{\pm 1\}$. Each node also has a degree heterogeneity parameter $\theta_i > 0$. The probability of forming an edge between nodes i and j is determined jointly by their mixed membership vectors and degree heterogeneity parameters (see Section 2.1). Given the adjacency matrix A, we are interested in estimating parameters of DCMM, especially the membership matrix $\Pi := [\pi_1, \pi_2, \ldots, \pi_n]'$. Estimation of Π is known as the problem of mixed membership estimation (Airoldi et al., 2008).

In the statistical literature of network data analysis, many works focused on community detection, which clusters nodes into K non-overlapping communities. Overlapping community detection (Gregory, 2010) allows the assignment of a node to more than one community. It is equivalent to a community detection problem with 2^K non-overlapping communities. Community detection is a clustering problem, so the methods and theory do not apply to mixed membership estimation. Airoldi et al. (2008) is a pioneer work on mixed membership estimation. They considered a special setting of DCMM with $\theta_1 = \theta_2 = \ldots = \theta_n$ (i.e., no degree heterogeneity) and assumed that π_i's are i.i.d. generated from a Dirichlet prior. They proposed a variational Bayes approach to computing the posterior of $\pi_1, \ldots, \pi_n$. However, in many real networks, degree heterogeneity is severe (Newman, 2003), so we must assume unequal θ_i's. Zhang et al. (2020) proposed the OCCAM algorithm for mixed membership estimation. OCCAM has the nice property of accommodating degree heterogeneity, but it requires a condition that the fraction of mixed nodes must be properly small, and so it does not work for networks with a large fraction of mixed nodes.

We propose a new method Mixed-SCORE for network mixed membership estimation. It is inspired by our discovery of a low-dimensional simplex geometry associated with the leading eigenvectors of A. Using linear algebra, we establish an explicit connection between this simplex and the target quantity Π. It leads to a fast spectral algorithm for estimating Π. Compared with the existing methods of mixed membership estimation (Airoldi et al., 2008; Zhang et al., 2020), Mixed-SCORE successfully deals with degree heterogeneity and allows for an arbitrary fraction of mixed nodes. Furthermore, we also give a characterization of the

error rate of Mixed-SCORE and show that it is rate-optimal for a wide range of settings. In comparison, the competitors either have no theoretical guarantees (Airoldi et al., 2008) or have non-optimal error rates (Zhang et al., 2020). Given $\hat{\Pi}$ from Mixed-SCORE, we also propose estimates of other parameters of DCMM.

1.3 Applications in network econometrics

We give a few examples of using our model and method in network econometrics.

Example 1: Economic outcomes are often affected by social influence. For example, a high school student's academic performance might depend on the attitudes and expectations of his/her friends and family. Such a social influence is not directly observed, and a popular solution is to collect network data and hope the unobserved social influence is revealed by linking behavior in the network (e.g., students with similar reported friendships may have similar family expectations (Auerbach, 2022)). Let $y_i \in \mathbb{R}$ be the outcome (e.g., academic performance of a student) and $X_i \in \mathbb{R}^p$ the observed features (e.g., school rating, family income, etc.). Consider an unobserved social influence such as the family expectation. We assume there are K extreme types of family expectation and the family expectation of a student is represented by a mixed membership vector $\pi_i \in \mathbb{R}^K$. We model the network by DCMM and the outcome by a regression $y_i = X_i'\beta + f(\pi_i) + \epsilon_i$. This model is similar to the model in Auerbach (2022), except that he models the network by graphon but we model it by DCMM. We can apply Mixed-SCORE to obtain $\hat{\pi}_i$ and plug them into the regression. Compared with the method in Auerbach (2022), our approach has some advantages: First, we allow the social feature π_i to have an arbitrary dimension K, but in a graphon, π_i is a scalar in $[0, 1]$. Second, our approach deals with severe degree heterogeneity and guarantees that the estimated social feature is not biased by the student's own friendship popularity.

Example 2: Understanding the social interactions or 'peer effects' in decision making is of great interest in economics. To estimate the peer effect, we propose a new linear-in-means model based on DCMM: Given a network generated from DCMM, let $y = (y_1, y_2, \ldots, y_n)'$ store the response at each node and $X = [X_1, X_2, \ldots, X_n]' \in \mathbb{R}^{n\times p}$ store the feature vectors. Define $G \in \mathbb{R}^{n\times n}$ by $G_{ij} = \pi_i'\pi_j/(\sum_{k:k\neq i} \pi_i'\pi_k)$, for $i \neq j$, and $G_{ii} = 0$. For some parameters $\alpha, \beta \in \mathbb{R}$ and $\gamma, \delta \in \mathbb{R}^p$, we model that $y = \alpha\mathbf{1}_n + \beta Gy + X\gamma + GX\delta + \epsilon$, where ϵ is the noise vector. This model differs from the standard linear-in-means model (Manski, 1993) in the definition of G. In the standard form, G is chosen as the normalized adjacency matrix. However, the adjacency matrix itself has stochastic errors. For example, two friends in real life may or may not be each other's Facebook friend. Our G allows for a possibly nonzero

peer effect between two nodes even when they are not directly connected by an edge. Under this model, we can apply Mixed-SCORE to obtain $\hat{\pi}_i$'s and then plug them into the model for y_i. A similar idea has been considered by Chen et al. (2020) for vector autoregression. They model the network with SBM, but we use the more general DCMM model.

Example 3: The dyadic regression model (Graham, 2020) is a popular network model. When there are unobserved covariates, how to make accurate parameter estimation is not fully understood. Inspired by Graham (2015), we assume that an unobserved dyadic covariate is a function of unobserved nodal covariates and propose a dyadic regression model with a DCMM-like structure. Let $X \in \mathbb{R}^{n\times n}$ be the adjacency matrix of a weighted network (e.g., in the international trade network, X_{ij} is the trade flow from country i to country j). Suppose $X_{ij} \sim \text{Poisson}(\lambda_{ij})$, with $\ln(\lambda_{ij}) = \sum_{m=1}^{M} \gamma_m \ln(Z_{m,i,j}) + \beta \ln(\pi_i' P \pi_j) + c_i + c_j$. Here $Z_1, \ldots, Z_M$ are the observed dyadic covariates, c_i is the fixed effect of node i, and $U_{ij} := \pi_i' P \pi_j$ is an unobserved dyadic covariate, with (π_i, P) similar to those in DCMM (to be introduced in Section 2.1). This model is connected to the model in Graham (2015): In his model, $U_{ij} = g(\xi_i, \xi_j; \delta_0)$, where $\xi_i \in \mathbb{R}$ is an unobserved nodal covariate and $g(\cdot, \cdot; \delta_0)$ is a symmetric distance function; in our model, the latent covariate π_i can take an arbitrary dimension K. We introduce a practical algorithm in Section 5.1: We first construct a network from the residuals of fitting a dyadic regression with only observed covariates; we then apply Mixed-SCORE to obtain $\hat{U}_{ij} = \hat{\pi}_i' \hat{P} \hat{\pi}_j$; last, we plug in $\hat{U}_{ij}$ and re-fit the dyadic regression. Although this approach is mainly from a practical perspective, it points out a new direction, that is, using spectral algorithms to learn unobserved covariates. Compared with the existing approaches such as Markov Chain Monte Carlo and triad probit (Graham, 2020), the spectral approach is computationally fast and allows for multidimensional π_i's.

Since the main focus of this paper is estimation of π_i, we leave a careful study of these examples to future work. One of the key requirements for plugging $\hat{\pi}_i$ into a downstream economic model is that the error on $\hat{\pi}_i$ can be well-controlled. In this paper, we provide not only a method for estimating π_i but also the explicit error bounds. In the case that the network is properly dense, the error bound reduces to $\mathbb{E}[n^{-1}\sum_{i=1}^{n} \|\hat{\pi}_i - \pi_i\|^2] = O(n^{-1}K^3)$, suggesting that the errors on $\hat{\pi}_i$ are negligible for downstream tasks (please see the discussions following Theorem 3.2).

The remaining of this paper is organized as follows. In Section 2, we formally introduce our model and method. In Section 3, we state the theoretical results. In Sections 4-5, we present the simulations and real data, respectively. We conclude the paper with discussions in Section 6. The technical proofs are relegated to the online supplementary material.

2 A spectral method for network membership estimation

2.1 The DCMM model

Consider an undirected network with n nodes. Suppose the network contains K communities. Each node has a mixed membership vector $\pi_i = (\pi_i(1), \pi_i(2), \ldots, \pi_i(K))'$, where the entries of π_i are nonnegative and sum to 1. We interpret $\pi_i(k)$ as the fractional weight that node i puts on community k. If node i puts 100% weight on community k, then $\pi_i(k) = 1$ and $\pi_i(\ell) = 0$ for all other $\ell \neq k$; we say that π_i is degenerate and call node i a pure node of community k. If node i is not a pure node of any community, we call it a mixed node. Each node also has a degree heterogeneity parameter $\theta_i > 0$. Let $P \in \mathbb{R}^{K,K}$ be a symmetric nonnegative matrix. Recall that $A \in \mathbb{R}^{n\times n}$ is the adjacency matrix of the network. Since we do not allow for self-edges, the diagonal entries of A are all zero. We assume that the upper triangle of A (excluding the diagonal) contains independent Bernoulli variables, where for any $1 \leq i, j \leq n$ and $i \neq j$,

$$\mathbb{P}(A_{ij} = 1) = \theta_i\theta_j \times \sum_{k=1}^{K}\sum_{\ell=1}^{K} \pi_i(k)\pi_j(\ell)P_{k\ell} = \theta_i\theta_j \times \pi_i' P \pi_j. \tag{2.2}$$

Take Model (1.1) for the political blog network for example. It is a special case with $K = 2$, $\pi_i = (\frac{1-p_i}{2}, \frac{1+p_i}{2})'$ and P being a 2×2 matrix whose diagonal entries are equal to $\alpha+\beta$ and the off-diagonal entries are equal to $\alpha - \beta$. The parameters in (2.2) are not identifiable. For identifiability, we assume that the diagonal entries of P are equal to 1 (see Section A.1 of the supplementary material for a proof of model identifiability).

We call (2.2) the degree-corrected mixed membership (DCMM) model. DCMM includes several popular network models as special cases. The stochastic block model (SBM) is a special DCMM where θ_i's are equal to each other (i.e., no degree heterogeneity) and all π_i's are degenerate (i.e., no mixed membership). The MMSBM model (Airoldi et al., 2008) is a special case with equal θ_i's (but π_i's can be non-degenerate). The DCBM model (Karrer and Newman, 2011) is a special case where all π_i's are degenerate (but θ_i's can be unequal). DCMM can also be viewed as an equivalence to the OCCAM model (Zhang et al., 2020), except that π_i's are re-normalized by their ℓ^2-norms in the OCCAM model.

It is convenient to express (2.2) in a matrix form. Write $\Theta = \mathrm{diag}(\theta_1, \theta_2, \ldots, \theta_n) \in \mathbb{R}^{n,n}$ and $\Pi = [\pi_1, \pi_2, \ldots, \pi_n]' \in \mathbb{R}^{n,K}$. Introduce an $n \times n$ matrix $\Omega = \Theta\Pi P \Pi'\Theta$. It is seen that $\Omega_{ij} = \theta_i\theta_j \cdot \pi_i' P \pi_j$. By Model (2.2), $\mathbb{E}[A_{ij}] = \Omega_{ij}$ for all $1 \leq i \neq j \leq n$. It follows that

$$A = \Omega - \mathrm{diag}(\Omega) + W, \qquad \text{with} \quad W := A - \mathbb{E}[A] \quad \text{and} \quad \Omega := \Theta\Pi P\Pi'\Theta. \tag{2.3}$$

We call Ω, $\mathrm{diag}(\Omega)$, and W the "main signal", "secondary signal" and "noise" respectively.

Remark 1: DCMM distinguishes from the latent space models (Handcock et al., 2007) or graphons (Lovász and Szegedy, 2006; Pensky, 2019) by not requiring exchangeability of nodes. In DCMM, we have no assumptions saying that θ_i's and π_i's are i.i.d. drawn from some distributions. We treat all of them as unknown parameters.

Remark 2: DCMM has an interesting connection to the dyadic regression model. In DCMM, we can view θ_i and θ_i as nodal covariates, and $\pi_i' P \pi_j$ as a dyadic covariate, but a major difference is that these covariates are unobserved.

2.2 The simplex structure in the spectral domain

We first consider an oracle case where we observe the "main signal" matrix Ω in (2.3). We would like to construct an estimate of Π from Ω. Note that Ω is a rank-K matrix. For each $1 \leq k \leq K$, let λ_k be the kth largest eigenvalue of Ω in magnitude, and let $\xi_k \in \mathbb{R}^n$ be the associated eigenvector. Write $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ and $\Xi = [\xi_1, \xi_2, \ldots, \xi_K]$. Jin (2015) proposed a normalization of eigenvectors called the SCORE normalization. It constructs a matrix $R \in \mathbb{R}^{n \times (K-1)}$ containing the entry-wise ratios of eigenvectors, where

$$R(i,k) = \xi_{k+1}(i)/\xi_1(i), \qquad 1 \leq i \leq n, \;\; 1 \leq k \leq K-1. \tag{2.4}$$

Let $r_i \in \mathbb{R}^{K-1}$ denote the i-th row of R. Viewing each r_i as a point in the $(K-1)$-dimension Euclidean space, there is a simplex structure for the point cloud $\{r_i\}_{1 \leq i \leq n}$: [3]

Lemma 2.1 (The simplex geometry in R)**.** *Consider Model* (2.2) *and assume that P is non-singular, $P(\Pi'\Theta^2\Pi)$ is irreducible, and each community has at least one pure node. The following statements are true: (1) All entries of ξ_1 are strictly positive, so that the matrix R in* (2.4) *is well-defined. (2) There exists a K-vertex simplex $\mathcal{S} \subset \mathbb{R}^{K-1}$, whose vertices are denoted by $v_1, v_2, \ldots, v_K$, such that each r_i is contained in $\mathcal{S}$ and that r_i falls on one vertex of $\mathcal{S}$ if and only if node i is a pure node. (3) Let $w_i \in \mathbb{R}^K_+$ contain the barycentric coordinates of r_i in $\mathcal{S}$. The vector w_i is connected to π_i through the equation $w_i = (\pi_i \circ b_1)/\|\pi_i \circ b_1\|_1$, where $b_1 \in \mathbb{R}^K$ is the vector defined by $b_1(k) = [\lambda_1 + v_k' \mathrm{diag}(\lambda_2, \ldots, \lambda_K) v_k]^{-1/2}$, $\lambda_1, \lambda_2, \ldots, \lambda_K$ are the nonzero eigenvalues of Ω, and $\circ$ denotes the entrywise product between two vectors.*

We call $\mathcal{S}$ the *Ideal Simplex*. Lemma 2.1 inspires a method to recover Π from Ω. Step 1: Obtain R from (2.4). Step 2: By the second claim of Lemma 2.1, we can retrieve the vertices

[3] By definition, the simplex $\mathcal{S}$ spanned by $v_1, v_2, \ldots, v_K$ is the set of points r such that $r = \sum_{k=1}^K \beta_k v_k$ for some nonnegative vector β with $\|\beta\|_1 = 1$. If $v_1, v_2, \ldots, v_K$ are affinely independent, $\mathcal{S}$ is non-degenerate; and we call $v_1, \ldots, v_K$ the vertices of $\mathcal{S}$ and β the barycentric coordinate vector of r.

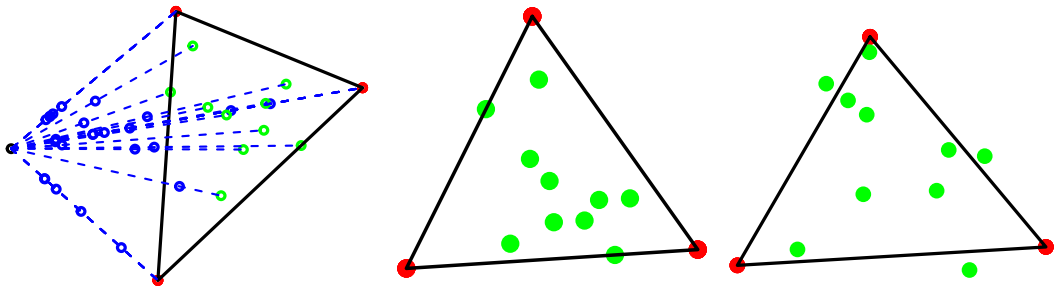


Figure 2: Illustration for why the simplex exists and the role of SCORE normalization ($K = 3$). Left: rows of Ξ (blue points). The point cloud is contained in a simplicial cone, and it is desirable to normalize the cone to a simplex. Middle: rows of R (red: pure nodes; green: mixed nodes). It shows that the SCORE normalization successfully produces a simplex. Right: rows of Ξ normalized by row-wise ℓ^1-norm (for visualization, we have projected these points to $\mathbb{R}^2$). This normalization fails to produce a simplex.

$v_1, \ldots, v_K$ by computing the convex hull of the point cloud $\{r_i\}_{1\le i\le n}$. Step 3: Given the vertices, we obtain the barycentric coordinate vector w_i for each node i (by solving a simple linear equation); we also compute the vector b_1 using the definition in Lemma 2.1; by the third claim of Lemma 2.1, we can recover π_i from $w_i \propto \pi_i \circ b_1$ and $\|\pi_i\|_1 = 1$.

Remark 3 *(Why the simplex exists and the crucial role of the SCORE normalization).* In the proof of Lemma 2.1, we will see that the rows of Ξ are contained in a simplicial cone with K supporting rays, where all the pure nodes in one community are on one supporting ray, and the mixed nodes are in the interior of the cone. The SCORE normalization (2.4) transforms the simplicial cone to a simplex and provides a direct link between the simplex and Π. Interestingly, other normalizations of eigenvectors (e.g., to normalize each row of Ξ by its own ℓ^1-norm) fail to produce a simplex structure. See Figure 2.

2.3 The Mixed-SCORE algorithm for estimating Π

We extend the aforementioned method of recovering Π to the real case where A, instead of Ω, is observed. For $1 \le k \le K$, let $\hat{\lambda}_k$ be the kth largest eigenvalue of A in magnitude, and let $\hat{\xi}_k \in \mathbb{R}^n$ be the associated eigenvectors. Write $\hat{\Lambda} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$ and $\hat{\Xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_K]$. We propose the following algorithm:

Mixed-SCORE algorithm for estimating Π. Input: A, K. Output: $\hat{\pi}_i$, $1 \le i \le n$.

- *SCORE step.* Fix a threshold $T > 0$ ($T = \log(n)$ by default). Obtain $(\hat{\lambda}_1, \hat{\xi}_1), \ldots, (\hat{\lambda}_K, \hat{\xi}_K)$ and define $\hat{R} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n]'$ as the matrix where for $1 \le i \le n$ and $1 \le k \le K-1$,

$$\hat{R}(i,k) = \mathrm{sign}(\hat{\xi}_{k+1}(i)/\hat{\xi}_1(i)) \cdot \min\big\{|\hat{\xi}_{k+1}(i)/\hat{\xi}_1(i)|,\ T\big\}. \tag{2.5}$$

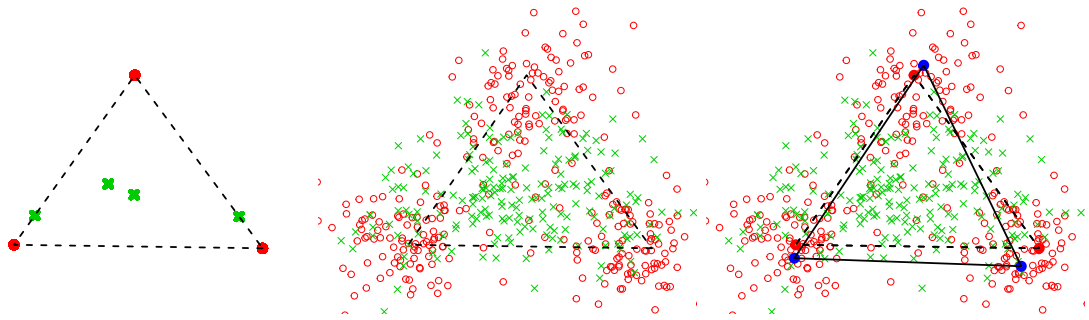


Figure 3: Left: rows of R (many rows are equal so a point may represent many rows). Middle: each point is a row of $\hat{R}$ (it is seen that we have strong noise and many outliers, so we may have poor results if we hunt for vertices directly). Right: same as the middle panel except that a triangle (solid blue) estimated by SVS is added. In all panels, dashed triangle is the Ideal Simplex, and red/green points correspond to pure/mixed nodes respectively. The figure suggests (a) the rows of $\hat{R}$ are quite noisy, with many outliers, and (b) SVS works reasonably well.

- *VH (vertex hunting) step.* Use the rows of $\hat{R}$ to estimate the vertices of Ideal Simplex (details below). Denote the estimated vertices by $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$.
- *MR (membership reconstruction) step.* Obtain an estimate of b_1 by

$$\hat{b}_1(k) = [\hat{\lambda}_1 + \hat{v}_k' \mathrm{diag}(\hat{\lambda}_2, \ldots, \hat{\lambda}_K)\hat{v}_k]^{-1/2}, \qquad 1 \le k \le K. \tag{2.6}$$

  For each $1 \le i \le n$, solve $\hat{w}_i \in \mathbb{R}^K$ from the linear equations: $\hat{r}_i = \sum_{k=1}^K \hat{w}_i(k)\hat{v}_k$, $\sum_{k=1}^K \hat{w}_i(k) = 1$. Define a vector $\hat{\pi}_i^* \in \mathbb{R}^K$ by $\hat{\pi}_i^*(k) = \max\{0, \hat{w}_i(k)/\hat{b}_1(k)\}$, $1 \le k \le K$. Estimate π_i by $\hat{\pi}_i = \hat{\pi}_i^*/\|\hat{\pi}_i^*\|_1$, $1 \le i \le n$.

In Step 1, $\hat{R}$ is an estimate of the matrix R in (2.4). In Step 3, $\hat{b}_1$ is an estimate of b_1 in Lemma 2.1. These two steps are similar to those in the oracle case. Step 2 is however very different from in the oracle case: The point cloud $\{\hat{r}_i\}_{1\le i\le n}$ is noisy. It is no longer possible to retrieve the vertices of the Ideal Simplex by simply computing the convex hull of these points. We call the estimation of $v_1, v_2, \ldots, v_K$ the vertex hunting (VH) problem. We introduce several VH algorithms. A summary of these algorithms is in Table 1.

The first possible VH approach is to use Successive Projection (SP) (Araújo et al., 2001). SP is a greedy algorithm. It starts by setting $\hat{v}_1$ as the data point $\hat{r}_i$ that has the largest Euclidean norm among $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$. Then, for $2 \le k \le K$ successively, it projects $\hat{r}_i$'s to the orthogonal complement of $\mathrm{Span}(\hat{v}_1, \ldots, \hat{v}_{k-1})$ and finds the data point with the largest Euclidean norm after projection; the estimated kth vertex $\hat{v}_k$ is set as the corresponding $\hat{r}_i$.

However, the SP algorithm frequently underperforms numerically. The Ideal Simplex is highly corrupted by noise and outliers (see Figure 3), but SP is well-known to be sensitive to outliers. To overcome the challenge, we propose *Sketched Vertex Search (SVS)*. SVS is a two-stage algorithm. In the denoise stage, we cluster n points into L clusters by k-means,

Table 1: Comparison of four versions of SVS (for completeness, we analyze all versions theoretically. Numerically, we recommend SVS and SVS* for they have better performances).

	Using exhaustive search in 2nd stage	Using SP in 2nd stage
$L < n$	SVS	SVS*
$L = n$	CVS	SP

for a tuning integer $K \ll L \ll n$. The center of each cluster (called a "local center") is the average of many nearby points and thus robust to outliers. In the second stage, we estimate K vertices from these L "local centers". The full algorithm is as follows:

Sketched Vertex Search (SVS) for vertex hunting. Input: K, a tuning integer $L \geq K$, the point cloud $\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n$. Output: vertices $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$.

- *Denoise.* Apply the classical k-means algorithm to $\{\hat{r}_i\}_{1\leq i\leq n}$ assuming there are L clusters. Denote the centers of the clusters by $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_L \in \mathbb{R}^{K-1}$.
- *Vertex search.* For any K distinct indices $1 \leq j_1 < \ldots < j_K \leq L$, let $\mathcal{H}(\hat{m}_{j_1}, \ldots, \hat{m}_{j_K})$ be the convex hull of $\hat{m}_{j_1}, \ldots, \hat{m}_{j_K}$, and

$$d_L(j_1, \cdots, j_K) = \max_{1\leq j\leq L} \text{distance}\big(\hat{m}_j,\ \mathcal{H}\{\hat{m}_{j_1}, \cdots, \hat{m}_{j_K}\}\big).^4 \tag{2.7}$$

  Find $1 \leq \hat{j}_1 < \hat{j}_2 < \ldots < \hat{j}_K \leq L$ that minimizes (2.7). Output $\hat{v}_k = \hat{m}_{\hat{j}_k}$, $1 \leq k \leq K$.

The tuning integer L can be chosen in a data-driven fashion. For each $L \in [K+1, 3K]$, let $d_L(\hat{R}) = d_L(\hat{j}_1, \cdots, \hat{j}_K)$ be the same as in (2.7) and $\delta_L(\hat{R}) = \min_{\{j_1,\ldots,j_K\}}\big(\max_{1\leq k\leq K}\{\|\hat{v}_{j_k}^{(L)} - \hat{v}_k^{(L-1)}\|\}\big)$, where the minimum is taken over all permutations of $\{1, 2, \ldots, K\}$. The quantity $\delta_L(\hat{R})$ tracks the change of estimated vertices when we increase the tuning parameter from $(L-1)$ to L. We select L by (if there is a tie, pick the largest integer):

$$\hat{L}_n^*(A) = \text{argmin}_{K+1\leq L\leq 3K}\{\delta_L(\hat{R})/(1 + d_L(\hat{R}))\}. \tag{2.8}$$

We also consider three variants of SVS. The first is SVS*, where in the second stage we apply SP to the L "local centers". The second is *Combinatorial Vertex Search (CVS)*, where we take $L = n$ in SVS (i.e., the denoise stage is skipped, so in the second stage, each $\hat{r}_i$ is viewed as a local center). In the last variant, we take $L = n$ in SVS*, so it reduces to SP. For practical use, we recommend SVS and SVS*; they have the denoise step by k-means, which is crucial for good numerical performance.

We view Mixed-SCORE a generic algorithm and treat VH as a "plug-in" step. For each VH approach, we can plug it in and obtain a different version of Mixed-SCORE. We denote

[4] For a point v and a set H, distance(v, H) is the Euclidean distance from v to H. When H is a simplex, this distance can be easily computed via a standard quadratic programming.

them by Mixed-SCORE-X, e.g., for X $\in$ {SVS, SVS*, CVS, SP}. Mixed-SCORE can also be used with other possible VH approaches.

The complexity of Mixed-SCORE mainly comes from obtaining the first K eigenvalues and eigenvectors of A, which is $O(nK^2)$, and the VH step, which is $O(nK^2)$ if we use the SP algorithm. Hence, Mixed-SCORE-SP is a polynomial-time algorithm. Mixed-SCORE-SVS is also a polynomial-time algorithm if (K, L) are both finite.

Remark 4 *(Comparison with the standard PCA).* The standard PCA approach creates a K-dimensional vector $x_i = \hat{\Xi}' e_i$ for each node i. These vectors do not have real meanings and are hard to interpret; moreover, each x_i is determined by all the parameters of DCMM and cannot faithfully represent the community structure among nodes. In comparison, the $\hat{\pi}_i$'s from Mixed-SCORE have clear interpretations.

2.4 Estimation of Θ and P

We are also interested in estimating the other parameters of DCMM. Among all the parameters, Π is the hardest to estimate. Once $\hat{\Pi}$ is obtained, estimation of (Θ, P) is comparably easy. Therefore, as a byproduct, we use the output of Mixed-SCORE to construct estimates of (Θ, P). Recall that $\lambda_1, \ldots, \lambda_K$ are the nonzero eigenvalues of Ω and $\xi_1, \ldots, \xi_K$ are the associated eigenvectors. Let $v_1, v_2, \ldots, v_K$ be the vertices of the Ideal Simplex and b_1 be as in Lemma 2.1. The next lemma is proved in the supplementary material.

Lemma 2.2. *Let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$, $V = [v_1, \ldots, v_K]$, and $B = \mathrm{diag}(b_1)[\mathbf{1}_K, V']$. If the conditions of Lemma 2.1 hold, then $P = B\Lambda B'$ and $\theta_i = \xi_1(i)/(\pi_i' b_1)$, $1 \le i \le n$.*

After running Mixed-SCORE, we collect the following quantities: (i) the leading eigenvector $\hat{\xi}_1$; (ii) the estimated vertices $\hat{V} = [\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K]$; (iii) a vector $\hat{b}_1$; (v) the estimated mixed membership vectors in $\hat{\Pi} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n]'$. Inspired by Lemma 2.2, we let

$$\hat{P} = \hat{B}\hat{\Lambda}\hat{B}', \qquad \text{and} \qquad \hat{\theta}_i = \hat{\xi}_1(i)/(\hat{\pi}_i'\hat{b}_1), \quad 1 \le i \le n. \tag{2.9}$$

3 Theoretical properties

We state some regularity conditions. Recall that θ_i's are the degree parameters in Model (2.2). Let $\theta_{\max} = \max_i \theta_i$, $\theta_{\min} = \min_i \theta_i$, $\bar{\theta} = n^{-1}\sum_{i=1}^n \theta_i$, and $\bar{\theta}_* = \sqrt{n^{-1}\sum_{i=1}^n \theta_i^2}$. Define

$$err_n = err_n(\Theta) = [(\theta_{\max}^{3/2}\bar{\theta}^{3/2})/(\theta_{\min}\bar{\theta}_*^2)] \cdot \sqrt{\log(n)/(n\bar{\theta}^2)}. \tag{3.10}$$

Assumption 1. $\theta_{\max} \le C$, and $err_n \to 0$.

Here, the interesting range for θ_i is from $O(n^{-1/2})$ (up to a multi-log(n) term) to $O(1)$, so the first condition is mild. To appreciate the second condition, note that when $\theta_{\max} \leq C\theta_{\min}$, $err_n \asymp \sqrt{\log(n)/(n\bar{\theta}^2)}$, where $n\bar{\theta}^2$ is the order of the expected average node degree. Therefore, the condition of $err_n \to 0$ is the same as that the average node degree grows to ∞ faster than $\log(n)$, which is mild. Introduce a $K \times K$ matrix $G = K\|\theta\|^{-2}(\Pi'\Theta^2\Pi)$.

Assumption 2. $\|P\|_{\max} \leq C$, $\|G\| \leq C$, and $\|G^{-1}\| \leq C$.

The first one is seen to be mild. For the other two conditions, it is instructive to consider a special case where all nodes are pure. In this case, $G = K\|\theta\|^{-2} \cdot \mathrm{diag}(\|\theta^{(1)}\|^2, \ldots, \|\theta^{(K)}\|^2)$, where $\|\theta^{(k)}\|^2 = \sum_{i \in \mathcal{C}_k} \theta_i^2$. Therefore, the two conditions reduce to that of $\max_k \|\theta^{(k)}\|^2 \leq C \min_k \|\theta^{(k)}\|^2$, which is only mild. Denote by $\lambda_k(PG)$ the k-th largest right eigenvalue of PG, and by $\eta_k \in \mathbb{R}^K$ the associated right eigenvector, $1 \leq k \leq K$.

Assumption 3. $|\lambda_2(PG)| \leq (1 - c_1)\lambda_1(PG)$, and $c_1\beta_n \leq |\lambda_K(PG)| \leq |\lambda_2(PG)| \leq c_1^{-1}\beta_n$, where $\beta_n \in (0, 1)$ and $c_1 \in (0, 1)$ is a constant.

The first item is a mild eigen-gap condition. In the second item, the quantity β_n captures the 'distinction' between communities and can be interpreted as the "signal strength" of the DCMM model, where $\beta_n = O(1)$ is the case of "strong signal" and $\beta_n = o(1)$ is the case of "weak signal" (β_n is a component in the error rate to be introduced). We assume $\lambda_2, \ldots, \lambda_K$ are at the same order. This is only for convenience and can be relaxed (e.g., $\lambda_2, \ldots, \lambda_K$ split into several groups and those in the same group are at the same order).

Assumption 4. $\min_{1 \leq k \leq K} \eta_1(k) > 0$, and $\frac{\max_{1 \leq k \leq K} \eta_1(k)}{\min_{1 \leq k \leq K} \eta_1(k)} \leq C$.

In Section A.2 of the supplementary material, we show that this assumption is satisfied in either of the following cases: As $n \to \infty$, (a) all entries of PG are lower bounded by a constant, (b) K is fixed and P tends to a fixed irreducible matrix P_0, (c) K is fixed and G tends to a fixed irreducible matrix G_0, and (d) the maximum and minimum row sums of P are at the same order and π_i's are i.i.d. generated from a Dirichlet distribution.

3.1 Large-deviation bounds for $\hat{R}$

The following entry-wise large-deviation bounds for matrix $\hat{R}$ plays a key role in our analysis. Let $\hat{R} = [\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n]'$ be as in (2.5). Let $R = [r_1, r_2, \ldots, r_n]'$ be as in (2.4).

Theorem 3.1 (Large-deviation bounds for $\hat{R}$)**.** *Consider the DCMM model where Assumptions 1-4 hold. Suppose $\sqrt{K\log(n)} \leq T \leq \infty$ for T in (2.5). Let err_n be as in (3.10) and β_n as in Assumption 3. With probability $1 - o(n^{-3})$, there exists an orthogonal matrix $H \in$*

$\mathbb{R}^{K-1,K-1}$ *such that* $\max_{1\le i\le n}\|H\hat{r}_i - r_i\| \le CK^{3/2}\beta_n^{-1}err_n$. *If, additionally,* $\theta_{\max} \le C\theta_{\min}$, *then with probability* $1-o(n^{-3})$, $\max_{1\le i\le n}\|H\hat{r}_i - r_i\| \le CK^{3/2}(n\bar{\theta}^2\beta_n^2)^{-1/2}\sqrt{\log(n)}$.

In Theorem 3.1, $(K, \beta_n, \bar{\theta})$ may all vary with n. Among them, β_n captures the "strength of community signals", where we either have $\beta_n = O(1)$ or $\beta_n \to 0$ reasonably fast, so the claims applies to both the cases of "strong signals" and "weak signals".

The proof of Theorem 3.1 is based on a row-wise large deviation bound for the eigenvectors of the adjacency matrix (Lemma D.2 in the supplement). In the literature, there were few results about row-wise deviation bounds for eigenvectors of a network adjacency matrix (Abbe et al., 2020; Fan et al., 2022, 2020). They focused on moderate degree heterogeneity and assumed that the nonzero population eigenvalues are at the same order, so they do not apply to our setting. We need non-trivial efforts to prove Lemma D.2 and Theorem 3.1.

3.2 Rates of Mixed-SCORE with a generic but efficient VH step

Mixed-SCORE has a plug-in VH step, and the goal of the VH step is to estimate the vertices $v_1, \ldots, v_K$ of the ideal simplex. In this section, we present the rate of Mixed-SCORE for a *generic but efficient* VH step. Next in Section 3.3, we discuss the rate of Mixed-SCORE for all 4 proposed VH step in Table 1 (where the rate can be much faster in some cases).

Definition 1 (*Efficient VH*). We call a VH step efficient if it satisfies that $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\max_{1\le i\le n}\|H\hat{r}_i - r_i\|$, where H is the orthogonal matrix in Theorem 3.1.

For our proposed VH methods in Table 1, CVS and SP are efficient under Assumptions 1-4, and SVS and SVS* are efficient if some additional conditions hold; see Section 3.3.

For any estimate $\hat{\Pi} = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n]'$ for Π, we measure the error by the mean squared error (MSE) $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \|\hat{\pi}_i - \pi_i\|^2]$. Recall that err_n is defined in (3.10).

Theorem 3.2 (Error of Mixed-SCORE)**.** *Consider the DCMM model where Assumptions 1-4 hold. Let* $\hat{\Pi}$ *be the estimate of* Π *by Mixed-SCORE with a generic but efficient VH step. Then,* $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \|\hat{\pi}_i - \pi_i\|^2] \le CK^3\beta_n^{-2}err_n^2 + o(n^{-2})$. *If additionally* $\theta_{\max} \le C\theta_{\min}$, *then* $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n \|\hat{\pi}_i - \pi_i\|^2] \le CK^3(n\bar{\theta}^2\beta_n^2)^{-1}\log(n) + o(n^{-2})$.

We now discuss the implication of Theorem 3.2 on economic applications. For simplicity, we consider a case where $\theta_{\max} \asymp \theta_{\min}$, $K = O(1)$ and $\beta_n \ge C$. By Theorem 3.2, the MSE is $O((n\bar{\theta}^2)^{-1}\log(n))$. For a dense network, $\bar{\theta} \asymp 1$, and the MSE becomes $O(n^{-1}\log(n))$, which is quite negligible. Suppose we have a downstream economic model $y_i = \alpha + \pi_{i(-1)}'\beta + \epsilon_i$, where y_i is an outcome of interest and $\pi_{i(-1)}$ is the sub-vector of π_i by dropping the last

coordinate (to remove co-linearity). We plug in the $\hat{\pi}_i$'s from Mixed-SCORE and let $\hat{\beta}$ be the least-squares coefficient. It can be shown that $|\hat{\beta}-\beta|^2 = O\big(n^{-1}\sum_{i=1}^n \|\hat{\pi}_i-\pi_i\|^2\big)+O_{\mathbb{P}}(n^{-1})$. Therefore, as long as $n\bar{\theta}^2 \gg \log(n)$, we have consistency on $\hat{\beta}$. Furthermore, using the faster rates in Section 3.3, we can further remove the $\log(n)$ factor in MSE; as a result, when the network is dense, we also have root-n consistency of $\hat{\beta}$.

Remark 5 *(Rate optimality).* Jin and Ke (2017) derived a minimax lower bound for the case where K is finite and that θ_i's are equal. They showed that for any estimate $\hat{\Pi}$, there is a constant $c_0 > 0$ such that $\frac{1}{n}\sum_{i=1}^n \|\hat{\pi}_i - \pi_i\|^2 \geq C/(n\bar{\theta}^2\beta_n^2)$ with probability $\geq c_0$. Comparing it with Theorem 3.2, the error rate of Mixed-SCORE is optimal (up to a $\log(n)$ factor) for DCMM with $\theta_{\max} \leq C\theta_{\min}$.

Remark 6 *(Comparison with the rate of the OCCAM algorithm (Zhang et al., 2020)).* Since the theory of OCCAM does not allow $\beta_n = o(1)$ or K diverging with n, we compare two methods only in the case that $K \leq C$ and $\beta_n \geq C$. The rate of Mixed-SCORE reduces to $(n\bar{\theta}^2)^{-1/2}$, but the rate of OCCAM cannot be faster than $(n\bar{\theta}^2)^{-1/5}$, which is strictly slower. Also, OCCAM works only if the fraction of mixed nodes is properly small (hinged in Assumption-B of Zhang et al. (2020)). For example, when $K = 3$, $P = 0.9I_3 + 0.1\mathbf{1}_3\mathbf{1}_3'$, and $\pi_i = \frac{1}{\sqrt{3}}\mathbf{1}_3$ for all mixed nodes, the fraction of mixed nodes has to be $< 1/4$.

Remark 7 *(Comparison with theory of community detection).* Community detection is a less challenging problem, where π_i's are known to be degenerate. It has exponential rates (Gao et al., 2018), but membership estimation only achieves polynomial rates (Jin and Ke, 2017). Consider an example with $K = 2$, $\pi_i \overset{iid}{\sim} \text{Dirichlet}(\alpha_0)$, and $P(A_{ij} = 1) = n^{-1}\pi_i' P \pi_j$, where $P_{km} = a\cdot 1\{k = m\} + b\cdot 1\{k \neq m\}$. As $n \to \infty$, α_0 is fixed but (a, b) can depend on n. This is equivalent to a DCMM with $\bar{\theta} \asymp n^{-1/2}\sqrt{a}$ and $\beta_n \asymp (a-b)/a$. Write $I = (a-b)^2/a$. The rate of Mixed-SCORE is $O(I^{-1/2}\sqrt{\log(n)})$, but when π_i's are all degenerate, the rate of community detection is $\exp(-O(I))$.

Given the results for $\hat{\Pi}$, we further study the estimates $(\hat{\Theta}, \hat{P})$ defined in Section 2.4.

Theorem 3.3 (Estimation of (Θ, P) in DCMM)**.** *Under the conditions of Theorem 3.2, with probability* $1-o(n^{-3})$, $\|\hat{P}-P\| \leq C(K^2+K^{3/2}\beta_n^{-1})err_n$ *and* $\|\hat{\Theta}-\Theta\|_F^2 \leq C\|\theta\|^2 K^3\beta_n^{-2}err_n^2$.

3.3 Rates for Mixed-SCORE with proposed VH steps, and faster rates

Section 3.2 analyzes a generic Mixed-SCORE algorithm with an efficient VH step. In this subsection, we discuss Mixed-SCORE with each specific VH approach in Table 1. First, we consider CVS and SP. The following theorem shows that CVS and SP are both efficient, and Mixed-SCORE-CVS and Mixed-SCORE-SP attain the rate in Theorem 3.2.

Theorem 3.4. *Consider the DCMM model where Assumptions 1-4 hold and each community has at least one pure node. Let H be the orthogonal matrix in Theorem 3.1. If we apply either CVS or SP to rows of $\hat{R}$, then with probability $1-o(n^{-3})$, $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\max_{1\le i\le n}\|H\hat{r}_i - r_i\|$, so both CVS and SP are efficient. Moreover, for Mixed-SCORE-CVS or Mixed-SCORE-SP, $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n\|\hat{\pi}_i-\pi_i\|^2] \le CK^3\beta_n^{-2}err_n^2 + o(n^{-2})$.*

Next, we consider Mixed-SCORE-SVS and Mixed-SCORE-SVS*. SVS and SVS* use a denoise stage, which provides a significant advantage in numerical performance, but also makes them harder to analyze. For this reason, we only consider two settings. In the first setting, we assume all π_i's for mixed nodes are *iid* drawn from a continuous distribution. In the second setting, π_i's form several *loose clusters*. Owing to space limit, we only present Setting 1 here. Setting 2 is in Section B of the supplementary material.

Setting 1. Let $\mathcal{S}_0 = \mathcal{S}_0(e_1, e_2, \ldots, e_K)$ be the standard simplex in $\mathbb{R}^K$, where the vertices $e_1, e_2, \ldots, e_K$ are the standard Euclidean basis vectors of $\mathbb{R}^K$. Fix a density g defined over $\mathcal{S}_0$. Let $\mathcal{R} = \{\pi \in \mathcal{S}_0 : g(\pi) > 0\}$ be the support of g. Suppose there is a constant $c_0 > 0$ such that $\mathcal{R}$ is an open subset of $\mathcal{S}_0$, and distance$(e_k, \mathcal{R}) \ge c_0$, $1 \le k \le K$. Let $\delta_v(\pi)$ be the point mass at $\pi = v$. Fixing constants $\epsilon_1, \ldots, \epsilon_K > 0$ with $\sum_{k=1}^K \epsilon_k < 1$, we invoke a random design model where π_i's are *iid* drawn from $f(\pi) = \sum_{k=1}^K \epsilon_k \cdot \delta_{e_k}(\pi) + \big(1 - \sum_{k=1}^K \epsilon_k\big)\cdot g(\pi)$. The following is similar to err_n in (3.10), and quantifies the "faster rate" aforementioned.

$$err_n^* = err_n^*(\Theta) = [(\theta_{\max}^{1/2}\bar{\theta}^{3/2})/(\theta_{\min}\bar{\theta}_*)]\cdot(n\bar{\theta}^2)^{-1/2}. \tag{3.11}$$

Theorem 3.5. *Consider the DCMM model where Assumptions 1-4 hold and π_i's are as in Setting 1. Let H be as in Theorem 3.1. There exists a constant $L_0(g, \epsilon_1, \ldots, \epsilon_K) > 0$ such that, if we apply SVS or SVS* to rows of $\hat{R}$ with $L \ge L_0$, then with probability $1-o(n^{-3})$, $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\big(n^{-1}\sum_{i=1}^n\|H\hat{r}_i - r_i\|^2\big)^{1/2}$. Moreover, for Mixed-SCORE-SVS or Mixed-SCORE-SVS*, $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n\|\hat{\pi}_i-\pi_i\|^2] \le CK^3\beta_n^{-2}(err_n^*)^2 + o(n^{-2})$.*

By Theorem 3.5, the rates of Mixed-SCORE-SVS and Mixed-SCORE-SVS* are faster than those of Mixed-SCORE-SP and Mixed-SCORE-CVS. In fact, by (3.10)-(3.11), we have $err_n^*/err_n = [\bar{\theta}_*/(\theta_{\max}\sqrt{\log(n)})]$. Since $\bar{\theta}_*/\theta_{\max} \le 1$ and $\bar{\theta}_*/\theta_{\max}$ may tend to 0 rapidly, we have the following observations: 1) The rate here is faster than that of Theorem 3.2 by at least a factor of $\log(n)$. 2) The rate here can be much faster than that of Theorem 3.2 if $\bar{\theta}_*/\theta_{\max} \to 0$ rapidly. As an example, suppose $\theta_1 = \ldots = \theta_{n-1} = \alpha_n$ and $\theta_n = n^\gamma\alpha_n$, where $0 < \gamma < 1/2$ is a constant; in this case, $err_n^*/err_n = \bar{\theta}_*/(\theta_{max}\sqrt{\log(n)}) \le n^{-\gamma}/\sqrt{\log(n)}$, and so the rate here is much faster than that of Theorem 3.2. Once we have a faster rate for $\hat{\Pi}$, we also enjoy a faster rate for the proposed $(\hat{\Theta}, \hat{P})$ in Section 2.4 (proof is omitted).

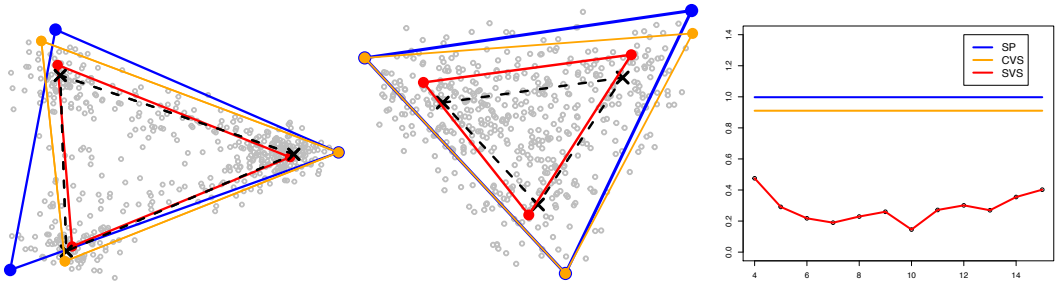


Figure 4: Comparison of VH methods (black: truth; blue: SP; yellow: CVS; red: SVS). Left: The case of weak noise. CVS and SVS perform well, but SP performs less satisfactorily (possible reason: SP is a greedy algorithm). Middle: The case of strong noise. SVS performs well, but SP and CVS perform unsatisfactorily. This is because SVS is much less sensitive to outliers. Right: Robustness of SVS to the choice of L (y-axis is $\max_k \|H\hat{v}_k - v_k\|^2$).

Remark 8. The faster rates here are because SVS and SVS* use a denoise stage, which improves the accuracy in vertex hunting and so in membership estimation. The improved rate is not due to the more strict setting considered here (in fact, in Setting 1 and Setting 2, if we use SP and CVS for VH in Mixed-SCORE, then we do not have a much faster rate). For more general settings, Mixed-SCORE-SVS or Mixed-SCORE-SVS* continue to enjoy this faster rate, as supported by numerical experiments in Section 4.

4 Simulations

Experiment 1 *(Comparison of VH approaches)*. We view Mixed-SCORE as a generic algorithm, where we can plug in any VH approach. In Table 1, we list four VH approaches. We now compare SP, CVS and SVS (the performance of SVS* is very similar to SVS, thus omitted). Fix $(n, K) = (500, 3)$. P is a matrix whose diagonals are 1 and off-diagonals are 0.3. Each community has 50 pure nodes. For π_i's of the remaining 350 nodes, half of them are *iid* drawn from Dirichlet$(0.6, 0.2, 0.2)$, and half are *iid* drawn from Dirichlet$(0.3, 0.4, 0.3)$. We consider two cases: (a) Weak noise ($\theta_i \equiv 0.7$, and the network is denser) (b) Strong noise ($\theta_i \equiv 0.4$, and the network is sparser). We choose L as in (2.8), but we also investigate SVS for all $L \in \{4, 5, 6, \ldots, 15\}$. We report the average of $\max_k \|H\hat{v}_k - v_k\|^2$ over 100 repetitions. The results are in Figure 4. We observe the following: (i) In the strong signal case, three methods perform similarly. (ii) In the weak signal case, CVS and SP are significantly worse than SVS. (iii) The performance of SVS is insensitive to the choice of L. The results confirm our claims in Section 2.3 and Section 3.3 that the de-noise stage in SVS plays a crucial role in improving the numerical performance.

Experiments 2-4 *(Performance of Mixed-SCORE-SVS)*. From now on, we fix the VH

approach as SVS. The tuning integer L is chosen from data using (2.8). In the literature, other mixed membership estimation approaches only work for MMSBM. The only exception is OCCAM Zhang et al. (2020). OCCAM assigns to each node a non-negative "membership" vector with unit ℓ_2-norm; we renormalize them by their ℓ_1-norms and use them as the estimated π_i. Fix $n = 500$ and $K = 3$. For $0 \leq n_0 \leq 160$, let each community have n_0 number of pure nodes. Fixing $x \in (0, 1/2)$, let the mixed nodes have four different memberships $(x, x, 1-2x)$, $(x, 1-2x, x)$, $(1-2x, x, x)$ and $(1/3, 1/3, 1/3)$, each with $(500-3n_0)/4$ number of nodes. Given $\rho \in (0, 1)$, P has diagonals 1 and off-diagonals ρ. Fixing $z \geq 1$, we generate the degree parameters such that $1/\theta_i \overset{iid}{\sim} U(1, z)$, where $U(1, z)$ denotes the uniform distribution on $[1, z]$. The tuning parameter L is selected as in (2.8). For each parameter setting, we report $n^{-1}\sum_{i=1}^{n} \|\hat{\pi}_i - \pi_i\|^2$ averaged over 100 repetitions.

Experiment 2 (fraction of pure nodes). Fix $(x, \rho, z) = (0.4, 0.1, 5)$ and let n_0 range in $\{40, 60, 80, 100, 120, 160\}$. As n_0 increases, the fraction of pure nodes increases from around 25% to around 95%. See Figure 5 (left). When the fraction of pure nodes is $< 70\%$, Mixed-SCORE significantly outperforms OCCAM; when the fraction of pure nodes is $> 70\%$, the two methods have similar performance.

Experiment 3 (purity of mixed nodes). We call $\max_{1\leq k\leq K}\{\pi_i(k)\}$ the "purity" of node i. Fix $(n_0, \rho, z) = (80, 0.1, 5)$ and let x range in $\{0.05, 0.1, 0.15, \cdots, 0.5\}$. In our settings, there are four types of mixed nodes. For the first three types, their purity is $(1-2x)1\{x \leq 1/3\} + x1\{x > 1/3\}$. Therefore, as x increases to $1/3$, these nodes become less pure; then, as x further increases, these nodes become more pure. See Figure 5 (middle). It suggests that membership estimation is harder as the purity of mixed nodes decreases. Mixed-SCORE outperforms OCCAM in almost all settings, especially when x is close to $1/3$.

Experiment 4 (degree heterogeneity). Fix $(x, n_0, \rho) = (0.4, 80, 0.1)$ and let z range in $\{1, 2, \cdots, 8\}$. Since $1/\theta_i \overset{iid}{\sim} U(1, z)$, a larger z means the lower average degree and more severe degree heterogeneity (so the problem is harder). See Figure 5 (right). Mixed-SCORE uniformly outperforms OCCAM. Interestingly, when z is small (so the problem is "easy"), Mixed-SCORE is very accurate, but the performance of OCCAM is unsatisfactory.

Experiments 5-8. For space limit, we have relegated them to the supplement. Experiment 5 studies settings where the matrix P varies. Experiment 6 studies settings where π_i's drawn from a continuous distribution. Experiment 7 further investigates robustness of Mixed-SCORE-SVS to the choice of L. Experiment 8 compares Mixed-SCORE with the latent space modeling of networks Handcock et al. (2007).

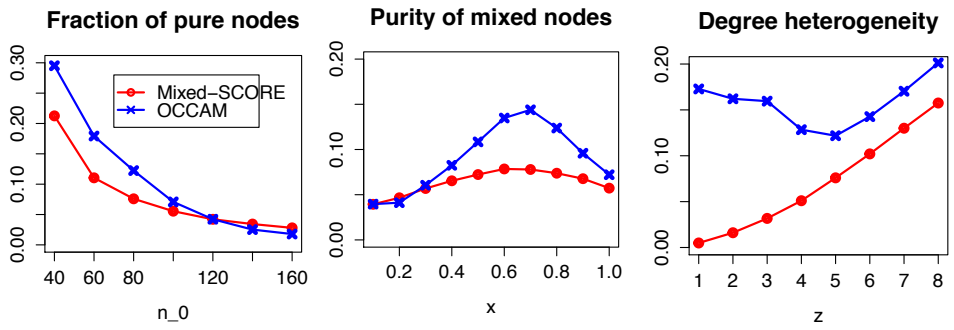


Figure 5: Estimation errors of Mixed-SCORE and OCCAM (y-axis: $n^{-1}\sum_{i=1}^{n}\|\hat{\pi}_i - \pi_i\|^2$).

5 Real data applications

5.1 The international trade networks and the trade triangles

There are two lines of literature on the analysis of international trade networks. The first is the gravity model (Anderson and Van Wincoop, 2003). It fits a generalized linear model for trade flows using countrywise 'size' covariates and pairwise 'trading cost' covariates. The second is in physics, which studies the topology of trade networks (Serrano and Boguná, 2003). Mixed-SCORE is useful in both approaches.

Combination of Mixed-SCORE and gravity models. Let $X(i,j)$ be the trade flow from country i to country j. The (general) gravity model assumes $X(i,j) \sim \text{Poisson}(\lambda(i,j))$, with $\ln(\lambda(i,j)) = \sum_{m=1}^{M}\alpha_m G_m(i) + \sum_{m=1}^{M}\beta_m G_m(j) + \sum_{s=1}^{S}\gamma_s D_s(i,j) + c_i + c_j$, where $G_1,\ldots,G_M$ are the (log) 'size' covariates, $D_1,\ldots,D_S$ are the (log) 'trading cost' covariates, and c_i's are the fixed effects of countries. We fit this model using Poisson pseudo maximum likelihood and let $\hat{\lambda}(i,j)$ denote the fitted value. We define two 'p-values' for each country pair: $Q_1(i,j) = \mathbb{P}(\text{Poisson}(\hat{\lambda}(i,j)) > X(i,j))$ and $Q_2(i,j) = \mathbb{P}(\text{Poisson}(\hat{\lambda}(i,j)) < X(i,j))$. A small value of $Q_1(i,j)$ implies that the observed trade flow is significantly higher than the fitted one, and a small value of $Q_2(i,j)$ indicates the opposite. We construct two undirected networks. In the first one, there is an edge between nodes i and j if $\min\{Q_1(i,j), Q_1(j,i)\} < 0.05$. In the second network, edges are defined similarly except that Q_1 is replaced by Q_2. We call them the *gravity-under-shooting (GUS)* network and *gravity-over-shooting (GOS)* network, respectively. For each network, we apply Mixed-SCORE to obtain $(\hat{\Pi}, \hat{\Theta}, \hat{P})$ and then construct a new nodal covariate, $U(i) = \ln(\hat{\theta}(i))$, and a new dyadic covariate, $H(i,j) = \ln(\hat{\pi}_i'\hat{P}\hat{\pi}_j)$. We use them as surrogates of those unobserved covariates in the gravity model and plug them back to re-fit the gravity model. As explained in Example 3 of Section 1.3, we assume here that the unobserved covariates have a DCMM-like structure, which has the same spirit as the model in Graham (2015). Our proposed 'Mixed-SCORE + refitting' is a

Table 2: Combination of Mixed-SCORE and gravity model. The bigger model has two new covariates created by Mixed-SCORE. The F statistic for model comparison is 928.56 (p-value $<$ 2.2e-16). We note that these coefficients are not supposed to be directly compared with the fitted coefficients in Column 2 of Table 2 in Head et al. (2010), because they use panel data but we only use one year's data (this also explains why our standard errors are considerably smaller).

		Before		After	
Covariate	Meaning	Coef.	Pval	Coef.	Pval
distw	weighted distance	-.832 (.012)	<2e-16 ***	-.722 (.011)	<2e-16 ***
rta	regional trade agreement dummy	.429 (.026)	<2e-16 ***	.429 (.022)	<2e-16 ***
contig	contiguity dummy	.415 (.022)	<2e-16 ***	.403 (.019)	<2e-16 ***
comlang_off	common official language dummy	.242 (.022)	<2e-16 ***	.181 (.019)	<2e-16 ***
comcur	common currency dummy	-.167 (.031)	7e-08 ***	.005 (.027)	.852
dyadic_GUS	new trade cost covariate (GUS)			1.294 (.033)	<2e-16 ***
dyadic_GOS	new trade cost covariate (GOS)			-.337 (.037)	<2e-16 ***

proxy approach to fitting the model we introduce there.

To test the performance of our approach, we use an edited version of the gravity data set in Head et al. (2010) (available in the R package `gravity`). The original data set contains the bilateral trade flows for 166 countries in 1948-2006. We only use the data in 2006. This edited version includes a nodal covariate, *gdp*, and five dyadic covariates, *distw*, *rta*, *contig*, *comlang_off* and *comcur* (their meanings are in Column 2 of Table 2). Compared with the original gravity model fitting in Head et al. (2010), this edited version does not provide all covariates, so it serves as a good example of *unobserved covariates*. Since there is only one year of data, we did not include any nodal covariate, because their effects will be absorbed into the fixed effect c_i; all five dyadic covariates were included. We constructed the GUS and GOS networks as above and ran Mixed-SCORE separately on these two networks. We set $K = 3$ for both networks.[5] It gave rise to two new dyadic covariates H^{GUS} and H^{GOS} (again, we did not include the new nodal covariates because of the fixed effects c_i). The results are in Table 2, where both new covariates created by Mixed-SCORE are significant. The other coefficients have mild changes and slightly smaller standard errors after re-fitting, except the coefficient of *comcur*. Initially, the coefficient of *comcur* is negative, with a very small p-value. This contradicts our common sense: sharing common currency should not have a significantly negative impact on trading. After adding the Mixed-SCORE covariates, the coefficient of *cumcur* becomes positive and insignificant. It suggests that our proposed approach is potentially useful in correcting the bias caused by unobserved covariates.

To appreciate what information Mixed-SCORE captures, we check the rows of $\hat{R}$ for the GUS and GOS networks. Owing to space limit, we only discuss the GUS network here but relegate the results of the GOS network to the supplementary material (see Section H). The

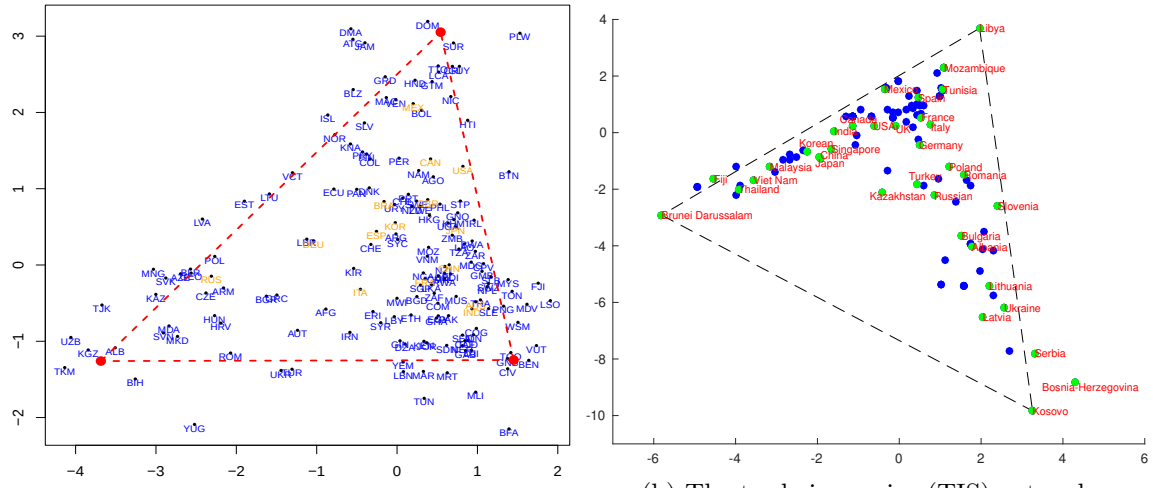
[5] We also tried other values of K. For different K, the networks and Mixed-SCORE output are different, but the newly created covariates and the subsequent gravity model fitting are similar.

edges in the GUS network indicate significant under-estimation of trade flows in the initial gravity model. Therefore, if $\hat{r}_i$ and $\hat{r}_j$ are close, the two countries may have unmodeled connections that benefit trade. The rows of $\hat{R}$ and the estimated simplex (which is a triangle since $K = 3$) for GUS are shown in Figure 6a. We have some observations: (a) The 3 vertices may be interpreted as *Caribbean* (top), *Former Soviet Union* (bottom left), and *Western African* (bottom right). (b) United States, Canada and Mexico are close. These countries are in the North American Free Trade Agreement (NAFTA). The benefit of NAFTA cannot be fully captured by the regional trade agreement dummy *rta* (Anderson and Yotov, 2016) and is further revealed in the covariates created by Mixed-SCORE. (c) United States and Russia are far away from each other - a consequence of the historical confrontation between two countries (Hufbauer and Oegg, 2003). (d) High-GDP countries tend to be in the interior of the triangle (i.e., they have low 'trading costs' with many countries). This is consistent with economic theory that good 'tradability' can boost economic growth (Waugh, 2010). (e) United States (with the highest GDP) is not in the deep interior of the triangle but on an edge. Interestingly, this position is farthest from the *Former Soviet Union* vertex.

Remark 9. In re-fitting the gravity model, an alternative approach is replacing $H(i, j)$ by $\ln(\widehat{\Omega}_{ij})$, where $\widehat{\Omega}$ is an arbitrary estimate of Ω. Using the output of Mixed-SCORE, we can obtain an estimate $\widehat{\Omega}^{\mathrm{MS}}$ by $\widehat{\Omega}^{\mathrm{MS}}_{ij} = \hat{\theta}_i\hat{\theta}_j \cdot \hat{\pi}_i'\hat{P}\hat{\pi}_j$. Since $\hat{\theta}_i$ and $\hat{\theta}_j$ will be absorbed into the fixed effects, this approach is equivalent to the approach we have used above. However, we may plug in a different estimate of Ω, such as $\widehat{\Omega}^{\mathrm{PCA}} = \sum_{k=1}^{K} \hat{\lambda}_k\hat{\xi}_k\hat{\xi}_k$, where $\hat{\lambda}_k$ and $\hat{\xi}_k$ are the kth eigenvalue and eigenvector of A. In Section I of the supplementary material, we compare the two estimates of Ω and find that $\widehat{\Omega}^{\mathrm{MS}}$ has much better numerical performance. The reason is that $\widehat{\Omega}^{\mathrm{MS}}$ utilizes the DCMM model structure, not just low-rankness of Ω.

Remark 10. In the recent literature of gravity modeling of trade data, it has become common to use panel data and to include the importer-year and exporter-year fixed effects (Weidner and Zylkin, 2021). We did not use panel data because Mixed-SCORE only applies to static networks. In a working paper, we extend Mixed-SCORE to dynamic networks. It will be useful for analysis of panel data. We leave this to future work.

Remark 11. In the analysis of panel data, an interesting approach is using the pairwise fixed effects (Weidner and Zylkin, 2021) to account for unobserved covariates. However, for our example here where we only use one year's data, this approach will introduce $n(n-1)/2$ free parameters, but we only have $n(n-1)$ observed trading flows; therefore, this approach will have the issue of over-fitting. In comparison, our Mixed-SCORE approach only allows for $O(nK)$ free parameters and does not have this over-fitting issue.



(a) The GUS network after gravity model fitting.

(b) The trade in service (TIS) network

Figure 6: Rows of $\hat{R}$ and the estimated simplex ($K = 3$, so the simplex is a triangle). Left: Orange dots are top 15 countries with highest GDPs. Right: Green dots are 35 manually-picked economies.

Using Mixed-SCORE for network analysis of the world trade web. Studying the network topology of the world trade web is a problem of interest (Serrano and Boguná, 2003). These works do not require observing any covariates. They build networks directly from trade flows and study the topology of these networks (e.g., power law degree distribution, latent community structure, centrality metric, clustering coefficient, etc.). We will show that Mixed-SCORE is useful for creating low-dimensional embeddings of countries in these networks. We downloaded the trade in services data from `https://data.wto.org/`. For each pair of economies (i, j), we aggregated the total service export from economy i to economy j during 2014-2018 (we used the numbers reported by economy i). There are 202 economies in total, but we removed *European Union* and *Extra EU Trade*, as their data partially overlap with the data of individual countries. This gave rise to a 200×200 weight matrix X. We symmetrize X to $Y = (X + X')/2$. Let $u = (u_1, u_2, \ldots, u_{200})'$ contain the row sums of Y. Define $Z = [\mathrm{diag}(u)]^{-1/2} Y [\mathrm{diag}(u)]^{-1/2}$, where each entry of Z is in $[0, 1]$. [6] Let μ and σ be the mean and standard deviation of all nonzero entries of Z. We construct an undirected network, where each economy is a node and there is an edge between i and j if and only if $Z(i, j) \geq \mu + \sigma$. We restrict it to the giant component, which has $n = 116$ nodes. We call this network the trade-in-service (TIS) network. We applied Mixed-SCORE

[6] One may use GDP or population to normalize, but here we are primarily interested in the case with no observed covariates. We follow the literature to use total trade flows to normalize.

with $K = 3$.[7] The rows of $\hat{R}$ are displayed in Figure 6b.[8] This creates an embedding of all economies into a 2-dimensional latent space. We have some noteworthy observations. (a) The point cloud fits well with a triangle, which we call the 'trade triangle'. The three vertices may be interpreted as three different regions: 'North Africa' (top vertex in Figure 6b), 'Southeast Asia' (bottom left vertex), and 'Central/South Europe' (bottom left vertex). (b) It agrees to economic theory that geographic proximity plays a key role in trade. In Figure 6b, countries that are geographically close tend to cluster together; e.g., countries in Southeast Asia (*Thailand*, *Vet Nam*, *Malaysia*, etc.), East Asia (*China*, *Japan*, *Korea*, etc.), North America (*USA*, *Canada*, *Mexico*, etc.), West Europe (*UK*, *France*, *Germany*, etc.), East Europe and West/Central Asia (*Russian*, *Kazakhstan*, *Turkey*, *Bulgaria*, etc.) and so on. (c) The node embedding contains more information than geographical proximity. For example, *Singapore* is geographically close to Southeast Asian countries, but it is closer to East Asian countries in the trade triangle; West European countries are geographically closer to East European countries, but they are closer to North American countries in the trade triangle. These can be explained by trading agreements and historical trading relationships. The above supports that Mixed-SCORE is useful for node embedding. Imagine that we are given the trade flows of a new product or service, with little known information; we can apply Mixed-SCORE to visualize the locations of countries in the embedded space and gain useful insights for next-step modeling.

5.2 The coauthor and citee network of statisticians, and Fan's group

The study of coauthorship networks and citation networks is common in applied social science (Barabâsi et al., 2002). The goal is using scientific publications in a field to study the development of the field itself. It is useful for discovering whether all sub-areas ('communities') are developed in a healthy and balanced way and whether any particular sub-area is under-developed and needs more allocation of resources (Foster et al., 2015). For example, Andrikopoulos et al. (2016) studied the coauthorship network for *Journal of Econometrics*. In this subsection, we use a data set from Ji and Jin (2016). It consists of bibtex and citation data of $3,248$ papers published in four top-tier statistics journals, *Annals of Statistics*, *Biometrika*, *Journal of American Statistical Association*, and *Journal of Royal Statistical*

[7] For the adjacency matrix, the scree plot shows the elbow point is either at $K = 3$ or $K = 4$. We applied Mixed-SCORE with both $K = 3$ and $K = 4$. It turns out that for $K = 3$, the plot of the rows of $\hat{R}$ (see (2.5)) fits better with the simplex structure, and the results are easier to interpret, so we choose $K = 3$. Furthermore, we set $T = 2\log(n)$ and $L = 25$ in Mixed-SCORE.

[8] The point associated with *Montenegro* is far away from the data cloud, which we treat as an outlier and do not show in the figure.

Society -Series B, during 2003–2012.

The coauthorship network. Ji and Jin (2016) defined a coauthorship network, where each node is an author, and two authors have an edge if they coauthored 2 or more papers in the data range. The giant component of the network contains 236 authors. Ji and Jin (2016) suggest that this is the "High Dimensional Data Analysis" group, which has a "Carroll-Hall" sub-group (including researchers in nonparametric and semi-parametric statistics and functional estimation) and a "North Carolina" sub-group (including researchers from Duke, North Carolina, and NCSU). In light of this, we consider a DCMM model assuming (a) there are $K = 2$ communities called "Carroll-Hall" and "North Carolina" respectively, and (b) some nodes have mixed memberships in two communities. We applied Mixed-SCORE, and the results are in Table 3. It was argued in Ji and Jin (2016) that the "Fan" group (Jianqing Fan and collaborators) has strong ties to both communities. Our results confirm such a finding but shed new light on the "Fan" group: many of the nodes (e.g., Yingying Fan, Rui Song, Yichao Wu, Chunming Zhang, Wenyang Zhang) have highly mixed memberships, and for each mixed node, we can quantify its weights in two communities. For example, both Runze Li (former graduate of UNC-Chapel Hill) and Jiancheng Jiang (former post-doc at UNC-Chapel Hill and current faculty member at UNC-Charlotte) have mixed memberships, but Runze Li is more on the "Carroll-Hall" community (weight: 73%) and Jiancheng Jiang is more on the "North Carolina" community (weight: 62%).

Table 3: Left and Middle: high-degree pure nodes in the "Carroll-Hall" community and the "North Carolina" community. Right: highly mixed nodes (data: Coauthorship network).

Name	Deg.	Name	Deg.	Name	Deg.	Estimated PMF
Peter Hall	21	Joseph G Ibrahim	14	Jianqing Fan	16	54% of Carroll-Hall
Raymond J Carroll	18	David Dunson	8	Jason P Fine	5	54% of Carroll-Hall
T Tony Cai	10	Donglin Zeng	7	Michael R Kosorok	5	57% of Carroll-Hall
Hans-Georg Muller	7	Hongtu Zhu	7	J S Marron	4	55% of North Carolina
Enno Mammen	6	Alan E Gelfand	5	Hao Helen Zhang	4	51% of North Carolina
Jian Huang	6	Ming-Hui Chen	5	Yufeng Liu	4	52% of North Carolina
Yanyuan Ma	5	Bing-Yi Jing	4	Xiaotong Shen	4	55% of North Carolina
Bani Mallick	4	Dan Yu Lin	4	Kung-Sik Chan	4	55% of North Carolina
Jens Perch Nielsen	4	Guosheng Yin	4	Yichao Wu	3	51% of Carroll-Hall
Marc G Genton	4	Heping Zhang	4	Yacine Ait-Sahalia	3	51% of Carroll-Hall
Xihong Lin	4	Qi-Man Shao	4	Wenyang Zhang	3	51% of Carroll-Hall
Aurore Delaigle	3	Sudipto Banerjee	4	Howell Tong	2	52% of North Carolina
Bin Nan	3	Amy H Herring	3	Chunming Zhang	2	51% of Carroll-Hall
Bo Li	3	Bradley S Peterson	3	Yingying Fan	2	52% of North Carolina
Fang Yao	3	Debajyoti Sinha	3	Rui Song	2	52% of Carroll-Hall
Jane-Ling Wang	3	Kani Chen	3	Per Aslak Mykland	2	52% of North Carolina
Jiashun Jin	3	Weili Lin	3	Bee Leng Lee	2	54% of Carroll-Hall

The citee network. Ji and Jin (2016) also defined a citee network: there is an edge between two authors if they have been cited at least once by the same author (other than themselves). The giant component of this network contains 1790 authors. Ji and Jin (2016) suggested that the network has three meaningful communities: "Large Scale Multiple Testing" (MulTest), "Spatial and Nonparametric Statistics" (SpatNon) and "Variable Selection"

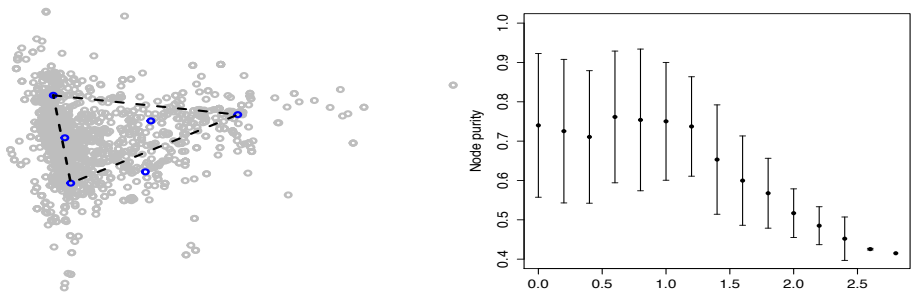


Figure 7: Left: rows of $\hat{R}$ and the estimated simplex. Right: node purity v.s. degree; x-axis is $\hat{\theta}(i)$ (grouped with an interval of .2; we plot the mean and standard deviation of $\|\hat{\pi}_i\|_\infty$ in each group).

Table 4: Estimated PMF of the 12 nodes with the highest degrees in the Citee network.

Name	Deg.	MulTest	SpatNon	VarSelect	Name	Deg.	MulTest	SpatNon	VarSelect
Jianqing Fan	977	0.365	0.220	0.415	Peter Buhlmann	742	0.527	0.121	0.352
Raymond Carroll	850	0.282	0.294	0.424	Hans-Georg Muller	714	0.413	0.237	0.350
Hui Zou	824	0.348	0.225	0.427	Yi Lin	693	0.417	0.137	0.446
Peter Hall	780	0.501	0.032	0.467	Nocolai Meinshausen	692	0.462	0.125	0.413
Runze Li	778	0.282	0.226	0.491	Peter Bickel	692	0.529	0.216	0.255
Ming Yuan	748	0.391	0.166	0.444	Jian Huang	677	0.572	0	0.428

(VarSelect). We thereby set $K = 3$ and apply Mixed-SCORE. Figure 7 (left) plots the rows of $\hat{R} \in \mathbb{R}^{n,2}$, where a simplex (triangle) is clearly visible in the cloud. Table 4 shows the estimated PMF of high degree nodes (please also see Table 3 in the supplementary material). The results confirm those in Ji and Jin (2016) (especially on the existence of three communities aforementioned), but also shed new light on the network. First, high-degree nodes in VarSelect are frequently observed to have an interest in MulTest, and this is not true the other way around (e.g., compare *Jianqing Fan*, *Hui Zou* with *Yoav Benjamini*, *Joseph Romano*). Second, the citations from SpatNon to either MulTest or VarSelect are comparably lower than those between MulTest and VarSelect. This fits well with our impression. Conceivably, a node with higher degree tends to be more senior and so tends to be more mixed. Figure 7 (right) is the plot of the node purity, $\max_{1\le k\le K}\{\hat{\pi}_i(k)\}$, versus the estimated degree heterogeneity parameter $\hat{\theta}(i)$. The results show a clear negative correlation between two quantities (especially on the right end, which corresponds to nodes with high degrees), which indicates that nodes with higher degrees tend to be more mixed.

6 Discussion

There have been independent interests on networks from both the econometric literature and the statistical literature. Recently, the use of statistical network models in economic problems has received increasingly more attention. However, the statistical models used in network econometrics are largely limited to the classical models, such as SBM and graphon.

Recent developments in statistical network analysis have suggested new ideas in network modeling, but such ideas are largely unknown in the area of network econometrics. In this paper, we make two contributions: 1) We provide a new tool for estimating community structure and creating nodal features from network data. 2) We offer a new network model that accommodates severe degree heterogeneity and mixed memberships and is more suitable for real data; we also equip it with a fast spectral algorithm for estimating parameters of this model. For many existing works in network econometrics that use SBM or graphon as the network model, we may improve the results by using the more realistic DCMM model introduced here. This will inspire interesting future research.

The design of our algorithm includes several novel ideas, e.g., discovering the simplex structure in the spectral domain and the correct steps to estimate Π from the simplex. We have also proposed new vertex hunting algorithms, which have much better numerical performance than the existing algorithms such as successive projection. Theoretically, we derive the explicit error bounds for $\hat{\Pi}$ and show that it is rate-optimal under some conditions.

For future research, first, it is unclear how to estimate K from data. Jin et al. (2022) proposed a stepwise goodness-of-fit procedure for estimating K when there is no mixed membership (i.e., the DCMM model reduces to DCBM). It is an interesting question how to combine Mixed-SCORE with this approach for estimating K under DCMM. Second, we mention several applications of our work in network econometrics (see Section 1.3). It is of great interest to study each application more carefully. For example, can we get a theoretical guarantee for using Mixed-SCORE in these problems? We briefly discuss it in the paragraph below Theorem 3.2, but more rigorous theoretical studies are needed. We leave these open problems to future work.

Data and code: The code for implementing Mixed-SCORE and different VH algorithms is available at `https://github.com/ZhengTracyKe/MixedSCORE`. This link also contains all the real networks used in this paper.

References

Abbe, Emmanuel, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong, 2020, Entrywise eigenvector analysis of random matrices with low expected rank, *Ann. Statist.* 48, 1452.

Adamic, Lada A, and Natalie Glance, 2005, The political blogosphere and the 2004 us election: divided they blog, in *Proceedings of the 3rd international workshop on Link discovery*, 36–43.

Airoldi, Edoardo, David Blei, Stephen Fienberg, and Eric Xing, 2008, Mixed membership stochastic blockmodels, *J. Mach. Learn. Res.* 9, 1981–2014.

Anderson, James E, and Eric Van Wincoop, 2003, Gravity with gravitas: A solution to the border puzzle, *Am. Econ. Rev.* 93, 170–192.

Anderson, James E, and Yoto V Yotov, 2016, Terms of trade and global efficiency effects of free trade agreements, 1990–2002, *J. Internat. Econ.* 99, 279–298.

Andrikopoulos, Andreas, Aristeidis Samitas, and Konstantinos Kostaris, 2016, Four decades of the Journal of Econometrics: Coauthorship patterns and networks, *J. Econometrics* 195, 23–32.

Araújo, Mário César Ugulino, Teresa Cristina Bezerra Saldanha, Roberto Kawakami Harrop Galvao, Takashi Yoneyama, Henrique Caldas Chame, and Valeria Visani, 2001, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, *Chemometrics and Intelligent Laboratory Systems* 57, 65–73.

Auerbach, Eric, 2022, Identification and estimation of a partially linear regression model using network data, *Econometrica* 90, 347–365.

Barabâsi, Albert-Laszlo, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek, 2002, Evolution of the social network of scientific collaborations, *Physica A: Statistical mechanics and its applications* 311, 590–614.

Bickel, Peter J, and Aiyou Chen, 2009, A nonparametric view of network models and newman–girvan and other modularities, *Proc. Nat. Acad. Sci.* 106, 21068–21073.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, 2009, Identification of peer effects through social networks, *Journal of Econometrics* 150, 41–55.

Chen, Elynn Y, Jianqing Fan, and Xuening Zhu, 2020, Community network auto-regression for high-dimensional time series, *arXiv:2007.05521* .

Fan, Jianqing, Yingying Fan, Xiao Han, and Jinchi Lv, 2020, Asymptotic theory of eigenvectors for random matrices with diverging spikes, *J. Amer. Statist. Soc.* 1–14.

Fan, Jianqing, Yingying Fan, Xiao Han, and Jinchi Lv, 2022, SIMPLE: Statistical inference on membership profiles in large networks, *J. R. Stat. Soc. Ser. B. (to appear)* .

Foster, Jacob G, Andrey Rzhetsky, and James A Evans, 2015, Tradition and innovation in scientists' research strategies, *American Sociological Review* 80, 875–908.

Gao, Chao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou, 2018, Community detection in degree-corrected block models, *Ann. Statist.* 46, 2153–2185.

Graham, Bryan S, 2015, Methods of identification in social networks, *Annu. Rev. Econ.* 7, 465–485.

Graham, Bryan S, 2020, Network data, in *Handbook of Econometrics*, volume 7, 111–218 (Elsevier).

Gregory, Steve, 2010, Finding overlapping communities in networks by label propagation, *New J. Phys.* 12, 103018.

Handcock, Mark, Adrian Raftery, and Jeremy Tantrum, 2007, Model-based clustering for social networks, *J. Roy. Statist. Soc. A* 170, 301–354.

Head, Keith, Thierry Mayer, and John Ries, 2010, The erosion of colonial trade linkages after independence, *J. Internat. Econ.* 81, 1–14.

Hindman, Matthew, Kostas Tsioutsiouliklis, and Judy A Johnson, 2003, Googlearchy: How a few heavily-linked sites dominate politics on the web, in *Annual Meeting of the Midwest Political Science Association*, volume 4, 1–33, Citeseer.

Hufbauer, Gary, and Barbara Oegg, 2003, The impact of economic sanctions on us trade: Andrew rose's gravity model, Technical report, Peterson Institute for International Economics.

Jackson, Matthew O, and Asher Wolinsky, 1996, A strategic model of social and economic networks, *Journal of Economic Theory* 71, 44–74.

Ji, Pengsheng, and Jiashun Jin, 2016, Coauthorship and citation networks for statisticians (with discussion), *Ann. Appl. Statist.* 10, 1779–1812.

Jin, Jiashun, 2015, Fast community detection by SCORE, *Ann. Statist.* 43, 57–89.

Jin, Jiashun, and Zheng Tracy Ke, 2017, A sharp lower bound for mixed-membership estimation, *arXiv:1709.05603* .

Jin, Jiashun, Zheng Tracy Ke, and Shengming Luo, 2021a, Optimal adaptivity of signed-polygon statistics for network testing, *Ann. Statist.* 49, 3408–3433.

Jin, Jiashun, Zheng Tracy Ke, Shengming Luo, and Minzhe Wang, 2022, Optimal estimation of the number of communities, *J. Amer. Statist. Soc. (to appear)* .

Jin, Jiashun, Shengming Luo, and Zheng Tracy Ke, 2021b, Improvements on SCORE, especially for weak signals, *Sankhya A* .

Karrer, Brian, and Mark Newman, 2011, Stochastic blockmodels and community structure in networks, *Phys. Rev. E* 83, 016107.

Lovász, László, and Balázs Szegedy, 2006, Limits of dense graph sequences, *J. Combin. Theory Ser. B* 96, 933–957.

Manski, Charles F, 1993, Identification of endogenous social effects: The reflection problem, *The review of economic studies* 60, 531–542.

Miyauchi, Yuhei, 2016, Structural estimation of pairwise stable networks with nonnegative externality, *Journal of econometrics* 195, 224–235.

Newman, Mark EJ, 2003, The structure and function of complex networks, *SIAM review* 45, 167–256.

Pensky, Marianna, 2019, Dynamic network models and graphon estimation, *Ann. Statist.* 47, 2378–2403.

Serrano, Ma Angeles, and Marián Boguná, 2003, Topology of the world trade web, *Phys. Rev. E* 68, 015101.

Tinbergen, Jan, 1962, *Shaping the world economy; suggestions for an international economic policy* (The Twentieth Century Fund, New York).

Waugh, Michael E, 2010, International trade and income differences, *Am. Econ. Rev.* 100, 2093–2124.

Weidner, Martin, and Thomas Zylkin, 2021, Bias and consistency in three-way gravity models, *Journal of International Economics* 132, 103513.

Zhang, Yuan, Elizaveta Levina, and Ji Zhu, 2020, Detecting overlapping communities in networks using spectral methods, *SIAM J. Math. of Data Sci,* 2, 265–283.