SIMPLE: Statistical Inference on Membership Profiles in Large Networks *

Jianqing Fan[1], Yingying Fan[2], Xiao Han[3] and Jinchi Lv[2]

Princeton University[1], University of Southern California[2]
and University of Science and Technology of China[3]

May 3, 2021

Abstract

Network data is prevalent in many contemporary big data applications in which a common interest is to unveil important latent links between different pairs of nodes. Yet a simple fundamental question of how to precisely quantify the statistical uncertainty associated with the identification of latent links still remains largely unexplored. In this paper, we propose the method of statistical inference on membership profiles in large networks (SIMPLE) in the setting of degree-corrected mixed membership model, where the null hypothesis assumes that the pair of nodes share the same profile of community memberships. In the simpler case of no degree heterogeneity, the model reduces to the mixed membership model for which an alternative more robust test is also proposed. Both tests are of the Hotelling-type statistics based on the rows of empirical eigenvectors or their ratios, whose asymptotic covariance matrices are very challenging to derive and estimate. Nevertheless, their analytical expressions are unveiled and the unknown covariance matrices are consistently estimated. Under some mild regularity conditions, we establish the exact limiting distributions of the two forms of SIMPLE test statistics under the null hypothesis and contiguous alternative hypothesis. They are the chi-square

*Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA (E-mail: jqfan@princeton.edu). Yingying Fan is Centennial Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *fanyingy@marshall.usc.edu*). Xiao Han is Professor, International Institute of Finance, Department of Statistics and Finance, University of Science and Technology of China, Hefei, China 230026 (E-mail: *xhan011@ustc.edu.cn*). Jinchi Lv is Kenneth King Stonier Chair in Business Administration and Professor, Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089 (E-mail: *jinchilv@marshall.usc.edu*). This work was supported by NIH grant R01-GM072611-16, NSF grants DMS-1712591, DMS-1953356, DMS-2052926, and DMS-2052964, NSF CAREER Award DMS-1150318, a grant from the Simons Foundation, Adobe Data Science Research Award, and a grant from NSF of China (No.12001518). Corresponding authors: Jianqing Fan and Xiao Han.

distributions and the noncentral chi-square distributions, respectively, with degrees of freedom depending on whether the degrees are corrected or not. We also address the important issue of estimating the unknown number of communities and establish the asymptotic properties of the associated test statistics. The advantages and practical utility of our new procedures in terms of both size and power are demonstrated through several simulation examples and real network applications.

Running title: SIMPLE

Key words: Network p-values; Statistical inference; Large networks; Clustering; Big data; Random matrix theory; Eigenvectors; Eigenvalues

1 Introduction

Large-scale network data that describes the pairwise relational information among objects is commonly encountered in many applications such as the studies of citation networks, protein-protein interaction networks, health networks, financial networks, trade networks, and social networks. The popularity of such applications has motivated a spectrum of research with network data. Popularly used methods include algorithmic ones and model-based ones, where the former uses algorithms to optimize some carefully designed criteria (e.g., Newman (2013a,b); Zhang and Moore (2014)), and the latter relies on specific structures of some probabilistic models (see, e.g., Goldenberg et al. (2010) for a review). This paper belongs to the latter group. In the literature, a number of probabilistic models have been proposed for modeling network data. As arguably the simplest model with planted community identity, the stochastic block model (SBM) (Holland et al., 1983; Wang and Wong, 1987; Abbe, 2017) has received a tremendous amount of attention in the last decade. To overcome the limitation and increase the flexibility in the basic stochastic block model, various variants have been proposed. To name a few, the degree-corrected SBM (Karrer and Newman, 2011) introduces a degree parameter for each node to make the expected degrees match the observed ones. The overlapping SBM, such as the mixed membership model (Airoldi et al., 2008), allows the communities to overlap by assigning each node a profile of community memberships. See also Newman and Peixoto (2015) for a review of network models.

An important problem in network analysis is to unveil the true latent links between different pairs of nodes, where nodes can be broadly defined such as individuals, economic entities, documents, or medical disorders in social, economic, text, or health networks. There is a growing literature on network analysis with various methods available for clustering the nodes into different communities within which nodes are more densely connected, based on the observed adjacency matrices or the similarity matrices constructed using the node information. These methods focus mainly on the clustering aspect of the problem, outputting subgroups with predicted membership identities. Yet the statistical inference aspect such as quantifying the statistical uncertainty associated with the identification of latent links

has been largely overlooked. This paper aims at filling this crucial gap by proposing new statistical tests for testing whether any given pair of nodes share the same membership profiles, and providing the associated p-values.

Knowing the statistical significance of membership profiles can bring more confidence to practitioners in decision making. Taking the stock market for example, investors often want to form diversified portfolios by including stocks with little or no correlation in their returns. The correlation matrix of stock returns can then be used to construct an affinity matrix, and stocks with relatively highly correlated returns can be regarded as in the same community. Obtaining the pairwise p-values of stocks can help investors form diversified portfolios with statistical confidence. For instance, if one is interested in the Apple stock, then the pairwise p-values of Apple and all other candidate stocks can be calculated, and stocks with the smallest p-values can be included to form portfolios. Another important application is in legislation. For example, illegal logging greatly conflicts with indigenous and local populations, contributing to violence, human rights abuses, and corruption. The DNA sequencing technology has been used to identify the region of logs. In such application, an affinity matrix can be calculated according to the similarity of DNA sequences. Then applying our method, p-values can be calculated and used in court as statistical evidence in convicting illegal logging.

To make the problem concrete, we consider the family of degree-corrected mixed membership models, which includes the mixed membership model and the stochastic block model as special cases. In the degree-corrected mixed membership model, node i is assumed to have a membership profile characterized by a community membership probability vector $\boldsymbol{\pi}_i \in \mathbb{R}^K$, where K is the number of communities and the kth entry of $\boldsymbol{\pi}_i$ specifies the mixture proportion of node i in community k (Airoldi et al., 2008; Zhao et al., 2012). For example, a book can be 30% liberal and 70% conservative. In addition, each node is allowed to have its own degree. For any given pair of nodes i and j, we investigate whether they have the same membership profile or not by testing the hypothesis $H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j$ vs. $H_a : \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j$. Two forms of statistical inference on membership profiles in large networks (SIMPLE) test are proposed. Under the mixed membership model where all nodes have the same degree, we construct the first form of SIMPLE test by resorting to the ith and jth rows of the spiked eigenvector matrix of the observed adjacency matrix. We establish the asymptotic null and alternative distributions of the test statistic, where under the null hypothesis the asymptotic distribution is chi-square with K degrees of freedom and under the alternative hypothesis, the asymptotic distribution is noncentral chi square with a location parameter determined by how distinct the membership profiles of nodes i and j are.

In the more general degree-corrected mixed membership model, where nodes are allowed to have heterogeneous degrees, we build the second form of SIMPLE test based on the ratio statistic proposed in Jin (2015). We show that the asymptotic null distribution is chi-

square with $K-1$ degrees of freedom, and under the alternative hypothesis and some mild regularity conditions, the test statistic diverges to infinity with asymptotic probability one. We prove that these asymptotic properties continue to hold even with estimated population parameters (including the number of communities K) provided that these parameters can be estimated reasonably well. We then suggest specific estimators of these unknown parameters and show that they achieve the desired estimation precision. These new theoretical results enable us to construct rejection regions that are pivotal to the unknown parameters for each of these two forms of the SIMPLE test, and to calculate p-values explicitly. Our method is more applicable than most existing ones in the community detection literature where K is required to be known. Although the second form of SIMPLE test can be applied to both cases with and without degree heterogeneity, we would like to point out that the first test is empirically more stable since it does not involve any ratio calculations. To the best of our knowledge, this paper is the first in the literature to provide quantified uncertainty levels in community membership estimation and inference.

Our test is most useful when one cares about local information of the network. For instance, if the interest is whether two (or several) nodes belong to the same community with quantified significance level, then SIMPLE can be used. Indeed, our statistics do not rely on any pre-determined membership information. Compared to community detection methods, our work has at least three advantages: 1) we do not need to assign memberships to nodes that are not of interests; 2) our method can provide the level of significance, which can be very important in scientific discoveries; and 3) if partial membership information is known in a network, then the nodes with missing membership information can be recovered with statistical confidence by applying our tests.

Both forms of SIMPLE test are constructed using the spectral information of the observed adjacency matrix. In this sense, our work is related to the class of spectral clustering methods, which is one of the most scalable tools for community detection and has been popularly used in the literature. See, e.g., von Luxburg (2007) for a tutorial of spectral clustering methods. See also Rohe et al. (2011); Lei and Rinaldo (2015); Jin (2015) among many others for the specifics on the implementation of spectral methods for community detection. In addition, the optimality for the case of two communities has been established by Abbe et al. (2017). Our work is related to but substantially different from the link prediction problem (Liben-Nowell and Kleinberg, 2007; Wu et al., 2018), which can be thought of as predicting pairs of nodes as linked or non-linked. The major difference is that in link prediction, only part of the adjacency matrix is observed and one tries to predict the latent links among the nodes which are unobserved. Moreover, link prediction methods usually do not provide statistical confidence levels.

Our work falls into the category of hypothesis testing with network data. In the literature, hypothesis testing has been used for different purposes. For example, Arias-Castro and

Verzelen (2014) and Verzelen and Arias-Castro (2015) formalized the problem of community detection in a given random graph as a hypothesis testing problem in dense and sparse random networks, respectively. Under the stochastic block model assumption, Bickel and Sarkar (2016) proposed a recursive bipartitioning algorithm to automatically estimate the number of communities using hypothesis test constructed from the largest principal eigenvalue of the suitably centered and scaled adjacency matrix. The null hypothesis of their test is that the network has only $K = 1$ community. Lei (2016) generalized their idea and proposed a test allowing for $K \geq 1$ communities in the stochastic block model under the null hypothesis. The number of communities can then be estimated by sequential testing. Wang and Bickel (2017) proposed a likelihood ratio test for selecting the correct K under the setting of SBM.

The rest of the paper is organized as follows. Section 2 introduces the model setting and technical preparation. We present the SIMPLE method and its asymptotic theory as well as the implementation details of SIMPLE in Section 3. Sections 4 and 5 provide several simulation and real data examples illustrating the finite-sample performance and utility of our newly suggested method. We discuss some implications and extensions of our work in Section 6. All the proofs and technical details are provided in the Supplementary Material.

2 Statistical inference in large networks

2.1 Model setting

Consider an undirected graph $\mathcal{N} = (V, E)$ with n nodes, where $V = \{1, \cdots, n\}$ is the set of nodes and E is the set of links. Throughout the paper, we use the notation $[n] = \{1, \cdots, n\}$. Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times n}$ be the symmetric adjacency matrix representing the connectivity structure of graph $\mathcal{N}$, where $x_{ij} = 1$ if there is a link connecting nodes i and j, and $x_{ij} = 0$ otherwise. We consider the general case when graph $\mathcal{N}$ may or may not admit self loops, where in the latter scenario $x_{ii} = 0$ for all $i \in [n]$. Under a probabilistic model, we will assume that x_{ij} is an independent realization from a Bernoulli random variable for all upper triangular entries of random matrix $\mathbf{X}$.

To model the connectivity pattern of graph $\mathcal{N}$, consider a symmetric binary random matrix $\mathbf{X}^*$ with the following latent structure

$$\mathbf{X}^* = \mathbf{H} + \mathbf{W}^*, \tag{1}$$

where $\mathbf{H} = (h_{ij}) \in \mathbb{R}^{n \times n}$ is the deterministic mean matrix (or probability matrix) of low rank $K \geq 1$ (see (5) later for a specification) and $\mathbf{W}^* = (w^*_{ij}) \in \mathbb{R}^{n \times n}$ is a symmetric random matrix with mean zero and independent entries on and above the diagonal. Assume that the observed adjacency matrix $\mathbf{X}$ is either $\mathbf{X}^*$ or $\mathbf{X}^* - \text{diag}(\mathbf{X}^*)$, corresponding to the cases

with or without self loops, respectively. In either case, we have the following decomposition

$$\mathbf{X} = \mathbf{H} + \mathbf{W}, \tag{2}$$

where $\mathbf{W} = \mathbf{W}^*$ in the presence of self loops and $\mathbf{W} = \mathbf{W}^* - \text{diag}(\mathbf{X}^*)$ in the absence of self loops. We can see that in either case, $\mathbf{W}$ in (2) is symmetric with independent entries on and above the diagonal. Our study will cover both cases. Hereafter to simplify the presentation, we will slightly abuse the notation by referring to $\mathbf{H}$ as the mean matrix and $\mathbf{W}$ as the noise matrix.

Assume that there is an underlying latent community structure that the network $\mathcal{N}$ can be decomposed into K latent disjoint communities

$$\mathcal{C}_1, \cdots, \mathcal{C}_K,$$

where each node i is associated with the community membership probability vector $\boldsymbol{\pi}_i = (\boldsymbol{\pi}_i(1), \cdots, \boldsymbol{\pi}_i(K))^T \in \mathbb{R}^K$ such that

$$P(\text{node } i \text{ belongs to community } \mathcal{C}_k) = \boldsymbol{\pi}_i(k), \quad k = 1, \cdots, K. \tag{3}$$

Throughout the paper, we assume that the number of communities K is *unknown* but bounded away from infinity.

For any given pair of nodes $i, j \in V$ with $i \neq j$, our goal is to infer whether they share the same community identity or not with quantified uncertainty level from the observed adjacency matrix $\mathbf{X}$ in the general model (2). In other words, for each pair of nodes $i, j \in V$ with $i \neq j$, we are interested in testing the hypothesis

$$H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j \quad \text{versus} \quad H_a : \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j. \tag{4}$$

Throughout the paper, we consider the preselected pair (i, j) and thus nodes i and j are fixed.

To make the problem more explicit, we consider the degree-corrected mixed membership (DCMM) model. Using the same formulation as in Jin et al. (2017), the probability of a link between nodes i and j with $i \neq j$ under the DCMM model can be written as

$$P(x_{ij} = 1) = \theta_i \theta_j \sum_{k=1}^{K} \sum_{l=1}^{K} \boldsymbol{\pi}_i(k) \boldsymbol{\pi}_j(l) p_{kl}. \tag{5}$$

Here, $\theta_i > 0$, $i \in [n]$, measures the degree heterogeneity, and p_{kl} can be interpreted as the probability of a typical member ($\theta_i = 1$, say) in community $\mathcal{C}_k$ connects with a typical member ($\theta_j = 1$, say) in community $\mathcal{C}_l$, as in the stochastic block model. Writing (5) in the matrix form, we have

$$\mathbf{H} = \mathbf{\Theta\Pi P\Pi}^T\mathbf{\Theta}, \tag{6}$$

where $\mathbf{\Theta} = \text{diag}(\theta_1, \cdots, \theta_n)$ stands for the degree heterogeneity matrix, $\mathbf{\Pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_n)^T \in \mathbb{R}^{n\times K}$ is the matrix of community membership probability vectors, and $\mathbf{P} = (p_{kl}) \in \mathbb{R}^{K\times K}$ is a nonsingular matrix with $p_{kl} \in [0, 1]$, $1 \le k, l \le K$.

The family of DCMM models in (6) contains several popularly used network models for community detection as special cases. For example, when $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$ and $\boldsymbol{\pi}_i \in \{\mathbf{e}_1, \cdots, \mathbf{e}_K\}$ with $\mathbf{e}_k$ a unit vector whose kth component is one and all other components are zero, the model reduces to the stochastic block model with non-overlapping communities. When $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$ and $\boldsymbol{\pi}_i$'s are general community membership probability vectors, the model becomes the mixed membership model. Each of these models has been studied extensively in the literature. Yet almost all these existing works have focused on the community detection perspective, which is a statistical estimation problem. In this paper, however we will concentrate on the statistical inference problem (4).

2.2 Technical preparation

When $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$, we have $\mathbf{H} = \theta\mathbf{\Pi P\Pi}^T$. Thus the column space spanned by $\mathbf{\Pi}$ is the same as the eigenspace spanned by the top K eigenvectors of matrix $\mathbf{H}$. In other words, the membership profiles of the network are encoded in the eigen-structure of the mean matrix $\mathbf{H}$. Denote by $\mathbf{H} = \mathbf{VDV}^T$ the eigen-decomposition of the mean matrix, where $\mathbf{D} = \text{diag}(d_1, \cdots, d_K)$ with $|d_1| \ge |d_2| \ge \cdots \ge |d_K| > 0$ is the matrix of all K nonzero eigenvalues and $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_K) \in \mathbb{R}^{n\times K}$ is the corresponding orthonormal matrix of eigenvectors. In practice, one replaces the matrices $\mathbf{D}$ and $\mathbf{V}$ by those of the observed adjacency matrix $\mathbf{X}$. Denote by $\widehat{d}_1, \cdots, \widehat{d}_n$ the eigenvalues of matrix $\mathbf{X}$ and $\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_n$ the corresponding eigenvectors. Without loss of generality, assume that $|\widehat{d}_1| \ge |\widehat{d}_2| \ge \cdots \ge |\widehat{d}_n|$ and let $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_K) \in \mathbb{R}^{n\times K}$. Denote by $\mathbf{W} = (w_{ij})$ and define $\alpha_n = \{\max_{1\le j\le n}\sum_{i=1}^n \text{var}(w_{ij})\}^{1/2}$, which is simply the maximum standard deviation of the column sums (node degrees).

The asymptotic mean of the empirical eigenvalue $\widehat{d}_k$ for $k \in [K]$ has been derived in Fan et al. (2020), which is a population quantity t_k and will be used frequently in our paper. Its definition is somewhat complicated which we now describe as follows. Let a_k and b_k be defined as

$$a_k = \begin{cases} \frac{d_k}{1+c_0/2} & \text{if } d_k > 0 \\ (1+c_0/2)d_k & \text{if } d_k < 0 \end{cases}, \quad b_k = \begin{cases} (1+c_0/2)d_k & \text{if } d_k > 0 \\ \frac{d_k}{1+c_0/2} & \text{if } d_k < 0 \end{cases},$$

where the eigen-ratio gap constant $c_0 > 0$ is given in Condition 1 in Section 3.1. For any deterministic real-valued matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of appropriate dimensions and complex

number $z \neq 0$, define

$$\mathcal{R}(\mathbf{M}_1, \mathbf{M}_2, z) = -\frac{1}{z}\mathbf{M}_1^T\mathbf{M}_2 - \sum_{l=2}^{L} \frac{1}{z^{l+1}}\mathbf{M}_1^T\mathbb{E}\mathbf{W}^l\mathbf{M}_2 \tag{7}$$

with L the smallest positive integer such that uniformly over $k \in [K]$,

$$\left(\frac{\alpha_n}{|z|}\right)^L \leq \min\left\{\frac{1}{n^4}, \frac{1}{|z|^4}\right\}, \quad z \in [a_k, b_k], \tag{8}$$

where $|z|$ denotes the modulus of complex number z. We can see that as long as $\frac{|d_K|}{\alpha_n} \geq n^\epsilon$ with some positive constant ϵ, which is guaranteed by Condition 1 and Condition 2 (or 4) in Section 3.1 (or Section 3.2), the existence of the desired positive integer L can be ensured.

We are now ready to define the asymptotic mean t_k of the sample eigenvalue $\widehat{d}_k$. For each $k \in [K]$, define t_k as the solution to equation

$$1 + d_k\left\{\mathcal{R}(\mathbf{v}_k, \mathbf{v}_k, z) - \mathcal{R}(\mathbf{v}_k, \mathbf{V}_{-k}, z)\left[\mathbf{D}_{-k}^{-1} + \mathcal{R}(\mathbf{V}_{-k}, \mathbf{V}_{-k}, z)\right]^{-1}\mathcal{R}(\mathbf{V}_{-k}, \mathbf{v}_k, z)\right\} = 0 \tag{9}$$

when restricted to the interval $z \in [a_k, b_k]$, where $\mathbf{V}_{-k}$ is the submatrix of $\mathbf{V}$ formed by removing the kth column and $\mathbf{D}_{-k}$ is formed by removing the kth diagonal entry of $\mathbf{D}$. Then as shown in Fan et al. (2020), for each $k \in [K]$, t_k is the asymptotic mean of the sample eigenvalue $\widehat{d}_k$ and $t_k/d_k \to 1$ as $n \to \infty$. See also Lemma 15 in Section C.6 of Supplementary Material, where the existence, uniqueness, and asymptotic property of t_k's are stated.

To facilitate the technical presentation, we further introduce some notation that will be used throughout the paper. We use $a \ll b$ to represent $a/b \to 0$. For a matrix $\mathbf{A} = (\mathbf{A}_{ij})$, denote by $\lambda_j(\mathbf{A})$ the jth largest eigenvalue, and $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}\mathbf{A}^T)}$, $\|\mathbf{A}\|_2$, and $\|\mathbf{A}\|_\infty = \max_{i,j}|\mathbf{A}_{ij}|$ the Frobenius norm, the spectral norm, and the entrywise maximum norm, respectively. In addition, we use $\mathbf{A}(k)$ to denote the kth row of a matrix $\mathbf{A}$, and $\mathbf{a}(k)$ to denote the kth component of a vector $\mathbf{a}$. For a unit vector $\mathbf{x} = (x_1, \cdots, x_n)^T$, let $d_{\mathbf{x}} = \max_{1\leq i\leq n}|x_i|$. Also define $\theta_{\max} = \max_{1\leq i\leq n}\theta_i$ and $\theta_{\min} = \min_{1\leq i\leq n}\theta_i$ as the maximum and minimum node degrees, respectively. For each $1 \leq k \leq K$, denote by $\mathcal{N}_k = \{i : 1 \leq i \leq n, \boldsymbol{\pi}_i(k) = 1\}$ the set of pure nodes in community k, where each pure node belongs to only a single community. Some additional definitions and notation are given at the beginning of Section A.

3 SIMPLE and its asymptotic theory

3.1 SIMPLE for mixed membership models

We first consider the hypothesis testing problem (4) in the mixed membership model without degree heterogeneity whose mean matrix takes the form (6) with $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$, that is,

$$\mathbb{E}\mathbf{X} = \mathbf{H} = \theta\mathbf{\Pi}\mathbf{P}\mathbf{\Pi}^T. \tag{10}$$

Here θ is allowed to converge to zero as $n \to \infty$. This model is a simple version of the mixed membership stochastic block (MMSB) model considered in Airoldi et al. (2008). As mentioned before, this model includes the stochastic block model with non-overlapping communities as a special case.

Under model (10), if $\boldsymbol{\pi}_i = \boldsymbol{\pi}_j$ then nodes i and j are exchangeable and it holds that $\mathbf{V}(i) = \mathbf{V}(j)$ by a simple permutation argument (see the beginning of the proof of Theorem 1 in Section A.1). Motivated by this observation, we consider the following test statistic for assessing the membership information of the ith and jth nodes

$$T_{ij} = \left[\widehat{\mathbf{V}}(i) - \widehat{\mathbf{V}}(j)\right]^T \mathbf{\Sigma}_1^{-1} \left[\widehat{\mathbf{V}}(i) - \widehat{\mathbf{V}}(j)\right], \tag{11}$$

where $\mathbf{\Sigma}_1$ is the asymptotic variance of $\widehat{\mathbf{V}}(i) - \widehat{\mathbf{V}}(j)$ that is challenging to derive and estimate. Nevertheless, we will show that $\mathbf{\Sigma}_1 = \mathrm{cov}[(\mathbf{e}_i - \mathbf{e}_j)^T\mathbf{W}\mathbf{V}\mathbf{D}^{-1}]$ whose expression is given in (28) later, and provide an estimator with required accuracy.

We need the following regularity conditions in establishing the asymptotic null and alternative distributions of test statistic T_{ij}.

Condition 1. *There exists some positive constant c_0 such that*

$$\min\{\frac{|d_i|}{|d_j|} \ : 1 \le i < j \le K, d_i \ne -d_j\} \ge 1 + c_0.$$

In addition, $\alpha_n \to \infty$ as $n \to \infty$.

Condition 2. *There exist some constants $0 < c_0 < 1$, $0 \le c_2 < 1/2$, $0 < c_1 < 1 - 2c_2$ such that $\lambda_K(\mathbf{\Pi}^T\mathbf{\Pi}) \ge c_0 n$, $\lambda_K(\mathbf{P}) \ge n^{-c_2}$, and $\theta \ge n^{-c_1}$.*

Condition 3. *As $n \to \infty$, all the eigenvalues of $\theta^{-1}\mathbf{D}\mathbf{\Sigma}_1\mathbf{D}$ are bounded away from 0 and ∞.*

The constant c_0 in Condition 1 can be replaced with some $o(1)$ term that vanishes as n grows at the cost of significantly more tedious calculations in our technical analysis. This condition is imposed to exclude the complicated case of multiplicity, which can lead to the singularity of $\mathbf{\Sigma}_1$, making our test ill-defined. A potential remedy is to use the Moore–Penrose generalized inverse of matrix $\mathbf{\Sigma}_1$ in defining T_{ij}, which we will leave to the future study due

to the extra technical challenge. We acknowledge that in some special models, results on community detection have been established allowing multiplicity (e.g., Gao et al. (2018)). Conditions 2 is a standard regularity assumption imposed for the case of mixed membership models. In particular, θ measures the degree density and is allowed to converge to zero at the polynomial rate n^{-c_1} with constant c_1 arbitrarily close to one. Condition 3 is a technical condition for establishing the asymptotic properties of T_{ij}. We provide sufficient conditions for ensuring Condition 3 in Section D of Supplementary file. As shown in the proof of Theorem 1, under Conditions 1 and 2, we have $\text{var}[(\mathbf{e}_i - \mathbf{e}_j)^T \mathbf{W} \mathbf{v}_k] \sim \theta$ for all $k = 1, \cdots, K$, which explains the normalization factor θ^{-1} in Condition 3. Our conditions accommodate the case where the magnitudes of spiked eigenvalues $|d_1|, \cdots, |d_K|$ are of different orders.

Example 1. *Consider SBM with $K = 2$ communities of equal sizes $n_1 = n_2 = n/2$ and $n^{-c_1} \leq \theta < 1$. Further assume that $\mathbf{P}$ has diagonal entries equal to a and off-diagonal entries equal to b, with a and b some positive constants satisfying $a > b$. Then we have $d_1 = n(a+b)\theta$ and $d_2 = n(a-b)\theta$. Some direct calculations show that Conditions 1–3 all hold.*

The following theorem summarizes the asymptotic distribution of test statistic T_{ij} under the null and alternative hypotheses.

Theorem 1. *Assume that Conditions 1–2 hold under the mixed membership model (10).*

i) Under the null hypothesis $H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j$, if in addition Condition 3 holds, then we have

$$T_{ij} \xrightarrow{\mathscr{D}} \chi^2_K \tag{12}$$

as $n \to \infty$, where χ^2_K is the chi-square distribution with K degrees of freedom.

ii) Under the contiguous alternative hypothesis $H_a : \boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j$ but $n^{1/2-c_2}\sqrt{\theta}\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| \to \infty$, then for arbitrarily large constant $C > 0$, we have

$$P(T_{ij} > C) \to 1 \tag{13}$$

as $n \to \infty$. Moreover, if Condition 3 holds, $c_2 = 0$, $\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| \sim \frac{1}{\sqrt{n\theta}}$, and $[\mathbf{V}(i) - \mathbf{V}(j)]^T \boldsymbol{\Sigma}_1^{-1} [\mathbf{V}(i) - \mathbf{V}(j)] \to \mu$ with μ some constant, then it holds that

$$T_{ij} \xrightarrow{\mathscr{D}} \chi^2_K(\mu) \tag{14}$$

as $n \to \infty$, where $\chi^2_K(\mu)$ is a noncentral chi-square distribution with mean μ and K degrees of freedom.

Remark 1. *Under the joint null hypotheses $H_{0,ij} : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j$ for all $1 \leq i \neq j \leq n$, we have*

in fact proved a uniform version of the result in (12)*:*

$$\lim_{n\to\infty} \sup_{1\le i\neq j\le n} |P(T_{ij} \le x) - P(X \le x)| = 0 \quad \textit{for all } x \in \mathbb{R}, \tag{15}$$

where $X \sim \chi^2_K$*. See Section E of Supplementary Material for more details.*

In the special case of stochastic block model with non-overlapping communities, we can see that $\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| = 0$ under the null hypothesis H_0, and $\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| = \sqrt{2}$ under the alternative hypothesis H_a. Thus under the null hypothesis H_0 and Conditions 1–3, the test statistic T_{ij} has asymptotic distribution (12). Under the alternative hypotheses H_a and Conditions 1–2, we have $\sqrt{n\theta}\|\boldsymbol{\pi}_i - \boldsymbol{\pi}_j\| \to \infty$ and thus the limiting result (13) holds.

The test statistic T_{ij} is, however, not directly applicable because of the unknown population parameters K and $\boldsymbol{\Sigma}_1$. We next show that for consistent estimators satisfying the following conditions

$$P(\widehat{K} = K) = 1 - o(1), \tag{16}$$

$$\theta^{-1}\|\mathbf{D}(\widehat{\mathbf{S}}_1 - \boldsymbol{\Sigma}_1)\mathbf{D}\|_2 = o_p(1), \tag{17}$$

the asymptotic results in Theorem 1 continue to hold.

Theorem 2. *Assume that estimators* $\widehat{K}$ *and* $\widehat{\mathbf{S}}_1$ *satisfy* (16) *and* (17)*, respectively. Let* $\widehat{T}_{ij}$ *be the test statistic constructed by replacing* K *and* $\boldsymbol{\Sigma}_1$ *in* (11) *with* $\widehat{K}$ *and* $\widehat{\mathbf{S}}_1$*, respectively. Then Theorem 1 holds with* T_{ij} *replaced by* $\widehat{T}_{ij}$ *under the same conditions.*

Theorem 2 suggests that at significance level α, to test the null hypothesis H_0 in (4), we can construct the following rejection region

$$\{\widehat{T}_{ij} > \chi^2_{\widehat{K},1-\alpha}\}, \tag{18}$$

where $\chi^2_{\widehat{K},1-\alpha}$ is the $100(1-\alpha)$th percentile of the chi-square distribution with $\widehat{K}$ degrees of freedom. The following corollary justifies the asymptotic size and power of our test.

Corollary 1. *Assume that* $\widehat{K}$ *and* $\widehat{\mathbf{S}}_1$ *satisfy* (16) *and* (17)*, respectively. Under the same conditions for ensuring* (12)*, event* (18) *holds with asymptotic probability* α*. Under the same conditions for ensuring* (13)*, event* (18) *holds with asymptotic probability one.*

3.2 SIMPLE for degree-corrected mixed membership models

In this section, we further consider the hypothesis testing problem (4) in the more general DCMM model (6). Degree heterogeneity in network models has been explored in the statistics literature. To name a few, Jin et al. (2017) considered the estimation of node membership assuming the average degree of the nodes to be much larger than $\log n$. Jin and Ke (2017)

established a sharp lower bound for the estimated node membership allowing the average node degree to diverge with the order $\log^2 n$ or faster. Zhang et al. (2020) proposed a spectral-based detection algorithm to recover the node membership assuming that $\theta_{\max}/\theta_{\min}$ is bounded by some positive constant. Our assumption on the degree heterogeneity is similar to that in Zhang et al. (2020) and will be presented in Condition 4 below.

The test statistic T_{ij} defined in Section 3.1 is no longer applicable due to the degree heterogeneity. A simple algebra shows that degree heterogeneity can be eliminated by the ratios of eigenvectors (columnwise division). Thus, following Jin (2015), to correct the degree heterogeneity we define the following componentwise ratio

$$Y(i,k) = \frac{\widehat{\mathbf{v}}_k(i)}{\widehat{\mathbf{v}}_1(i)}, \quad 1 \le i \le n,\ 2 \le k \le K, \tag{19}$$

where $0/0$ is defined as 1 by convention. Note that the division here is to get rid of the degree heterogeneity and the equality

$$\frac{\mathbf{v}_k(i)}{\mathbf{v}_1(i)} = \frac{\mathbf{v}_k(j)}{\mathbf{v}_1(j)}, \qquad 2 \le k \le K \tag{20}$$

holds under the null hypothesis, which is due to the exchangeability of nodes i and j under the mixed membership model; see (A.18) at the beginning of the proof of Theorem 3 in Section A.4. Denote by $\mathbf{Y}_i = (Y(i,2), \cdots, Y(i,K))^T$. Our new test statistic will be built upon $\mathbf{Y}_i$.

To test the null hypothesis $H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j$, using (19) and (20), we propose to use the following test statistic

$$G_{ij} = (\mathbf{Y}_i - \mathbf{Y}_j)^T \boldsymbol{\Sigma}_2^{-1} (\mathbf{Y}_i - \mathbf{Y}_j) \tag{21}$$

for assessing the null hypothesis H_0 in (4), where $\boldsymbol{\Sigma}_2$ is the asymptotic variance of $\mathbf{Y}_i - \mathbf{Y}_j$. This is even much harder to derive and estimate. Nevertheless, we will show $\boldsymbol{\Sigma}_2 = \mathrm{cov}(\mathbf{f})$ with $\mathbf{f} = (f_2, \cdots, f_K)^T$ and

$$f_k = \frac{\mathbf{e}_i^T \mathbf{W} \mathbf{v}_k}{t_k \mathbf{v}_1(i)} - \frac{\mathbf{e}_j^T \mathbf{W} \mathbf{v}_k}{t_k \mathbf{v}_1(j)} - \frac{\mathbf{v}_k(i) \mathbf{e}_i^T \mathbf{W} \mathbf{v}_1}{t_1 \mathbf{v}_1^2(i)} + \frac{\mathbf{v}_k(j) \mathbf{e}_j^T \mathbf{W} \mathbf{v}_1}{t_1 \mathbf{v}_1^2(j)}. \tag{22}$$

The entries of $\boldsymbol{\Sigma}_2$ are given by (29) later that also involves the asymptotic mean of $\widehat{d}_k$.

The following conditions are needed for investigating the asymptotic properties of test statistic G_{ij}.

Condition 4. *There exist some constants $c_2 \in [0, 1/2)$, $c_3 \in (0, 1 - 2c_2)$, $c_5 \in (0,1)$ and $c_4 > 0$ such that $\lambda_K(\mathbf{P}) \ge n^{-c_2}$, $\min_{1 \le k \le K} |\mathcal{N}_k| \ge c_5 n$, $\theta_{\max} \le c_4 \theta_{\min}$, and $\theta_{\min}^2 \ge n^{-c_3}$.*

Condition 5. *Matrix $\mathbf{P} = (p_{kl})$ is positive definite, irreducible, and has unit diagonal entries. Moreover $n \min_{1 \le k \le K,\, t=i,j} \mathrm{var}(\mathbf{e}_t^T \mathbf{W} \mathbf{v}_k) \sim n\theta_{\max}^2 \to \infty$.*

Condition 6. *It holds that all the eigenvalues of* $(n\theta_{\max}^2)^{-1}\mathbf{D}\mathrm{cov}(\mathbf{f})\mathbf{D}$ *are bounded away from 0 and* ∞.

Condition 7. *Let* $\boldsymbol{\eta}_1$ *be the first right singular vector of* $\mathbf{P}\boldsymbol{\Pi}^\top\boldsymbol{\Theta}^2\boldsymbol{\Pi}$. *It holds that*

$$\min_{1\le k\le K}\boldsymbol{\eta}_1(k) > 0, \quad \textit{and} \quad \frac{\max_{1\le k\le K}\boldsymbol{\eta}_1(k)}{\min_{1\le k\le K}\boldsymbol{\eta}_1(k)} \le C,$$

for some positive constant C, *where* $\boldsymbol{\eta}_1(k)$ *is the* k*-th entry of* $\boldsymbol{\eta}_1$.

Conditions 4–7 are similar to those in Jin et al. (2017). In particular, Conditions 4, 5 and 7 are special cases of (2.13), (2.14) and (2.16) therein. Same as in the previous section, the degree density is measured by $\theta_{\min}^2$ and is allowed to converge to zero at rate n^{-c_3}, and our conditions accommodate the case where $|d_1|, \cdots, |d_K|$ are of different orders.

Theorem 3. *Assume that Conditions 1 and 4–7 hold under the degree-corrected mixed membership model (6).*

i) Under the null hypothesis $H_0 : \boldsymbol{\pi}_i = \boldsymbol{\pi}_j$, *we have as* $n \to \infty$,

$$G_{ij} \xrightarrow{\mathscr{D}} \chi^2_{K-1}. \tag{23}$$

ii) Under the contiguous alternative hypothesis with $\lambda_2(\boldsymbol{\pi}_i\boldsymbol{\pi}_i^T + \boldsymbol{\pi}_j\boldsymbol{\pi}_j^T) \gg \frac{1}{n^{1-2c_2}\theta_{\min}^2}$, *we have for any arbitrarily large constant* $C > 0$,

$$P(G_{ij} > C) \to 1 \ \textit{as}\ n \to \infty. \tag{24}$$

A uniform result similar to (15) has also been proved in Section E of Supplementary Material under the DCMM. The test statistic G_{ij} is not directly applicable in practice due to the presence of the unknown population parameters K and $\boldsymbol{\Sigma}_2$. Nevertheless, certain consistent estimators can be constructed and the results in Theorem 3 remain valid. In particular, for the estimator $\widehat{K}$ of K, we require condition (16) and for the estimator $\widehat{\mathbf{S}}_2$ of $\boldsymbol{\Sigma}_2$, we need the following property

$$(n\theta_{\max}^2)^{-1}\|\mathbf{D}(\widehat{\mathbf{S}}_2 - \boldsymbol{\Sigma}_2)\mathbf{D}\|_2 = o_p(1). \tag{25}$$

Theorem 4. *Assume that the estimators* $\widehat{K}$ *and* $\widehat{\mathbf{S}}_2$ *of parameters* K *and* $\boldsymbol{\Sigma}_2$ *satisfy (16) and (25), respectively. Let* $\widehat{G}_{ij}$ *be the test statistic constructed by replacing* K *and* $\boldsymbol{\Sigma}_2$ *with* $\widehat{K}$ *and* $\widehat{\mathbf{S}}_2$, *respectively. Then Theorem 3 holds with* G_{ij} *replaced by* $\widehat{G}_{ij}$ *under the same conditions.*

Theorem 4 suggests that with significance level α, the rejection region can be constructed as

$$\{\widehat{G}_{ij} > \chi^2_{\widehat{K}-1,1-\alpha}\}. \tag{26}$$

We have similar results to Corollary 1 regarding the type I and type II errors of the above rejection region.

Corollary 2. *Assume that* $\widehat{K}$ *and* $\widehat{\mathbf{S}}_2$ *satisfy* (16) *and* (25)*, respectively. Under the same conditions for ensuring* (23)*, event* (26) *holds with asymptotic probability* α*. Under the same conditions for ensuring* (24)*, event* (26) *holds with asymptotic probability one.*

It is worth mentioning that since the DCMM model (6) is more general than the mixed membership model (10), the test statistic $\widehat{G}_{ij}$ can be applied even under model (10). However, as will be shown in our simulation studies in Section 4, the finite-sample performance of $\widehat{T}_{ij}$ can be better than that of $\widehat{G}_{ij}$ in such a model setting, which is not surprising since the latter involves ratios (see (19)) in its definition and has two sources of variations from both numerators and denominators. This is also reflected in losing one degree of freedom in (26)

3.3 Estimation of unknown parameters

We now discuss some consistent estimators of K, $\mathbf{\Sigma}_1$, and $\mathbf{\Sigma}_2$ that satisfy conditions (16), (17), and (25), respectively. There are some existing works concerning the estimation of parameter K. For example, Lei (2016); Chen and Lei (2018); Daudin et al. (2008); Latouche et al. (2012); Saldana et al. (2017); Wang and Bickel (2017), among others. Most of these works consider specific network models such as the stochastic block model or degree-corrected stochastic block model.

In our paper, since we consider the general DCMM model (6) which allows for mixed memberships, the existing methods are no longer applicable. To overcome the difficulty, we suggest a simple thresholding estimator defined as

$$\widehat{K} = \Big|\Big\{\widehat{d}_i : \ \widehat{d}_i^2 > 2.01(\log n)\check{d}_n, i \in [n]\Big\}\Big|, \tag{27}$$

where $|\cdot|$ stands for the cardinality of a set, the constant 2.01 can be replaced with any other constant that is slightly larger than 2, and $\check{d}_n = \max_{1\le l\le n}\sum_{j=1}^n X_{lj}$ is the maximum degree of the network. That is, we count the number of eigenvalues of matrix $\mathbf{X}$ whose magnitudes exceed a certain threshold. The following lemma justifies the consistency of $\widehat{K}$ defined in (27) as an estimator of the true number of communities K.

Lemma 1. *Assume that Condition 1 holds,* $|d_K| \gg \sqrt{\log(n)}\alpha_n$ *and* $\alpha_n \ge n^{c_5}$ *for some positive constant* c_5*. Then* $\widehat{K}$ *defined in* (27) *is consistent, that is, it satisfies condition* (16)*.*

Observe in Theorems 1–4 that we need the condition of $K \ge 1$ for test statistic $\widehat{T}_{ij}$ and the condition of $K \ge 2$ for test statistic $\widehat{G}_{ij}$. Motivated by such an observation, we propose to use $\max\{\widehat{K}, 1\}$ and $\max\{\widehat{K}, 2\}$ as the estimated number of communities in implementing test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$, respectively.

We next discuss the estimation of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. The following two lemmas provide the expansions of these two matrices which serve as the foundation for our proposed estimators.

Lemma 2. *The (a,b)th entry of matrix $\mathbf{\Sigma}_1$ is given by*

$$\frac{1}{d_a d_b}\left\{\sum_{t\in\{i,j\}}\sum_{l=1}^{n}\sigma_{tl}^2\mathbf{v}_a(l)\mathbf{v}_b(l)-\sigma_{ij}^2\left[\mathbf{v}_a(j)\mathbf{v}_b(i)+\mathbf{v}_a(i)\mathbf{v}_b(j)\right]\right\}, \tag{28}$$

where $\sigma_{ab}^2=\mathrm{var}(w_{ab})$ for $1\le a,b\le n$.

Lemma 3. *The (a,b)th entry of matrix $\mathbf{\Sigma}_2$ is given by*

$$\begin{aligned}
&\frac{1}{t_1^2}\Big\{\sum_{l=1,\,l\neq j}^{n}\sigma_{il}^2\left[\frac{t_1\mathbf{v}_{a+1}(l)}{t_{a+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{a+1}(i)\mathbf{v}_1(l)}{\mathbf{v}_1(i)^2}\right]\left[\frac{t_1\mathbf{v}_{b+1}(l)}{t_{b+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{b+1}(i)\mathbf{v}_1(l)}{\mathbf{v}_1(i)^2}\right]\\
&\quad+\sum_{l=1,\,l\neq i}^{n}\sigma_{jl}^2\left[\frac{t_1\mathbf{v}_{a+1}(l)}{t_{a+1}\mathbf{v}_1(j)}-\frac{\mathbf{v}_{a+1}(j)\mathbf{v}_1(l)}{\mathbf{v}_1(j)^2}\right]\left[\frac{t_1\mathbf{v}_{b+1}(l)}{t_{b+1}\mathbf{v}_1(j)}-\frac{\mathbf{v}_{b+1}(j)\mathbf{v}_1(l)}{\mathbf{v}_1(j)^2}\right]\\
&\quad+\sigma_{ij}^2\left[\frac{t_1\mathbf{v}_{a+1}(j)}{t_{a+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{a+1}(i)\mathbf{v}_1(j)}{\mathbf{v}_1(i)^2}-\frac{t_1\mathbf{v}_{a+1}(i)}{t_{a+1}\mathbf{v}_1(j)}+\frac{\mathbf{v}_{a+1}(j)\mathbf{v}_1(i)}{\mathbf{v}_1(j)^2}\right]\\
&\quad\times\left[\frac{t_1\mathbf{v}_{b+1}(j)}{t_{b+1}\mathbf{v}_1(i)}-\frac{\mathbf{v}_{b+1}(i)\mathbf{v}_1(j)}{\mathbf{v}_1(i)^2}-\frac{t_1\mathbf{v}_{b+1}(i)}{t_{b+1}\mathbf{v}_1(j)}+\frac{\mathbf{v}_{b+1}(j)\mathbf{v}_1(i)}{\mathbf{v}_1(j)^2}\right]\Big\}.
\end{aligned} \tag{29}$$

The above expansions in Lemmas 2–3 suggest that the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ can be estimated by plugging in the sample estimates to replace the unknown population parameters. In particular, $\mathbf{v}_a$ and d_a can be estimated by $\widehat{\mathbf{v}}_a$ and $\widehat{d}_a$, respectively, and the last result in Lemma 15 suggests that t_k can be estimated by $\widehat{d}_k$ very well. The estimation of σ_{ab}^2 is more complicated and we will discuss it in more details below.

Recall that $\sigma_{ab}^2=\mathrm{var}(w_{ab})$. With estimated $\widehat{K}$, a naive estimator of σ_{ab}^2 is $\widehat{w}_{0,ab}^2$ with $\widehat{\mathbf{W}}_0=(\widehat{w}_{0,ab})=\mathbf{X}-\sum_{k=1}^{\widehat{K}}\widehat{d}_k\widehat{\mathbf{v}}_k\widehat{\mathbf{v}}_k^T$. The good news is that it appears in (28) and (29) in the form of the average and hence the variance will be averaged out. However, this estimator is not good enough to make (17) and (25) hold due to the well-known fact that $\hat{d}_k$ is biased up. Thus we propose the following one-step refinement procedure to estimate σ_{ab}^2, which is motivated from the higher-order asymptotic expansion of empirical eigenvalue $\widehat{d}_k$ in our theoretical analysis and shrinks $\widehat{d}_k$ to make the bias at a more reasonable level.

1). Calculate the initial estimator $\widehat{\mathbf{W}}_0=\mathbf{X}-\sum_{k=1}^{\widehat{K}}\widehat{d}_k\widehat{\mathbf{v}}_k\widehat{\mathbf{v}}_k^T$.

2). With the initial estimator $\widehat{\mathbf{W}}_0$, update the estimator of eigenvalue d_k as

$$\widetilde{d}_k=\Big[\frac{1}{\widehat{d}_k}+\frac{\widehat{\mathbf{v}}_k^T\mathrm{diag}(\widehat{\mathbf{W}}_0^2)\widehat{\mathbf{v}}_k}{\widehat{d}_k^3}\Big]^{-1}.$$

3). Then update the estimator of $\mathbf{W}$ as $\widehat{\mathbf{W}}\equiv(\widehat{w}_{ij})=\mathbf{X}-\sum_{k=1}^{\widehat{K}}\widetilde{d}_k\widehat{\mathbf{v}}_k\widehat{\mathbf{v}}_k^T$ and estimate σ_{ab}^2 as $\widehat{\sigma}_{ab}^2=\widehat{w}_{ab}^2$.

To summarize, we propose to estimate matrix $\mathbf{\Sigma}_1$ by replacing d_k, $\mathbf{v}_k$, and σ_{ab}^2 with $\widehat{d}_k$, $\widehat{\mathbf{v}}_k$, and $\widehat{\sigma}_{ab}^2$, respectively, in (28). The covariance matrix $\mathbf{\Sigma}_2$ can be estimated in a similar

way by replacing t_k, $\mathbf{v}_k$, and σ_{ab}^2 with $\widehat{d}_k$, $\widehat{\mathbf{v}}_k$, and $\widehat{\sigma}_{ab}^2$, respectively, in (29). Denote by $\widehat{\mathbf{S}}_1$ and $\widehat{\mathbf{S}}_2$ the resulting estimators, respectively. The following lemma justifies the effectiveness of these two estimators.

Theorem 5. *Under Conditions 1–3, estimator* $\widehat{\mathbf{S}}_1$ *satisfies condition* (17). *Under Conditions 1 and 4–7, estimator* $\widehat{\mathbf{S}}_2$ *satisfies condition* (25).

4 Simulation studies

We use simulation examples to examine the finite-sample performance of our new SIMPLE test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ with true and estimated numbers of communities K, respectively. In particular, we consider the following two model settings.

Model 1: the mixed membership model (10). We consider $K = 3$ communities, where there are n_0 pure nodes within each community. Thus for the kth community, the community membership probability vector for each pure node is $\boldsymbol{\pi} = \mathbf{e}_k \in \mathbb{R}^K$. The remaining $n - 3n_0$ nodes are divided equally into 4 groups, where within the lth group all nodes have mixed memberships with community membership probability vector $\mathbf{a}_l$, $l = 1, \cdots, 4$. We set $\mathbf{a}_1 = (0.2, 0.6, 0.2)^T$, $\mathbf{a}_2 = (0.6, 0.2, 0.2)^T$, $\mathbf{a}_3 = (0.2, 0.2, 0.6)^T$, and $\mathbf{a}_4 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^T$. Matrix $\mathbf{P}$ has diagonal entries one and (i, j)th entry equal to $\frac{\rho}{|i-j|}$ for $i \neq j$. We experiment with two sets of parameters $(\rho, n, n_0) = (0.2, 3000, 500)$ and $(0.2, 1500, 300)$, and vary the value of θ from 0.2 to 0.9 with step size 0.1. It is clear that parameter θ has direct impact on the average degree and hence measures the signal strength.

Model 2: the DCMM model (6). Both matrices $\mathbf{\Pi}$ and $\mathbf{P}$ are the same as in Model 1. For the degree heterogeneity matrix $\mathbf{\Theta} = \mathrm{diag}(\theta_1, \cdots, \theta_n)$, we simulate $\frac{1}{\theta_i}$ as independent and identically distributed (i.i.d.) random variables from the uniform distribution on $[\frac{1}{r}, \frac{2}{r}]$ with $r \in (0, 1]$. We consider different choices of r with $r^2 \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We can see that as parameter r^2 increases, the signal becomes stronger.

4.1 Hypothesis testing with K known

Recall that our test statistics are designed to test the membership information for each preselected pair of nodes (i, j) with $1 \leq i \neq j \leq n$. To examine the empirical size of our tests, we preselect (i, j) as two nodes with community membership probability vector $(0.2, 0.6, 0.2)^T$. To examine the empirical power of our tests, we preselect i as a node with community membership probability vector $(0.2, 0.6, 0.2)^T$ and j as a node with community membership probability vector $(0, 1, 0)^T$. The nominal significance level is set to be 0.05 when calculating the critical points and the number of repetitions is chosen as 500.

We first generate simulated data from Model 1 introduced above and examine the empirical size and power of test statistic $\widehat{T}_{ij}$ with estimated $\mathbf{\Sigma}_1$, but with the true value of K. Then we consider Model 2 and examine the empirical size and power of test statistic $\widehat{G}_{ij}$

with estimated $\mathbf{\Sigma}_2$ and the true value of K. The empirical size and power at different signal levels are reported in Tables 1 and 2, corresponding to sample sizes $n = 1500$ and 3000, respectively. As shown in Tables 1 and 2, the size and power of our tests converge quickly to the nominal significance level 0.05 and the value of one, respectively, as the signal strength θ (related to effective sample size) increases. As demonstrated in Figure 1, the empirical null distributions are well described by our theoretical results. These results provide stark empirical evidence supporting our theoretical findings, albeit complicated formulas (28) and (29).

Table 1: The size and power of test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ when the true value of K is used. The nominal level is 0.05 and sample size is $n = 1500$.

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.058	0.046	0.06	0.05	0.05	0.058	0.036	0.05
	Power	0.734	0.936	0.986	0.998	1	1	1	1
Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.076	0.062	0.072	0.062	0.074	0.046	0.044	0.056
	Power	0.426	0.562	0.696	0.77	0.89	0.93	0.952	0.976

Table 2: The size and power of test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ when the true value of K is used. The nominal level is 0.05 and sample size is $n = 3000$.

Model 1	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.082	0.066	0.052	0.052	0.044	0.042	0.038	0.062
	Power	0.936	0.994	1	1	1	1	1	1
Model 2	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	Size	0.082	0.06	0.062	0.058	0.062	0.066	0.064	0.06
	Power	0.67	0.842	0.918	0.972	0.99	1	1	1

Figure 1 presents how the asymptotic null distributions change with sample size n when $\theta = 1/(2\log n)$ and $r^2 = 1/(2\log n)$, respectively, for Model 1 and Model 2. It is seen that the network become sparser as its size increases. The top panel shows the histogram plots when $n = 1500$ and the bottom panel corresponds to $n = 3000$. One can observe that as sample size increases, the χ^2-distribution fits the empirical null distribution better, which is consistent with our theoretical results.

4.2 Hypothesis testing with estimated K

We now examine the finite-sample performance of our test statistics $\widehat{T}_{ij}$ and estimated $\widehat{G}_{ij}$ with estimated K. The simulation settings are identical to those in Section 4.1 except that we explore only the setting with sample size $n = 3000$.

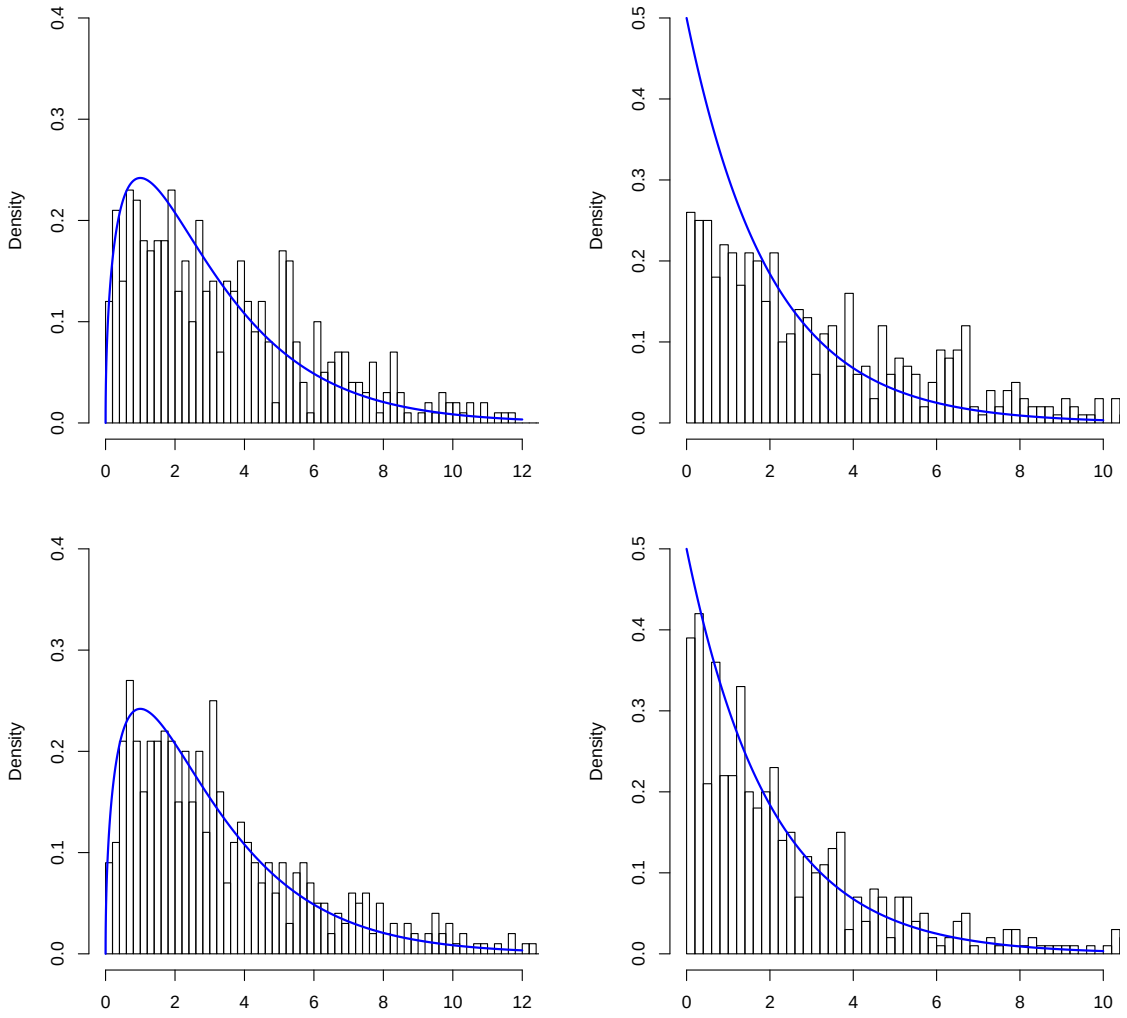

Figure 1: Left: the histogram of test statistic $\widehat{T}_{ij}$ under null hypothesis with known K when $\theta = \frac{1}{2\log n}$. Blue curve is the density function of χ^2_3. Right: the histogram of test statistic $\widehat{G}_{ij}$ under null hypothesis with known K when $r^2 = \frac{1}{2\log n}$. Blue curve is the density function of χ^2_2. Top panel is for sample size $n = 1500$ and bottom panel is for sample size $n = 3000$. Here $n_0 = \frac{n}{5}$.

In Table 3, we report the proportion of correctly estimated K using the thresholding rule (27) in both simulation settings of Models 1 and 2. It is seen that as the signal becomes stronger (i.e., as θ or r^2 increases), the estimation accuracy becomes higher. We also observe that for relatively weak signals, the thresholding rule in (27) tends to underestimate K, resulting in low estimation accuracy. We can see from the same table that over all repetitions, K is either correctly estimated or underestimated. The critical values are constructed based on these estimated values of K.

Same as in Section 4.1, we also examine the empirical size and power of our tests at different levels of signal strength. The results are presented in Table 4. It is seen that the

Table 3: Estimation accuracy of K using the thresholding rule (27)

	θ or r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Model 1	$P(\widehat{K} = K)$	1	1	1	1	1	1	1	1
	$P(\widehat{K} \leq K)$	1	1	1	1	1	1	1	1
Model 2	$P(\widehat{K} = K)$	0	0	0	1	1	1	1	1
	$P(\widehat{K} \leq K)$	1	1	1	1	1	1	1	1

Table 4: The size and power of test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ when the estimated value of K is used. The nominal level is 0.05 and sample size is $n = 3000$.

	θ	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Model 1	Size	0.082	0.066	0.052	0.052	0.044	0.042	0.038	0.062
	Power	0.936	0.994	1	1	1	1	1	1
	r^2	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Model 2	Size	0.054	0.058	0.062	0.058	0.062	0.066	0.064	0.06
	Power	0.074	0.042	0.918	0.972	0.99	1	1	1

performance of $\widehat{T}_{ij}$ is identical to that in Table 2, and the performance of $\widehat{G}_{ij}$ is the same as in Table 2 for all $r^2 > 0.3$. This is expected because of the nearly perfect estimation of K as shown in Table 3 in these scenarios and/or the relatively strong signal strength. When $r^2 \leq 0.3$, $\widehat{G}_{ij}$ has poor power because of the underestimated K (see Table 3). Nevertheless, we observe the same trend as the signal strength increases, which provides support for our theoretical results. We have also applied our tests to nodes with more distinct membership probability vectors $(0.2, 0.6, 0.2)^T$ and $(0, 0, 1)^T$, and the impact of estimated K is much smaller. These additional simulation results are available upon request.

5 Real data applications

5.1 U.S. political data

The U.S. political data set consists of 105 political books sold by an online bookseller in the year of 2004. Each book is represented by a node and links between nodes represent the frequency of co-purchasing of books by the same buyers. The network was compiled by V. Krebs (source: `http://www.orgnet.com`). The books have been assigned manually three labels (conservative, liberal, and neutral) by M. E. J. Newman based on the reviews and descriptions of the books. Note that such labels may not be very accurate. In fact, as argued in multiple papers (e.g., Koutsourelakis and Eliassi-Rad (2008)), the mixed membership model may better suit this data set.

Since our SIMPLE tests $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ do not differentiate network models with or without

mixed memberships, we will view the network as having $K = 2$ communities (conservative and liberal) and treat the neutral nodes as having mixed memberships. To connect our results with the literature, we consider the same 9 books reported in Jin et al. (2017). Another reason of considering the same 9 books as in Jin et al. (2017) is that our test statistic $\widehat{G}_{ij}$ is constructed using the SCORE statistic which is closely related to Jin et al. (2017). The book names as well as labels (provided by Newman) are reported in Table 5. The p-values based on test statistics $\widehat{T}_{ij}$ and $\widehat{G}_{ij}$ for testing the pairwise membership profiles of these 9 nodes are summarized in Tables 6 and 7, respectively.

From Table 7, we see that our results based on test statistic $\widehat{G}_{ij}$ are mostly consistent with the labels provided by Newman and also very consistent with those in Table 5 of Jin (2015). For example, books 59 and 50 are both labeled as "conservative" by Newman and our tests return large p-values between them. These two books generally have much smaller p-values with books labeled as "neutral." Book 78, which was labeled as "conservative" by Newman, seems to be more similar to some neutral books. This phenomenon was also observed in Jin et al. (2017), who interpreted this as a result of having a liberal author. Among the nodes labeled by Newman as "neutral," "All the Shah's Men," or book 29, has relatively larger p-values with conservative books. However, this book has even larger p-values with some other neutral books such as book 104, "The Future of Freedom," which is consistent with the results in Jin et al. (2017) who reported that these two books have very close membership probability vectors. In summary, our SIMPLE method provides statistical significance for the membership probability vectors estimated in Jin et al. (2017).

For a summary of our testing results, we also provide the multidimensional scaling map of the nodes based on test statistics $\widehat{G}_{ij}$ on the left panel of Figure 2. The graph on the right panel of Figure 2 is defined by the pairwise p-value matrix calculated from $\widehat{G}_{ij}$. Specifically, we first apply the hard-thresholding to the p-value matrix by setting all entries below 0.05 to 0. Denote by $\widetilde{P}$ the resulting matrix. Then we plot the graph using the entries of $\widetilde{P}$ as edge weights so that zeros correspond to unconnected pairs of nodes and larger entries mean more closely connected nodes with thicker edges. The nodes in both graphs are color coded according to Newman's labels, with red representing "conservative," blue representing "liberal," and orange representing "neutral." It is seen that both graphs are mostly consistent with Newman's labels, with a few exceptions as partially discussed before. We also would like to mention that the hard-thresholding step in p-value graph is to make the graph less dense and easier to view. In fact, a small perturbation of the threshold does not change much of the overall layout of the graph.

5.2 Stock data

We consider a larger network of stocks in this section. Specifically, daily prices of stocks in the S&P 500 from the period of January 2, 2009 to December 30, 2019 were collected and

Table 5: Political books with labels

Title	Label (by Newman)	Node index
Empire	Neutral	105
The Future of Freedom	Neutral	104
Rise of the Vulcans	Conservative	59
All the Shah's Men	Neutral	29
Bush at War	Conservative	78
Plan of Attack	Neutral	77
Power Plays	Neutral	47
Meant To Be	Neutral	19
The Bushes	Conservative	50

Table 6: P-values based on test statistics $\widehat{T}_{ij}$. The labels provided by Newman are in the parentheses.

Node No.	105(N)	104(N)	59(C)	29(N)	78(C)	77(N)	47(N)	19(N)	50(C)
105(N)	1.0000	0.6766	0.0298	0.3112	0.0248	0.0000	0.0574	0.1013	0.0449
104(N)	0.6766	1.0000	0.0261	0.2487	0.0204	0.0000	0.0643	0.1184	0.0407
59(C)	0.0298	0.0261	1.0000	0.1546	0.2129	0.0013	0.0326	0.0513	0.9249
29(N)	0.3112	0.2487	0.1546	1.0000	0.3206	0.0034	0.0236	0.0497	0.2121
78(C)	0.0248	0.0204	0.2129	0.3206	1.0000	0.0991	0.0042	0.0084	0.2574
77(N)	0.0000	0.0000	0.0013	0.0034	0.0991	1.0000	0.0000	0.0000	0.0035
47(N)	0.0574	0.0643	0.0326	0.0236	0.0042	0.0000	1.0000	0.9004	0.0834
19(N)	0.1013	0.1184	0.0513	0.0497	0.0084	0.0000	0.9004	1.0000	0.1113
50(C)	0.0449	0.0407	0.9249	0.2121	0.2574	0.0035	0.0834	0.1113	1.0000

Table 7: P-values based on test statistics $\widehat{G}_{ij}$. The labels provided by Newman are in the parentheses.

Node No.	105(N)	104(N)	59(C)	29(N)	78(C)	77(N)	47(N)	19(N)	50(C)
105(N)	1.0000	0.4403	0.1730	0.4563	0.8307	0.5361	0.0000	0.0000	0.1920
104(N)	0.4403	1.0000	0.0773	0.9721	0.3665	0.6972	0.0000	0.0000	0.1144
59(C)	0.1730	0.0773	1.0000	0.0792	0.1337	0.0885	0.0000	0.0000	0.8141
29(N)	0.4563	0.9721	0.0792	1.0000	0.4256	0.7624	0.0000	0.0000	0.1153
78(C)	0.8307	0.3665	0.1337	0.4256	1.0000	0.5402	0.0000	0.0000	0.1591
77(N)	0.5361	0.6972	0.0885	0.7624	0.5402	1.0000	0.0000	0.0000	0.1294
47(N)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.9778	0.0000
19(N)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9778	1.0000	0.0000
50(C)	0.1920	0.1144	0.8141	0.1153	0.1591	0.1294	0.0000	0.0000	1.0000

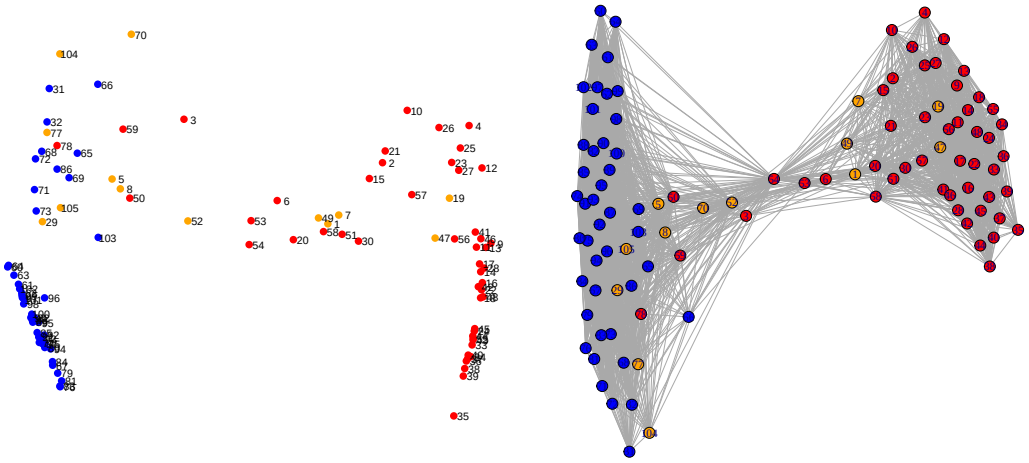

Figure 2: Left panel: the multidimensional scaling map of the nodes based on test statistics $\widehat{G}_{ij}$. Right panel: the connectivity graph generated from the thresholded p-valuate matrix based on $\widehat{G}_{ij}$. The nodes are color coded according to Newman's labels, with red representing "conservative," blue representing "liberal," and orange representing "neutral."

converted into log returns. After some pre-processing (e.g., removing stocks with missing values or very low node degrees), we ended up with 404 stocks. All data analyses in this section were conducted using those 404 stocks. It is well known that much variation in stock excess returns can be captured by factors such as the Fama–French three factors. We first remove these common factors by fitting a factor model, and then the adjacency matrix of stocks is constructed as the correlation matrix of idiosyncratic components from the factor model.

Since stocks are commonly believed to have heterogeneous node degrees, we only apply $\widehat{G}_{ij}$ to the constructed adjacency matrix. The estimated number of communities is $\widehat{K} = 3$. For each pair of stocks, we calculate its p-value using $\widehat{G}_{ij}$ and the asymptotic null distribution χ^2_2. This forms a p-value matrix, denoted as $\mathbf{A}$. To better visualize the results, we provide the multiscale plot of the distance matrix $\mathbf{1} - \mathbf{A}$ with $\mathbf{1}$ the matrix with all entries being 1, and present the results in Figure 3. It is seen that the scatter plot roughly has three legs and a central cluster. The three legs can be interpreted as the three communities with nodes having relatively more pure membership profiles, and the central cluster can be understood as for nodes with mixed membership profiles. For easier visualization, we provide zoomed plots for the three legs and the central cluster in Figure 4. The first three subplots a)–c) correspond to the three legs, and the last subplot d) corresponds to the central cluster. We observe some interesting clustering effects. Figure 4a) corresponds to the top leg in Figure 3. When it is far away from the central cluster (i.e., top left of this subplot), we have stocks mostly related to the retail and restaurant industry (e.g., TGT, HD, LOW, DRI), and when it moves closer to the central cluster (i.e, bottom right of this subplot), the companies are mostly in the real estate (e.g., EXR, VTR, PSA, AVB, PLD). Figure

	HD	L	AAPL	INTC	MCHP	AEE	NEE	EVRG	ADBE
TGT	0.29643	0.71361	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00033
HD	1.00000	0.14934	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00025
L	0.14934	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00031
AAPL	0.00000	0.00000	1.00000	0.00780	0.01933	0.00004	0.00003	0.00010	0.00395
INTC	0.00000	0.00000	0.00780	1.00000	0.00024	0.00000	0.00000	0.00000	0.00148
MCHP	0.00000	0.00000	0.01933	0.00024	1.00000	0.00000	0.00000	0.00000	0.00202
AEE	0.00000	0.00000	0.00004	0.00000	0.00000	1.00000	0.93719	0.46490	0.00467
NEE	0.00000	0.00000	0.00003	0.00000	0.00000	0.93719	1.00000	0.24407	0.00467
EVRG	0.00000	0.00000	0.00010	0.00000	0.00000	0.46490	0.24407	1.00000	0.00465
ADBE	0.00025	0.00031	0.00395	0.00148	0.00202	0.00467	0.00467	0.00465	1.00000

Table 8: The p-value matrix for selected stocks.

4b) mostly consists of tech companies such as AAPL, MCHP, MU, INTC, XLNX, QCOM, ADI, among many others in similar category. Figure 4c) roughly has two subclusters. The left cluster mostly consists of companies in or related to the health industry such as DGX, VAR, GLW, MDT, CERN, TEL, UNH, PFE, BMY, and many other similar ones. The right cluster has predominately companies in the energy industry such as AEE, NEE, EVRG, PNW, DUK, LNT, LNT, ES. Figure 4d) is a zoomed plot that roughly shows the central cluster. It contains a wide range of companies including, but not limited to, risk management and investment companies (BEN, HIG, NDAQ), transportation industry (AAL, NSC, UAL), and communication industry (CTL, VRSN, CTXS).

In Table 8 below, we also present the p-value matrix for selected stocks. The first three stocks (TGT, HD, L) are all in the retail industry, the next three stocks (APPL, INTC, MCHP) are all in the tech industry, stocks 7 to 9 (AEE, NEE, EVRG) are all in the energy industry, and the remaining one (ADBE) is taken from the central cluster. It is seen that the first three groups of stocks have high pairwise p-values within groups, but almost zero p-values with stocks from other groups. In particular, Adobe (ADBE) seems to be connected to most of these selected stocks, which is consistent with the common sense. We would also like to point out that these results were obtained after removing the three common factors from the stock returns, and the clustering structure discovered here should be interpreted as complementary to the ones already captured by the factors.

6 Discussions

In this paper, we have asked a simple yet practical question of how to determine whether any given pair of nodes in a network share the same profile of latent community memberships for large-scale social, economic, text, or health network data with precise statistical significance. Our work represents a first attempt to partially address such an important question. The suggested method of statistical inference on membership profiles in large networks (SIMPLE)

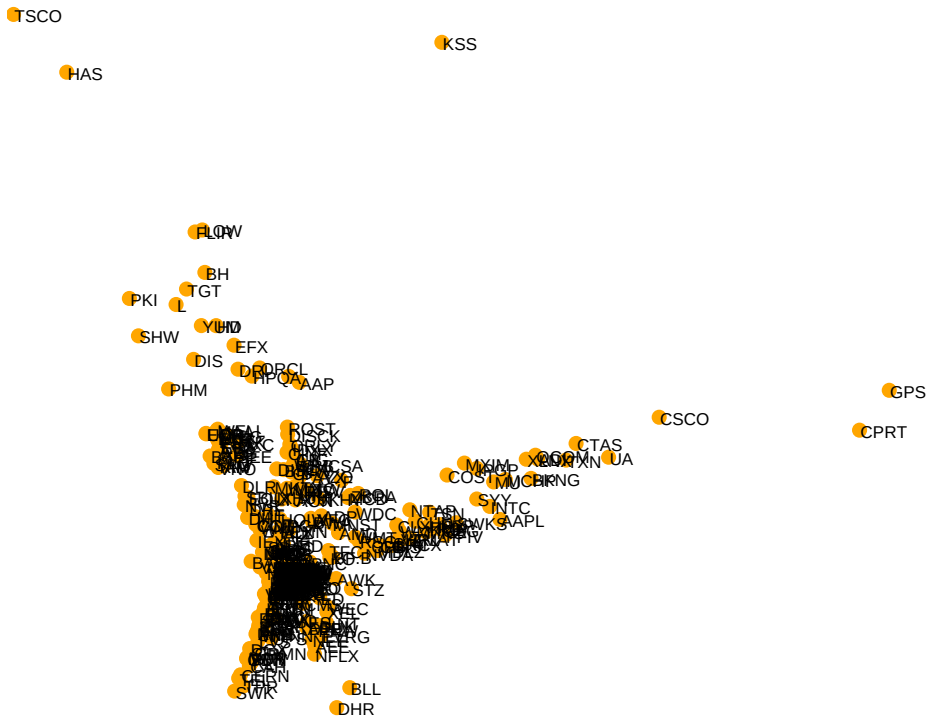

Figure 3: Multiscale plot based on the distance matrix $\mathbf{1} - \mathbf{A}$, where $\mathbf{1}$ is the matrix with all entries being 1 and $\mathbf{A}$ is the p-value matrix based on $\widehat{G}_{ij}$. It is seen that the scatter plot roughly has three legs and a central cluster.

provides theoretically justified network p-values in our context for both settings of mixed membership models and degree-corrected mixed membership models. We have formally shown that the two forms of SIMPLE test statistics can enjoy simple limiting distributions under the null hypothesis and appealing power under the contiguous alternative hypothesis. In particular, the tuning-free feature of SIMPLE makes it easy to use by practitioners. Our newly suggested method and established theory lay the foundation for practical policies or recommendations rooted on statistical inference for network data with quantifiable impacts.

To illustrate the key ideas of SIMPLE and simplify the technical analysis, we have focused our attention on the hypothesis testing problem for any preselected pair of nodes. It would be interesting to study the problem when one of or each of the nodes is replaced by a selected set of nodes. For example, in certain applications one may have some additional knowledge that all the nodes within the selected set indeed share the same membership profile information. It would also be interesting to quantify and control the statistical inference error rates when one is interested in performing a set of hypothesis tests simultaneously for network data. Moreover, it would be interesting to investigate the hypothesis testing problem for more general network models as well as for statistical models beyond network data such as for large collections of text documents.

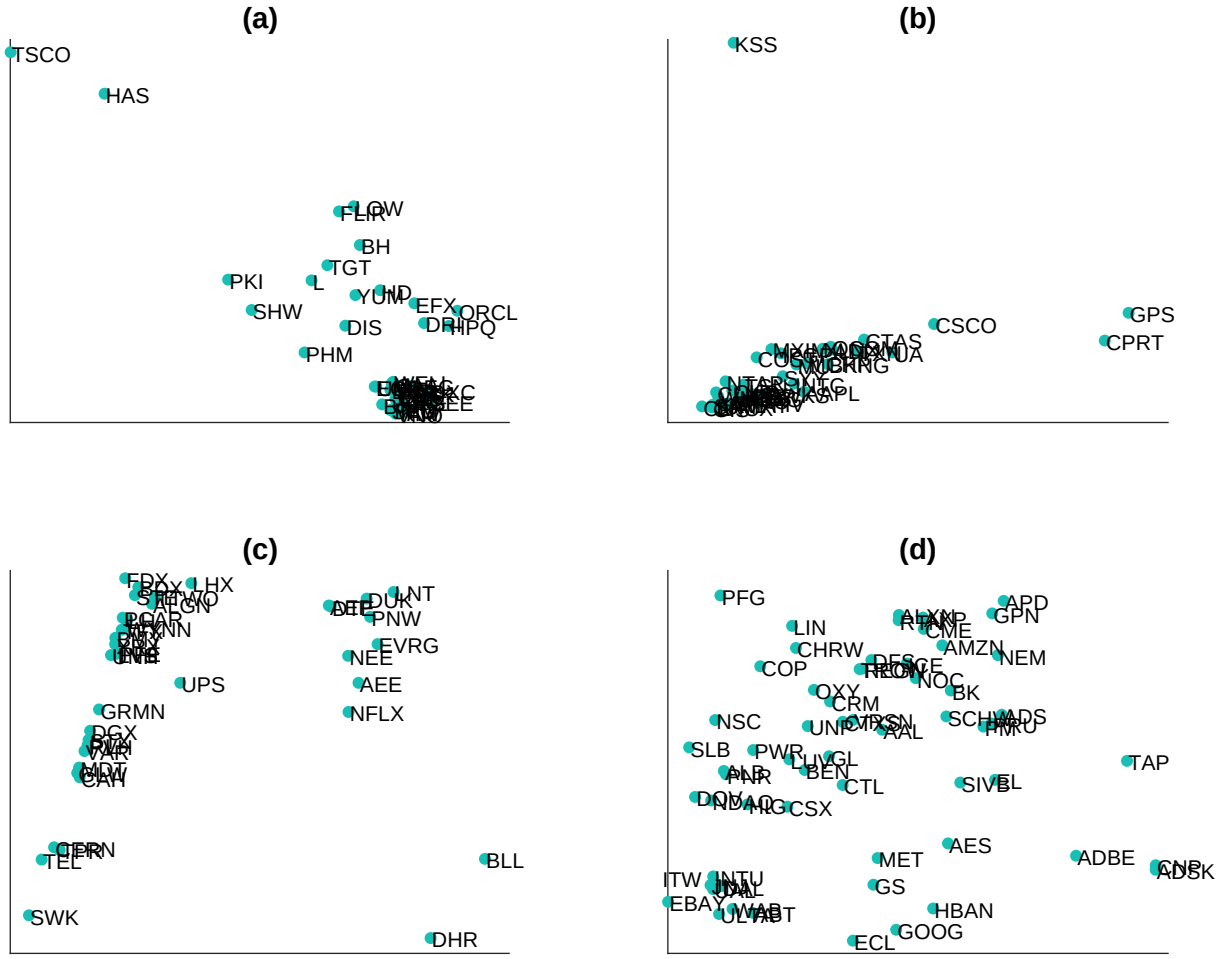

Figure 4: Zoomed multiscale plots based on the distance matrix $\mathbf{1} - \mathbf{A}$, where $\mathbf{A}$ is the p-value matrix based on $\widehat{G}_{ij}$.

In addition, it would be interesting to connect the growing literature on sparse covariance matrices and sparse precision matrices with that on network models. Such connections can be made via modeling the graph Laplacian through a precision matrix or covariance matrix (Brownlees et al., 2019). A natural question is then how well the network profiles can be inferred from a panel of time series data. The same question also arises if the panel of time series data admits a factor structure (**?**Fan et al., 2013). These problems and extensions are beyond the scope of the current paper and will be interesting topics for future research.

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research 18*, 177:1–177:86.

Abbe, E., J. Fan, K. Wang, and Y. Zhong (2017). Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*.

Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research 9*, 1981–2014.

Arias-Castro, E. and N. Verzelen (2014, 06). Community detection in dense random networks. *Ann. Statist. 42*(3), 940–969.

Bickel, P. J. and P. Sarkar (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society Series B 78*, 253–273.

Billingsley, P. (1995). *Probability and Measure*. Wiley.

Brownlees, C., G. G. Stefan, and G. Lugosi (2019). Community detection in partial correlation network models. *Manuscript*.

Chen, K. and J. Lei (2018). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association 113*, 241–251.

Daudin, J.-J., F. Picard, and S. Robin (2008). A mixture model for random graphs. *Statistics and Computing 18*, 173–183.

Fan, J., Y. Fan, X. Han, and J. Lv (2020). Asymptotic theory of eigenvectors for random matrices with diverging spikes. *Journal of the American Statistical Association*, to appear.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society Series B 75*, 603–680.

Gao, C., Z. Ma, A. Y. Zhang, and H. H. Zhou (2018, 10). Community detection in degree-corrected block models. *Ann. Statist. 46*(5), 2153–2185.

Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2010). A survey of statistical network models. *Found. Trends Mach. Learn. 2*, 129–233.

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

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

Jin, J. and Z. T. Ke (2017). A sharp lower bound for mixed-membership estimation. *arXiv preprint arXiv:1709.05603*.

Jin, J., Z. T. Ke, and S. Luo (2017). Estimating network memberships by simplex vertex hunting. *arXiv preprint arXiv:1708.07852*.

Karrer, B. and M. E. J. Newman (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E 83*, 016107.

Koutsourelakis, P.-S. and T. Eliassi-Rad (2008). Finding mixed-memberships in social networks. *AAAI Spring Symposium: Social Information Processing*, 48–53.

Latouche, P., E. Birmelé, and C. Ambroise (2012). Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling 12*, 93–115.

Lei, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist. 44*, 401–424.

Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist. 43*, 215–237.

Liben-Nowell, D. and J. Kleinberg (2007). The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol. 58*, 1019–1031.

Newman, M. E. J. (2013a). Community detection and graph partitioning. *EPL (Europhysics Letters) 103*, 28003.

Newman, M. E. J. (2013b). Spectral methods for community detection and graph partitioning. *Phys. Rev. E 88*, 042822.

Newman, M. E. J. and T. P. Peixoto (2015). Generalized communities in networks. *Phys. Rev. Lett. 115*, 088701.

Raič, M. (2019). A multivariate Berry-Esseen theorem with explicit constants. *Bernoulli 25*(4A), 2824 – 2853.

Rohe, K., S. Chatterjee, and B. Yu (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics 39*, 1878–1915.

Saldana, D., Y. Yu, and Y. Feng (2017). How many communities are there? *Journal of Computational and Graphical Statistics 26*, 171–181.

Tropp, J. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math. 12*, 389–434.

Verzelen, N. and E. Arias-Castro (2015, 12). Community detection in sparse random networks. *Ann. Appl. Probab. 25*(6), 3465–3510.

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

Wang, Y. J. and G. Y. Wong (1987). Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association 82*, 8–19.

Wang, Y. X. R. and P. J. Bickel (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist. 45*, 500–528.

Wu, Y.-J., E. Levina, and J. Zhu (2018). Link prediction for egocentrically sampled networks. *arXiv preprint arXiv:1803.040845*.

Zhang, P. and C. Moore (2014). Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences 111*, 18144–18149.

Zhang, Y., E. Levina, and J. Zhu (2020). Detecting overlapping communities in networks using spectral methods. *SIAM Journal on Mathematics of Data Science 2*(2), 265–283.

Zhao, Y., E. Levina, and J. Zhu (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist. 40*, 2266–2292.