Statistica Sinica **34** (2024), 87-110
doi:https://doi.org/10.5705/ss.202021.0306

A SPECTRAL-BASED FRAMEWORK FOR HYPOTHESIS TESTING IN POPULATIONS OF NETWORKS

Li Chen[1,2], Nathaniel Josephs*[3], Lizhen Lin[4],
Jie Zhou[2] and Eric D. Kolaczyk[5]

[1]*Southwest Minzu University,* [2]*Sichuan University,* [3]*North Carolina State University,* [4]*University of Maryland and* [5]*McGill University*

Abstract: We propose a new spectral-based approach to hypothesis testing for populations of networks. The primary goal is to develop a test to determine whether two given samples of networks come from the same random model or distribution. Our test statistic is based on the trace of a centered and scaled adjacency matrix to the third power, which we prove converges to the standard normal distribution as the number of nodes tends to infinity. We also provide the asymptotic power guarantee of the test. We explore the relationship between the number of networks and the number of nodes in each network when characterizing the theoretical properties of the proposed test statistic. Our test can be applied to both binary and weighted networks, operates under a very general framework in which the networks are allowed to be large and sparse, and can be extended to multiple-sample testing. We present a simulation study that demonstrates the superior performance of our tests over that of existing methods, and apply our tests to three real data sets.

Key words and phrases: Hypothesis testing, populations of networks, random matrix theory.

1. Introduction

In this work, we consider an inference problem related to populations of networks in which each sample or data point is a network. Most existing works on statistical network analysis focus on models and algorithms that can be used to analyze a single network. However, the increasing prevalence of multiple-network data sets, in which the network is the fundamental data object, has motivated the need for statistical inference methods for populations of networks, from which we can extract useful scientific information.

For example, in the brain network data examined in Section 5, one may be interested in testing whether a brain network structure from a group of individuals with schizophrenia differs from that of a group of healthy controls. Given a collection or sample of such networks, one might also be interested in estimating some mean network feature, which would enable us to average networks or cluster networks into groups (Mukherjee, Sarkar and Lin (2017)). These are all inference

*Corresponding author.

tasks for one or two samples of network objects, both of which have been explored in the literature.

Ginestet et al. (2017) consider two-sample testing for networks, with applications to functional neuroimaging. Kolaczyk et al. (2020) extended this work using a geometric and statistical framework for inference on populations of unlabeled networks. They did so by providing a geometric characterization of the space of unlabeled networks and deriving a central limit theorem for the sample Fréchet mean. Supervised and unsupervised learning, such as clustering, regression, and classification for network objects, have also been considered in the literature. See, for example, Arroyo Relión et al. (2019) and Josephs et al. (2020); the former consider network classification in neuroimaging, and the latter use Bayesian methods for classification, anomaly detection, and survival analysis.

Here, we focus on the problem of two-sample hypothesis testing for populations of networks. There are several such hypothesis tests in the literature, but these typically make assumptions on the network model. For example, Tang et al. (2017) study whether or not two networks ($m = 2$) defined on different vertex sets are generated from the same random dot product graph model. Ghoshdastidar et al. (2020) study two-sample problems from a minimax perspective that test whether two samples of binary networks of n nodes are generated from the same link probability matrix, against an alternative that the two link probability matrices are ρ apart with respect to some matrix norm. Their work focuses on a theoretical characterization of minimax separation with respect to the number of networks m, the number of nodes n, and different matrix norms. Ghoshdastidar and von Luxburg (2018) apply the same test statistic, and prove that it converges to a normal distribution asymptotically. Recently, Yuan and Wen (2021) modified the test statistic in Ghoshdastidar and von Luxburg (2018), proposing a new test for weighted graph two-sample hypothesis testing.

One straightforward alternative to two-sample testing for networks is to convert the networks into vector values, and then to apply a two-sample, high-dimensional mean test. This strategy has been widely studied in the literature (Chen and Qin (2010); Cai, Liu and Xia (2014); Xu et al. (2016)). Although this approach is model free, it may lose information in the conversion process, which essentially ignores the interconnectedness that defines the network data. We return to this discussion in Section 4.

In contrast to most existing works, such as Ginestet et al. (2017), in which the number of nodes is fixed, we consider a general framework that allows both the number of nodes and the sample size (the number of networks) to grow. Our test statistics are spectral based and not restricted to a given network structure. We use the trace of the third power of a centered and scaled adjacency matrix, which is proven to converge to the standard normal distribution as the number of nodes tends to infinity. In addition, we show that the asymptotic power tends to one as the number of nodes increases. Because we also want to understand

the limiting behavior as the sample size increases, we explore the relationship between the asymptotics in the number of networks and in the number of nodes for each network when characterizing the theoretical properties of our proposed test statistics. These statistics are conceptually simple and computational friendly, and we discuss an extensive simulation study that we conducted under various models to demonstrate the superior performance of our test over that of existing methods. In almost all cases we examine here, the proposed test statistics achieve the nominal rejection rate under the null, and a power close to one under the alternative. We also apply our test to three real data sets of weighted and binary networks.

The idea of applying a spectral method based on random matrix theory to network data is a natural one, because network data (e.g., the adjacency or Laplacian matrix) can naturally be viewed as a random matrix. Our method is motivated by Dong, Wang and Liu (2020), who propose a spectral-based hypothesis test for testing the community structure within a single network. The authors prove that their test statistic, which is similar to that in Bickel and Sarkar (2016), converges quickly to the normal distribution. However, it is limited to testing the presence of a community structure in a single network versus the null Erdős–Rényi model. In our work, we extend the statistic to test the difference between arbitrary network models. The proposed statistic can be applied to either binary or weighted networks in both two-sample and multiple-sample frameworks. A spectral-based test based on a Tracy–Widom law for hypothesis testing of populations of networks and change-point detection in networks can also be found in Chen, Lin and Zhou (2020) and Chen, Zhou and Lin (2021). Compared with these two works, our spectral-based test has an asymptotic standard normal distribution, and a much faster convergence rate under the null compared with the slow convergence of tests based on a Tracy–Widom law. Furthermore, our test statistics require much milder conditions for the theoretical performance guarantees: we need an error estimate of the link probability estimates of $o_p(1)$, compared with the error condition of $o_p(n^{-2/3})$ required by Chen, Lin and Zhou (2020).

The remainder of the paper is organized as follows. In Section 2, we describe our proposed spectral-based test statistic, and derive its asymptotic null distribution and an asymptotic power result. We extend our test for weighted networks and multiple-sample testing in Section 3. In Section 4, we report the results of extensive simulation studies, and in Section 5, we analyze three real network data sets. We conclude the paper in Section 6.

2. A New Spectral-Based Test for Binary Networks

In this section, we first propose a new spectral-based test for testing the difference between distributions of two samples of binary networks. Specifically,

we consider two samples of networks on the same n nodes with possibly different sample sizes' m_1 and m_2, respectively. We assume that we observe the independent and identically distributed (i.i.d.) symmetric binary adjacency matrices $A_1^{(1)}, \ldots, A_1^{(m_1)}$, with conditionally independent entries generated from a symmetric link probability matrix P_1, that is,

$$A_{1,ij}^{(k)} \sim \text{Bernoulli}(P_{1,ij}),$$

for $k = 1, 2, \ldots, m_1$ and $i, j = 1, 2, \ldots, n$. Similarly, we observe a second sample of adjacency matrices $A_2^{(1)}, \ldots, A_2^{(m_2)}$ with

$$A_{2,ij}^{(k)} \sim \text{Bernoulli}(P_{2,ij}),$$

generated from the same model with link probability matrix P_2. Assume that there are no self-loops, that is, $A_{u,ii}^{(k)} = 0$, for $u = 1, 2$, $i = 1, \ldots, n$, and $k = 1, \ldots, m_u$. Our goal is to test whether the two samples of networks have the same graph structure, which is equivalent to testing

$$H_0 : P_1 = P_2 \text{ versus } H_1 : P_1 \neq P_2. \tag{2.1}$$

To address this, we propose a new statistic that uses results from random matrix theory. For some background on the spectral properties of inhomogeneous networks, which are used heavily in this work, see the online Supplementary Material.

2.1. New spectral test for binary networks

Given two samples of networks $\{A_1^{(k)}\}_{k=1}^{m_1}$ and $\{A_2^{(k)}\}_{k=1}^{m_2}$, sampled from the link probability matrices P_1 and P_2, respectively, we introduce the normalized matrix with elements as follows:

$$Z_{ij} = \begin{cases} \frac{\bar{A}_{1,ij} - \bar{A}_{2,ij}}{\sqrt{n\{(1/m_1)P_{1,ij}(1-P_{1,ij}) + (1/m_2)P_{2,ij}(1-P_{2,ij})\}}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases}, \tag{2.2}$$

where $\bar{A}_u$ is the sample average of the adjacency matrices in the uth group, for $u = 1, 2$,

$$\bar{A}_u = \frac{1}{m_u} \sum_{k=1}^{m_u} A_u^{(k)}, \tag{2.3}$$

and B is an $n \times n$ diagonal matrix with, B_{ii} given by i.i.d. random variables such that

$$P\left(B_{ii} = -\frac{1}{\sqrt{n}}\right) = P\left(B_{ii} = \frac{1}{\sqrt{n}}\right) = 1/2, \tag{2.4}$$

for $i = 1, \ldots, n$.

Consider the test statistic

$$\theta = \frac{1}{\sqrt{15}}\mathrm{Tr}(Z^3), \tag{2.5}$$

where $\mathrm{Tr}(\cdot)$ represents the trace operator. This statistic is an extension of that in Dong, Wang and Liu (2020), which was inspired by a result in Bai and Silverstein (2010). Under the null hypothesis, we have the following theorem on the asymptotic distribution of θ.

Theorem 1. *Let Z be given as in* (2.2). *Assume the sample size satisfies $m_u = O(n^{\alpha_u})$, for some $\alpha_u > 0, u = 1, 2$. Then, under the null hypothesis $P_1 = P_2$, for the scaled test statistic $\theta = (1/\sqrt{15})\,Tr(Z^3)$, we have*

$$\theta \xrightarrow{d} \mathcal{N}(0,1) \quad as \quad n \to \infty, \tag{2.6}$$

where $\xrightarrow{d}$ denotes weak convergence.

We defer the details of the proof to the Supplementary Material. However, an overview of the argument is as follows. First, under the null hypothesis of $P_1 = P_2$, Z is a Wigner matrix satisfying $\mathrm{E}(Z_{ij}) = 0$ and $\mathrm{Var}(Z_{ij}) = 1/n$. Then, we verify that $X = \sqrt{n}Z$ satisfies conditions (1)–(3) of Lemma 1 in the Supplementary Material, after which, the asymptotic normality of θ follows. Lastly, we obtain the mean and variance following Dong, Wang and Liu (2020).

To formalize a testing framework using θ in (2.5), we need to account for the fact that the diagonal matrix B in (2.4) is random. We do so by employing a Monte Carlo procedure, which we describe in Algorithm 1. Our output is an empirical significance level, which is the rejection rate based on the test statistics computed from the Monte Carlo samples of B.

Remark 1. In Algorithm 1, we deliberately do not output a p-value. For $Q = 1$, we can obtain a p-value using $2P\big(\theta > |\theta_{obs}^{(Q=1)}|\big)$, as in Bickel and Sarkar (2016) and Dong, Wang and Liu (2020), where $\theta_{obs}^{(Q=1)}$ is the sample test statistic and θ follows the null distribution of the test statistic. However, in this case, the p-value is implicitly conditional on B, and the authors' simulations reveal that the randomness of B leads to highly variable p-values. Instead, for our test, we propose computing many $\theta_{obs}^{(q)}$ in parallel to reduce the noise induced by B. The analogous p-value estimate combining these Monte Carlo test statistics is $2P\big(\theta > |\bar{\theta}_{obs}|\big)$, where $\bar{\theta}_{obs} = (1/Q)\sum_{q=1}^{Q}\theta_{obs}^{(q)}$.

Algorithm 1: Procedure for testing using the statistic in (2.5). The output is an empirical significance level based on Monte Carlo test statistics, where $I(\cdot)$ is an indicator function and $\mu_{\alpha/2}$ is the $\alpha/2$ upper quantile of $\mathcal{N}(0,1)$.

New Spectral-Based Hypothesis Test $\big(\{A_1^{(k)}\}_{k=1}^{m_1}, \{A_2^{(k)}\}_{k=1}^{m_2}, \alpha, Q\big)$;

Input : Adjacency matrices $\{A_1^{(k)}\}_{k=1}^{m_1}$ and $\{A_2^{(k)}\}_{k=1}^{m_2}$ for groups 1 and 2
Significance level α
Number of Monte Carlo samples Q

Output: Empirical significance level rej_rate

Compute $\bar{A}_u$ for $u = 1, 2$ using (2.3);

for $q = 1, \ldots, Q$ **do in parallel**

  Sample $B^{(q)}$ satisfying (2.4);

  Compute $Z^{(q)}$ in (2.2) using $B^{(q)}$;

  Compute $\theta^{(q)}$ in (2.5) using $Z^{(q)}$;

end

rej_rate $= (1/Q)\sum_{q=1}^{Q} I\big(|\theta^{(q)}| > \mu_{\alpha/2}\big)$

Remark 2. The rejection rate from our Monte Carlo estimator has the property that its expectation under the null is the nominal significance level:

$$\mathrm{E}\left(\frac{1}{Q}\sum_{q=1}^{Q} I\big(|\theta^{(q)}| > \mu_{\alpha/2}\big)\right) = P\big(|\theta^{(q)}| > \mu_{\alpha/2}\big) = \alpha.$$

2.2. Test statistic based on estimated link probability matrices

Theorem 1 assumes that the true link probability matrices P_1 and P_2 are known, which is not the case in practice. Therefore, θ cannot be used directly as a test statistic. A natural alternative is to plug in appropriate estimates of P_1 and P_2, with the hope that the plug-in estimator for the test statistic retains its asymptotic normality.

We denote the plug-in estimates of P_1 and P_2 by $\hat{P}_1$ and $\hat{P}_2$, respectively. Then, the empirical version of the normalized matrix Z in (2.2) can be written as

$$\hat{Z}_{ij} = \begin{cases} \dfrac{\bar{A}_{1,ij} - \bar{A}_{2,ij}}{\sqrt{n\{(1/m_1)\hat{P}_{1,ij}(1-\hat{P}_{1,ij}) + (1/m_2)\hat{P}_{2,ij}(1-\hat{P}_{2,ij})\}}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases}. \tag{2.7}$$

The resulting test statistic is

$$\hat{\theta} = \frac{1}{\sqrt{15}}\mathrm{Tr}(\hat{Z}^3), \tag{2.8}$$

which has the following limiting law.

Theorem 2. *Under the two-sample framework of binary networks, let $\hat{Z}$ be given in* (2.7)*. As before, assume the sample size $m_u = O(n^{\alpha_u})$, and $\hat{P}_u$ is some*

estimate of P_u, for some $\alpha_u > 0$, $u = 1, 2$. If $\max_{i,j} |\hat{P}_{u,ij} - P_{u,ij}| = o_p(1)$, then, under the null hypothesis $P_1 = P_2$, we have the following asymptotic distribution of the scaled test statistic $\hat{\theta} = (1/\sqrt{15})\, Tr(\hat{Z}^3)$:

$$\hat{\theta} \xrightarrow{d} \mathcal{N}(0, 1) \quad as \quad n \to \infty.$$

Again, we defer the proof to the Supplementary Material, which relies on rewriting

$$\mathrm{Tr}(\hat{Z}^3) = \mathrm{Tr}(Z^3) + 3\mathrm{Tr}\big\{Z^2(Z \circ H)\big\} + 3\mathrm{Tr}\big\{Z(Z \circ H)^2\big\} + \mathrm{Tr}\big\{(Z \circ H)^3\big\},$$

where $\circ$ denotes the Hadamard product, and H is an $n \times n$ matrix with entries $H_{ij} = o_p(1)$. Each term on the right-hand side of this equality can be proven to be $o_p(1)$.

2.3. Estimating link probability matrices

As it is, our test statistic in (2.8) is really for a two-sample matrix testing problem for a difference of means. However, this becomes a network test when we estimate the link probability matrices. Here, we propose three methods that satisfy the conditions in Theorem 2, which require that the sample sizes of the observed networks grow with n at a rate of n^α, for any $\alpha > 0$, and $\max_{i,j} |\hat{P}_{u,ij} - P_{u,ij}| = o_p(1)$.

The simplest estimator of $P_{u,ij}$ is the sample mean of all (i, j) elements in the adjacency matrices of group u. We refer to this spectral method based on simple averages as SPE-AVG. It is not difficult to see that $\max_{i,j} |\hat{P}_{u,ij} - P_{u,ij}| = o_p(m_u^{-1/2} \log n)$ by applying Bernstein's inequality (Bernstein (1946)). Intuitively, SPE-AVG requires large sample sizes to achieve good performance. This is confirmed empirically by our extensive simulation studies, in which SPE-AVG typically yields inferior performance compared with that of other methods.

Another possible average estimator of $P_{u,ij}$ is based on the stochastic block model (SBM). The key idea is to approximate a graph with an SBM, which, for large networks, is reasonable, by Szemerédi's regularity lemma (Lovász (2012)). The membership vector of nodes can be obtained using community algorithms, such as the method proposed in Ng, Jordan and Weiss (2002). After the membership vector has been estimated, we can simply approximate $P_{u,ij}$ using the sample mean of all entries in the submatrix over all $A_u^{(k)}$, for $k = 1, 2, \ldots, m_u$, restricted to the corresponding block consisting of the communities of i and j. We refer to this test method based on an SBM as SPE-SBM. Assuming the true community number is K_u, the estimation error satisfies $\max_{i,j} |\hat{P}_{u,ij} - P_{u,ij}| = o_p(K_u m_u^{-1/2} n^{-1} \log n)$. Thus, the rate of SPE-SBM is better than that of SPE-AVG as long as $K_u < m_u^{1/2} n^{1-\beta}$, with β a small positive number, which is very easily satisfied. However, the property may be limited by the assumption that

the network topologies follow an SBM structure.

Finally, we introduce a new estimation method based on the modified neighborhood smoothing (MNBS) proposed in Zhao, Chen and Lin (2019). The idea is to perform neighborhood smoothing on the matrix $\bar{A}$, which is the weighted average of m networks, and then to apply the smoothing procedure to a shrunken neighborhood size. This results in a better bias–variance trade-off leading to a better estimate of the link probability matrix, with a smaller error. Note that MNBS is essentially an NBS method applied to $\bar{A}$ instead of to the adjacent matrix of a single network, and with a shrunken neighborhood size. This reduces the variance due to the multiple networks available in the each sample. From Lemma 9.3 in Zhao, Chen and Lin (2019), the size of a neighborhood is $O_p((n \log n/m_u)^{1/2})$. Using this and Bernstein's inequality, the estimation error of the link probability is $|\hat{P}_{u,ij} - P_{u,ij}| = \max\big(O_p\big((m_u n \log n)^{-1/4}\big), O_n(n^{-1}\log n), O_n((m_u n/\log n)^{-1/2})\big)$. For the technical details, see Section S4.1 in the Supplementary Material. We refer to this test method based on MNBS as SPE-MNBS. Note that SPE-MNBS places no structure conditions on the networks. Therefore, we expect the method to be generally applicable.

2.4. Asymptotic power guarantee

Next, we consider the power of the test based on $\hat{\theta}$ in (2.8), which we summarize in the following theorem.

Theorem 3. *Consider the alternative model of $P_1 \neq P_2$ under the assumptions of Theorem* 1. *Let Z'' be an $n \times n$ matrix with zero diagonals and, for any $i \neq j$,*

$$Z''_{ij} = \frac{P_{1,ij} - P_{2,ij}}{\sqrt{n\{(1/m_1)P_{1,ij}(1-P_{1,ij}) + (1/m_2)P_{2,ij}(1-P_{2,ij})\}}}. \tag{2.9}$$

Define the partition $\{1,\ldots,n\}^3 = S_a \cup S_b \cup S_c$, where $(i,k,l) \in S_a$, S_b, and S_c indicates that $Z''_{ik}Z''_{kl}Z''_{li} > 0$, $Z''_{ik}Z''_{kl}Z''_{li} < 0$, and $Z''_{ik}Z''_{kl}Z''_{li} = 0$, respectively. Let $|S_a| = an^3$, $|S_b| = bn^3$, and $|S_c| = cn^3$, with $a,b,c \in [0,1]$ satisfying $a+b+c=1$. If either

$$(i) \quad an^3 \min_{(i,k,l)\in S_a} (Z''_{ik})^3 + bn^3 \min_{(i,k,l)\in S_b} (Z''_{ik})^3 > 0 \text{ or}$$

$$(ii) \quad -an^3 \max_{(i,k,l)\in S_a} (Z''_{ik})^3 - bn^3 \max_{(i,k,l)\in S_b} (Z''_{ik})^3 > 0$$

is satisfied, then

$$\lim_{n\to\infty} P(|\hat{\theta}| > \mu_{\alpha/2}) = 1, \ \alpha > 0.$$

The proof is given in the Supplementary Material.

Remark 3. Note that there is a slight abuse of notation in our conditions (i) and (ii), where the minimum or maximum operator is taken over all pairs of indices among (i, j, k). These conditions characterize the minimum signal difference between P_1 and P_2 required for Theorem 3 to hold, which implies that the power is asymptotically one when either of the sets S_a or S_b is sufficiently large. For a better understanding of this, consider the case in which $P_{1,ij} \geq P_{2,ij}$, for all i and j, that is, $Z''_{ij} \geq 0$, and Theorem 3 holds as long as $a > O((m_u n)^{-3/2})$, which is a very mild condition.

Remark 4. The separation conditions in Theorem 3 arise in our proof as a characterization of the signal difference between two link probability matrices. Importantly, this characterization is on the whole network, rather than on the method of network moments or on motifs for network data (the frequencies of particular patterns such as triangles, stars, or wheels), which are studied in Gao and Lafferty (2017), Banerjee and Ma (2017), Jin, Ke and Luo (2021), Zhang and Xia (2020), and Bhattacharya, Das and Mukherjee (2020).

3. Extending our Test to Other Settings

In this section, we extend our test for weighted networks, and for multiple samples, in a manner analogous to a one-way analysis of variance (ANOVA).

3.1. Extension to weighted networks

We now consider a more general framework that focuses on weighted networks. Let $F_1 = \{F_{1,ij}\}$ and $F_2 = \{F_{2,ij}\}$, for $i, j = 1, \ldots, n$, be two sequences of distributions defined on bounded intervals and specified by some parameters. Let $A_1^{(1)}, \ldots, A_1^{(m_1)} \overset{i.i.d.}{\sim} F_1$ and $A_2^{(1)}, \ldots, A_2^{(m_2)} \overset{i.i.d.}{\sim} F_2$ be symmetric weighted adjacency matrices for networks that are undirected and without self-loops, that is, $A_{u,ii}^{(k)} = 0$, for $u = 1, 2$, $i = 1, \ldots, n$, and $k = 1, \ldots, m_u$. Let Σ_u denote an $n \times n$ matrix in which the (i, j) element is the variance of $A_{u,ij}^{(k)}$. Note that its diagonal elements are zero, because $A_{u,ii}^{(k)} = 0$. Finally, let $\hat{\Sigma}_{u,ij}$ be an estimate of $\Sigma_{u,ij}$.

Our approach for weighted networks is to replace $P_{u,ij}(1 - P_{u,ij})$ in (2.2) and $\hat{P}_{u,ij}(1 - \hat{P}_{u,ij})$ in (2.7) with $\Sigma_{u,ij}$ and $\hat{\Sigma}_{u,ij}$, respectively. Just as in Section 2.3, the estimates $\hat{\Sigma}_{u,ij}$ can be obtained using various methods, which are discussed later. For simplicity, we use the same notation as in Section 2.3.

For the weighted case, the testing problem in (2.1) is equivalent to

$$H_0 : F_1 = F_2 \text{ versus } H_1 : F_1 \neq F_2. \tag{3.1}$$

We define the normalized matrix Z as

$$Z_{ij} = \begin{cases} \frac{\bar{A}_{1,ij} - \bar{A}_{2,ij}}{\sqrt{n\{(1/m_1)\Sigma_{1,ij} + (1/m_2)\Sigma_{2,ij}\}}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases}, \tag{3.2}$$

where B is defined as in (2.4). Then, the asymptotic distribution of $\theta = (1/\sqrt{15})\mathrm{Tr}(Z^3)$ follows a standard normal distribution under the null hypothesis, as stated in the following theorem.

Theorem 4. *Under the two-sample framework of weighted networks, let Z be given in (3.2). Assume a sample size $m_u = O(n^{\alpha_u})$, for some $\alpha_u > 0$, $u = 1, 2$. Then, under the null hypothesis $F_1 = F_2$, for the scaled test statistic $\theta = (1/\sqrt{15})\,Tr(Z^3)$, we have*

$$\theta \xrightarrow{d} \mathcal{N}(0, 1) \quad as \quad n \to \infty. \tag{3.3}$$

The proof is omitted because it is similar to that of Theorem 1.

Remark 5. Although the two-sample testing framework for binary networks is a special case of that in (3.1), we discuss the two cases separately. In the binary case, our test statistic is obtained by plugging in an estimate of the link probability matrix P, whereas our test statistic for the weighted networks requires a plug-in estimate of the variance of each edge weight. Hence, the estimation methods differ for these two cases.

For practical applications, we need to estimate the covariance matrices Σ_1 and Σ_2, assuming some conditions to ensure that the asymptotic normality of the new test statistic still holds. For $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, the plug-in estimates of Σ_1 and Σ_2, respectively, the empirical normalized matrix of Z in (3.2) can be written with entries as

$$\hat{Z}_{ij} = \begin{cases} \frac{\bar{A}_{1,ij} - \bar{A}_{2,ij}}{\sqrt{n\{(1/m_1)\hat{\Sigma}_{1,ij} + (1/m_2)\hat{\Sigma}_{2,ij}\}}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases}. \tag{3.4}$$

Therefore, our test statistic is

$$\hat{\theta} = \frac{1}{\sqrt{15}}\mathrm{Tr}(\hat{Z}^3). \tag{3.5}$$

Then, we have the following limiting law.

Theorem 5. *Under the two-sample framework of weighted networks, let $\hat{Z}$ be given as in (3.4). Assume the sample size $m_u = O(n^{\alpha_u})$ and $\hat{\Sigma}_u$ is some estimate of Σ_u, for some $\alpha_u > 0$, $u = 1, 2$. If $\max_{i,j} |\hat{\Sigma}_{u,ij} - \Sigma_{u,ij}| = o_p(1)$, then under the null hypothesis $F_1 = F_2$, we have the following asymptotic distribution of the scaled test statistic $\hat{\theta} = (1/\sqrt{15})\,Tr(\hat{Z}^3)$:*

$$\hat{\theta} \xrightarrow{d} \mathcal{N}(0, 1) \quad as \quad n \to \infty.$$

The proof is similar to that of Theorem 2, so we include only the key differences in the Supplementary Material; the remainder of the proof can be completed in a straightforward manner.

We consider two estimates of $\Sigma_{u,ij}$. The first is obtained simply as the sample variance of each element over all adjacency matrices in the same group. For convenience, we still refer to this method as SPE-AVG. Then, we have

$$\max_{i,j} |\hat{\Sigma}_{u,ij} - \Sigma_{u,ij}| = o_p(m_u^{-1/2} \log m_u). \tag{3.6}$$

The proof of (3.6) is available in the Supplementary Material (see Section S4.6). The order of the error is the same as the binary case, which implies that SPE-AVG is suitable for large sample sizes.

The second estimate of $\Sigma_{u,ij}$ is obtained similarly to SPE-SBM for unweighted networks: assume each network comes from an SBM, approximate the community membership vector, and compute the sample covariance within each community as the sample variance of the nodes corresponding to that community block (rather than the sample mean). Again, we refer to this method as SPE-SBM, as in the binary case. Using a similar argument to that in the proof in the Supplementary Material, we have $\max_{i,j} |\hat{\Sigma}_{u,ij} - \Sigma_{u,ij}| = o_p(K_u m_u^{-1/2} n^{-1} \log n)$. Therefore, the error condition in Theorem 5 is satisfied as long as $K_u < m_u^{1/2} n^{1-\beta}$, with β a small positive number, which should hold for most cases.

The power of the test for weighted networks is presented in the following theorem.

Theorem 6. *Under the assumptions of Theorem* 4 *and the alternative model* $F_1 \neq F_2$, *let* Z'' *be an* $n \times n$ *matrix with zero diagonals, and for any* $i \neq j$,

$$Z''_{ij} = \frac{P_{1,ij} - P_{2,ij}}{\sqrt{n(m_1{}^{-1}\Sigma_{1,ij} + m_2{}^{-1}\Sigma_{2,ij})}}.$$

Define S_a *and* S_b *as in Theorem* 3, *based on the above* Z''. *If either*

$$(i) \quad an^3 \min_{(i,k,l)\in S_a} (Z''_{ik})^3 + bn^3 \min_{(i,k,l)\in S_b} (Z''_{ik})^3 > 0 \text{ or}$$
$$(ii) \quad -an^3 \max_{(i,k,l)\in S_a} (Z''_{ik})^3 - bn^3 \max_{(i,k,l)\in S_b} (Z''_{ik})^3 > 0$$

is satisfied, then

$$\lim_{n\to\infty} P(|\hat{\theta}| > \mu_{\alpha/2}) = 1, \ \alpha > 0.$$

Again, the proof is omitted, because it is similar to that of Theorem 3.

3.2. Extension to multiple-sample testing

Finally, we consider the case when $S > 2$ groups are present. Assume we observe the symmetric binary adjacency matrices $A_s^{(1)}, \ldots, A_s^{(m_s)}$ that are

generated from a symmetric link probability matrix P_s, that is,

$$A_{s,ij}^{(k)} \sim \text{Bernoulli}(P_{s,ij}),$$

for $s = 1, \ldots, S$, $k = 1, \ldots, m_s$, and $i, j = 1, \ldots, n$. Our goal is to test whether there are any differences in the distributions of the S groups, which is equivalent to testing

$$H_0 : P_1 = P_2 = \cdots = P_S \quad \text{versus} \quad H_1 : P_s \text{ are not all equal.} \tag{3.7}$$

This is analogous to a one-way ANOVA.

We define the pairwise normalized matrices with elements as follows:

$$Z_{ij}^{(s)} = \begin{cases} \frac{\bar{A}_{s,ij} - \bar{A}_{ij}}{\sqrt{n\{(1/m_s - 2/m)P_{s,ij}(1-P_{s,ij}) + (1/m^2)\sum_{s=1}^{S} m_s P_{s,ij}(1-P_{s,ij})\}}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases}, \tag{3.8}$$

where $\bar{A}_s$ is the sample average of the adjacency matrices in group s, as in (2.3), $\bar{A}$ is the overall sample average of all the adjacency matrices,

$$\bar{A} = \frac{1}{m} \sum_{s=1}^{S} \sum_{k=1}^{m_s} A_s^{(k)},$$

m is the total sample size,

$$m = \sum_{s=1}^{S} m_s,$$

and B is defined as in (2.4).

If $\theta^{(s)} = (1/\sqrt{15})\text{Tr}\{(Z^{(s)})^3\}$, then, under the null distribution and appropriate conditions on m_s, Theorem 2 gives

$$\theta^{(s)} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as} \quad n \to \infty,$$

and it follows that

$$\left(\theta^{(s)}\right)^2 \xrightarrow{d} \chi^2(1) \quad \text{as} \quad n \to \infty.$$

Unfortunately, $\theta^{(1)}, \ldots, \theta^{(S)}$ are not independent, so the sum of their squares is not $\chi^2(S)$. However, Ferrari (2019) shows that the sum of dependent χ^2 random variables can be approximated by a gamma distribution. Therefore, we have

$$\theta \equiv \sum_{s=1}^{S} \left(\theta^{(s)}\right)^2 \approx \Gamma\left(\frac{S}{u}, u\right) \quad \text{as} \quad n \to \infty, \tag{3.9}$$

where the scale parameter u is given by

$$u = 2\left(1 + \frac{2\sum_{q\neq r}^{S} \rho_{qr}}{S}\right),$$

with ρ_{qr} the pairwise correlation between the statistics $\left(\theta^{(q)}\right)^2$ and $\left(\theta^{(r)}\right)^2$.

As before, the true link probability matrices P_s are unknown and need to be estimated. We can estimate each P_s as in Section 2.3, and then substitute these estimates into $Z^{(s)}$ in (3.8). Furthermore, although the pairwise correlations ρ_{qr} are not analytically tractable, they can be estimated easily using the Monte Carlo simulations in Algorithm 1, which does not add to the computational complexity. The simulation results in the Supplementary Material (see Section S1.3) demonstrate that using these estimates in the approximation in (3.9) is very accurate, even for small m and n.

Moreover, using this setup, it is possible to follow the same development of Theorem 2 to prove the convergence of the plug-in estimator $\hat{\theta}$ that uses the estimated link probability matrices and estimated pairwise correlations. Similarly, (3.8) can be extended to weighted networks, as in Section 3.1.

4. Simulation Studies

In this section, we demonstrate the performance of our proposed tests by means of a simulation study. For binary networks, we evaluate three plug-in estimators for the link probability matrices (AVG, SBM, and MNBS), and compare the results with those of the test proposed in Ghoshdastidar and von Luxburg (2018). The latter test involves an estimated distance between two network distributions based on the Frobenius measure for binary networks that allows n to go to infinity. To evaluate the approach of conducting a high-dimensional mean test directly on the vectorized networks, we compare our method with that of Chen and Qin (2010), which is based on sum-of-squares-type statistics. We refer to these five tests as SPE-AVG, SPE-SBM, SPE-MNBS, DFRO, and VEC, respectively. We do not include the test proposed in Ginestet et al. (2017) in the comparison, because their results are asymptotic in the sample size with a fixed number of nodes, and the authors expect that their test will lose power in larger dimensions, that is, with more nodes.

We evaluate the test performance by estimating the power when the alternative is true, as well as the null rejection rate (rejection rate under the null). We also vary the number of nodes, $n \in \{100, 200, \ldots, 1000\}$, and the sample sizes, $m_1 = m_2 = m \in \{10, 50\}$. In each example, we set the significance level $\alpha = 0.05$. We follow the procedure described in Algorithm 1, with $Q = 1$, and report the empirical significance level as the average rejection rate on 5,000 separate samples of networks from the underlying distributions. Note that sampling new networks allows us to use $Q = 1$, but the results are similar if we use 5,000 separate samples of networks with $Q > 1$.

Using this design, we consider three types of random graph model for sampling binary networks. In the Supplementary Material, we also include additional simulations for weighted networks, networks from an exponential random graph model that introduce edge dependencies, and multiple-sample testing. The conclusions are as follows. Overall, it appears that SPE-MNBS is the most robust to different network structures and sample sizes. If the networks are drawn from an SBM, then, unsurprisingly, SPE-SBM is suitable. Throughout, SPE-AVG shows significant improvement as the sample size increases. Finally, all three plug-in estimates of the link probability matrices yield superior results for our test compared with those for DFRO and VEC.

Finally, note that in our simulations, VEC always rejects H_0, even for the null settings. Furthermore, VEC is too computationally expensive for networks with many nodes, for example, vectorizing a network with $n = 200$ results in a dimension of almost 20,000. For these reasons, we omit the results of VEC from our figures, and conclude that this approach is inadequate for two-sample testing of nontrivial networks.

4.1. Stochastic block model (SBM)

In the first example, we consider an SBM structure with a block matrix given as

$$P_{\mathrm{SBM}} = \begin{bmatrix} 0.5 + \varepsilon_1 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}, \tag{4.1}$$

where ε_1 depends on our hypothesis. The membership of the ith node is

$$M(i) = I\left(1 \le i \le \left\lfloor \frac{n}{3} \right\rfloor\right) + 2I\left(\left\lfloor \frac{n}{3} \right\rfloor + 1 \le i \le n\right),$$

where $\lfloor \cdot \rfloor$ is the floor operator.

The first group of networks, $\{A_1^{(k)}\}_{k=1}^{m_1}$, is generated from P_{SBM} with $\varepsilon_1 = 0$. In the null setting, the second group of networks, $\{A_2^{(k)}\}_{k=1}^{m_2}$, is also generated from P_{SBM} with $\varepsilon_1 = 0$, whereas $\varepsilon_1 = 1/(5 \log m)$ in the alternative setting. The results are shown in the first row of Figure 1.

To investigate the performance of the tests for sparser networks, we consider the same setting, except now with $\varepsilon_1 = 2/(5 \log m)$ and with the link probability matrix P_{SBM} scaled by a factor $\rho = 10 \log(n)/n$. The corresponding results are shown in the second row of Figure 1.

In the first row of Figure 1, where the networks are dense, SPE-SBM and SPE-MNBS are close to the nominal level $\alpha = 0.05$ under H_0, and both achieve good power under H_1. Furthermore, SPE-AVG is the most powerful under H_1, but its rejection rates are too high under H_0 when $m = 10$. However, this issue is mitigated when we increase the sample size to $m = 50$, even though this makes

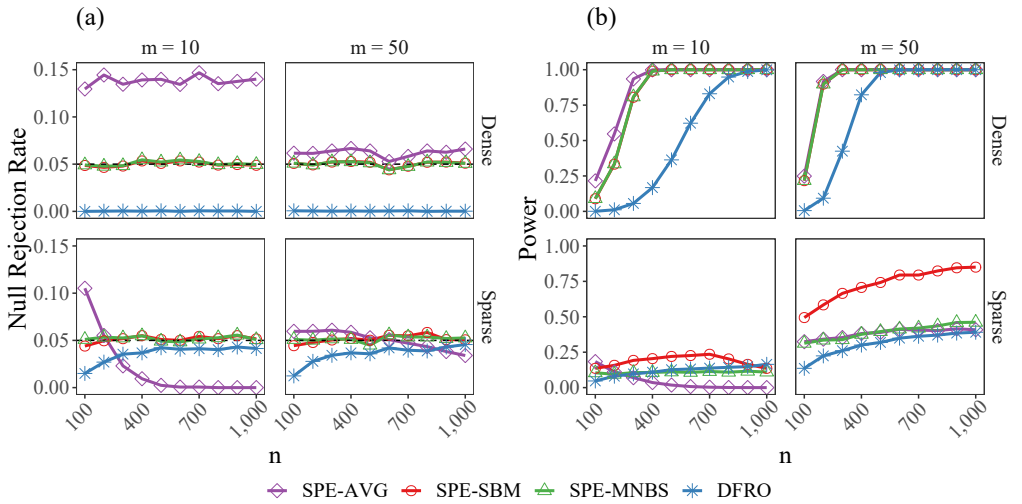

Figure 1. Simulation results for testing networks with an SBM structure for different network orders and sample sizes. The first and second rows represent dense and sparse networks, respectively. (a) Null rejection rate. (b) Power under the alternative.

ε_1 smaller, that is, the underlying SBM structures are more similar. DFRO has a zero rejection rate under H_0, and increases to unit power more slowly than our proposed tests do.

In the sparser settings, shown in the bottom row, similar results hold for SPE-SBM and SPE-MNBS, except for small $m = 10$, which is also difficult for the other methods. Moreover, DFRO performs comparably with SPE-SBM and SPE-MNBS, and SPE-AVG suffers a low rejection rate under H_0 with increasing n.

4.2. Graphon

In the second example, we focus on graphon structures, which have found applications in hierarchical clustering (Eldridge, Belkin and Wang (2016)) and link probability estimation (Zhang, Levina and Zhu (2017)). A graphon f is defined as follows.

Definition 1 (Graphon (Zhang, Levina and Zhu (2017)))**.** For any network with a link probability matrix P and number of nodes n, there exists a function $f : [0,1] \times [0,1] \to [0,1]$ and a set of i.i.d. random variables $\xi_i \sim \text{Uniform}[0,1]$, such that

$$P_{ij} = f(\xi_i, \xi_j),$$

with $i, j = 1, \ldots, n$.

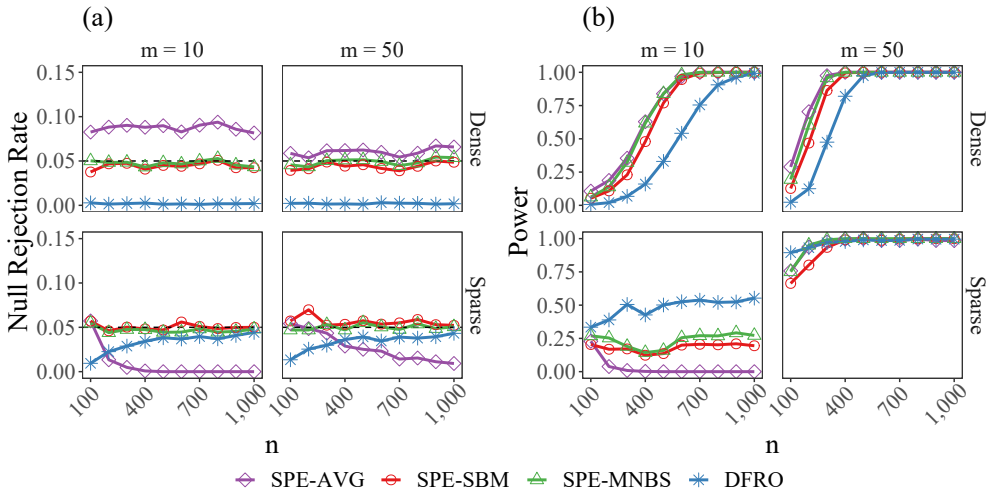

Figure 2. Simulation results for testing networks with a graphon structure. The first and second rows represent dense and sparse networks, respectively. (a) Null rejection rate. (b) Power under the alternative.

In our simulation, we consider a graphon structure from Zhang, Levina and Zhu (2017), in which

$$f(u,v) = \frac{u^2+v^2}{3\cos\{1/(u^2+v^2)\}} + 0.15.$$

We generate $\{A_1^{(k)}\}_{k=1}^{m_1}$ from the probability matrix P_1 according to f. For the second group of networks, under the null hypothesis, we again sample from f to generate $\{A_2^{(k)}\}_{k=1}^{m_2}$. Under the alternative hypothesis, we first randomly choose a subset $S \subset \{1,2,\ldots,n\}$, with $|S| = \lfloor n/10 \rfloor$, and then generate $\{A_2^{(k)}\}_{k=1}^{m_2}$ from P_2, with $P_{2,ij} = P_{1,ij} - \varepsilon_2$, where

$$\varepsilon_2 = \begin{cases} \frac{1}{8\log m} & \text{if } i,j \in S \\ 0 & \text{if } i,j \notin S \end{cases}.$$

The results are presented in the first row of Figure 2. As before, we set $\varepsilon_2 = 2/(5\log m)$, for $i,j \in S$, and scale the link probability matrices P_1 and P_2 by $\rho = 12\log n/n$ to yield sparser networks. The results are shown in the second row of Figure 2.

Figure 2 shows that SPE-MNBS outperforms the other tests in terms of both the null rejection rate and power, except for small $m = 10$ and a sparse structure. Furthermore, SPE-SBM exhibits a lower rejection rate than the nominal level in the dense case, which suggests that the method is more sensitive to network topologies that deviate from an SBM. SPE-AVG and DFRO behave similarly to those in the first example, as we continue to see subpar performance, especially

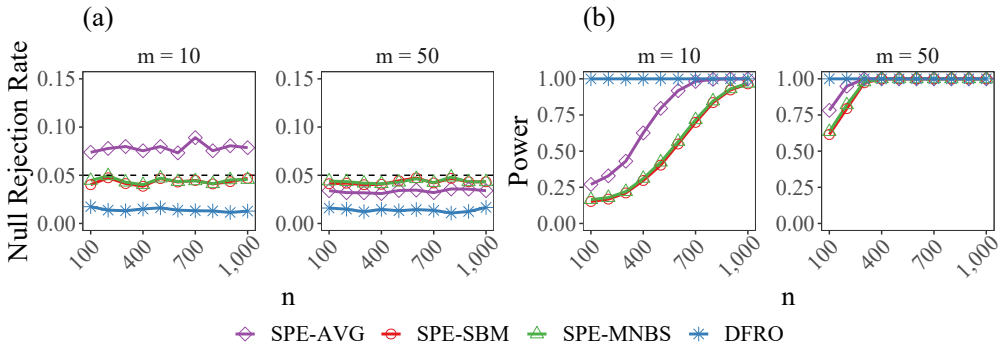

Figure 3. Simulation results for testing networks with a correlated ER structure. (a) Null rejection rate. (b) Power under the alternative.

for small m.

4.3. Correlated Erdős–Rényi model

In the third experiment, we study the robustness of the four tests to dependency. For this, we consider the correlated Erdős–Rényi (ER) model of Pedarsani and Grossglauser (2011). We begin by sampling two independent ER networks, $A_1 \sim ER(n, p_1)$ and $A_2 \sim ER(n, p_2)$. We generate $\{A_1^{(k)}\}_{k=1}^{m_1}$ with a parameter ε_3 as follows:

$$A_{1,ij}^{(k)} \sim \begin{cases} \text{Bernoulli}(\varepsilon_3) & \text{if } A_{1,ij} = 1 \\ 0 & \text{if } A_{1,ij} = 0 \end{cases}.$$

This yields m_1 networks that are marginally $ER(n, p_1\varepsilon_3)$, but whose edge sets are correlated. We similarly generate $\{A_2^{(k)}\}_{k=1}^{m_2}$ conditional on A_2 with parameter ε_4. We set $\varepsilon_3 = \varepsilon_4 = 0.8$ and $p_1 = 0.9$. Under the null hypothesis, we set $p_2 = p_1 = 0.9$, and $p_2 = 0.83$ under the alternative hypothesis. The results are shown in Figure 3.

DFRO exhibits consistently high power in the alternative setting for the entire range of n, which is matched only by our tests as n increases, with SPE-AVG outperforming both SPE-SBM and SPE-MNBS. However, the rejection rate under the null is below the nominal level for DFRO, whereas both SPE-SBM and SPE-MNBS are very close to $\alpha = 0.05$. SPE-AVG has a higher rejection rate than expected when the sample size is $m = 10$, but this improves when $m = 50$. Overall, it appears that SPE-SBM and SPE-MNBS are robust to the independence violation when n is large.

5. Real-Data Examples

In this section, we apply our tests to three real data sets representing three settings of interest within the biological research community: the StarPlus, COBRE, and MB data sets. The first two are networks constructed from fMRI data that represent two distinct streams of fMRI usage, the former being task based and the latter being a case/control study. The third data set is derived from microbial measurements, an area in which network-based representations have recently emerged as a popular technique for studying the bacteria present within a microbiome (Layeghifard, Hwang and Guttman (2017)). A description of the data sets can be found in the Supplementary Material.

In all three cases, the networks are weighted. Therefore, we present results from our tests for weighted networks in Section 3.1. To understand the performance of our tests for binary networks from Section 2.1, we also present the results as a function of thresholding the weights to binarize the networks (as often occurs in practice).

5.1. Results for weighted tests

We begin by applying our tests for weighted networks from Section 3.1. We also include the method of Yuan and Wen (2021), which we refer to as WRG. We test whether the groups defined by their respective labels, that is, picture/sentence, schizophrenic/control, preterm/term, are different. To do so, we specify a null hypothesis that states that the underlying random distributions are equal against the alternative that states that they are different. We refer to this as the "alternative setting", because the two samples differ with respect to their group label.

As is, we cannot apply WRG directly, because it requires that the sample sizes for both groups be the same, which is not true of the COBRE and MB data sets. We address this by following the authors original solution, which is to randomly sample m_2 networks from group one (assuming $m_1 > m_2$), and then to compare this subgroup with group two.

For $\alpha = 0.05$ and $Q = 1000$, we find that for the StarPlus networks, SPE-AVG, SPE-SBM, and WRG correctly reject the null with rejection rates of 1, 0.726, and 1, respectively. This is consistent with the findings of previous research on distinguishing the cognitive states of looking at a picture and a sentence (Mitchell et al. (2004); Wang, Hutchinson and Mitchell (2003); Mitchell et al. (2003)). For the COBRE and MB data sets, we find a rejection rate of one for both SPE-AVG and SPE-SBM. On the other hand, WRG rejects the null with rates of 1 and 0.749 for the COBRE and MB data sets, respectively.

Next, we perform an in silico experiment using the real data by subsampling within one of the classes. We refer to this as the "null setting." The rationale for this setup is that we do not actually know whether the groups are generated by

different underlying distributions, for example, one for schizophrenic and another for non-schizophrenic. Therefore, we want to check whether the null rejection rate is close to the nominal level in an experiment in which all of the networks are from the same group.

To do so, we test the entire NetP, non-schizophrenic, and term delivery groups against a subsample (with half of the original sample size) of the same group for the StarPlus, COBRE, and MB data sets. For WRG, we test two subsamples of the two groups, both with half of the original sample size.

After 1,000 random subsamples of the networks and $Q = 1$ for each subsample, for the StarPlus networks, SPE-SBM fails to reject the null hypothesis, with a rejection rates of 0.006, which is expected, because the samples are drawn from the same population. However, SPE-AVG and WRG reject the null with inflated rates of 0.12 and 0.873, respectively. For the COBRE networks, we obtain null rejection rates of 0.763, 0.668, and 1 for SPE-AVG, SPE-SBM, and WRG, respectively. The null rejection rates improve for the MB networks, with 0.655, 0.489, and 0.956 for SPE-AVG, SPE-SBM, and WRG, respectively. Although SPE-AVG and SPE-SBM outperform WRG, the null rejection rates are still very inflated compared with the nominal $\alpha = 0.05$.

We speculate that this is happening because, even within one class, there is a lot of variation. That is, one subsample of brain networks with schizophrenia may look very different to another sample of brain networks with schizophrenia, because we are not controlling for potential factors such as age and sex. We refer to this issue as having too much heterogeneity within a class. This heterogeneity can lead to inflated null rejection rates, because the underlying distributions of the two samples are different, but the difference is not the one we are trying to isolate.

5.2. Results for binary tests

Because the results for the weighted tests showed inflated rejection rates in our simulated null setting, there is reason to believe that the networks are too heterogeneous within each class. Furthermore, many of the weights could represent spurious correlations. Therefore, this is a setting in which binarizing the weights could improve the signal-to-noise ratio. This idea is related to a common problem in the neuroscience literature related to the issue of sensitivity to thresholding edges (Ginestet, Fournel and Simmons (2014); Garrison et al. (2015)).

To evaluate this, we apply the binary tests from Section 2.1 by binarizing the weights, which are all correlation values in $[-1, 1]$, based on thresholding their magnitude. Specifically, we set the adjacency matrix entries to one when the absolute values of the corresponding weights are larger than the threshold, and zero otherwise. This threshold relates directly to the density of the networks. Note that WRG is for weighted networks, and is therefore excluded. Using the

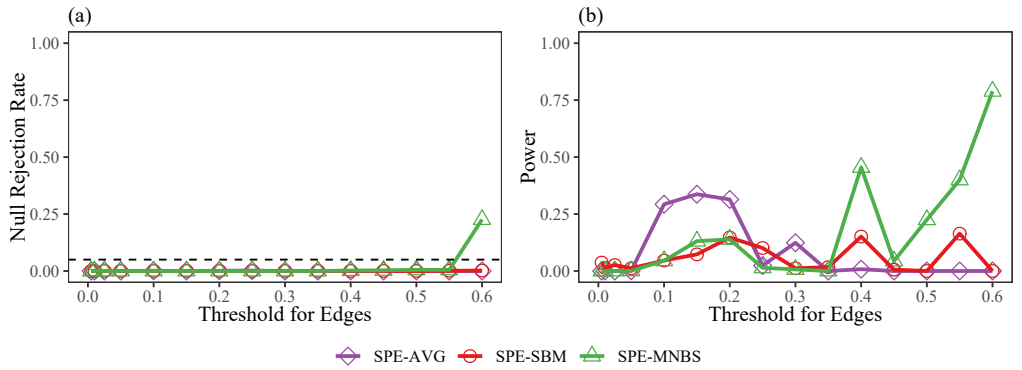

Figure 4. Figures (a) and (b) show the null rejection rate and power, respectively, for different thresholds for binarizing the StarPlus networks.

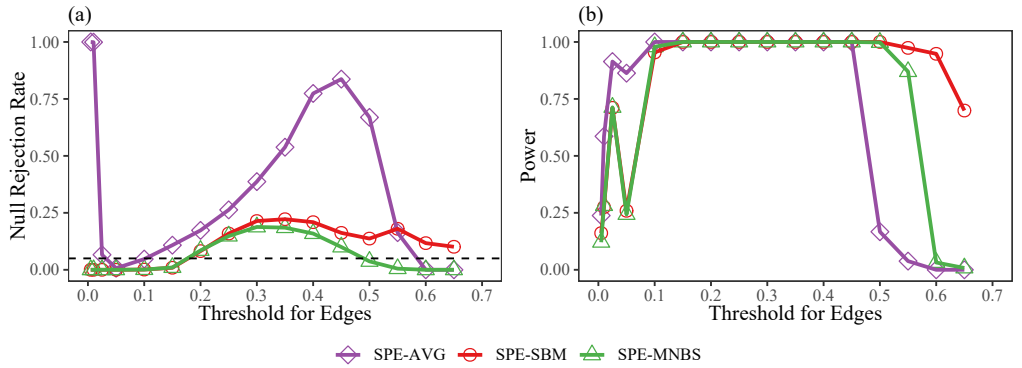

Figure 5. Figures (a) and (b) show the null rejection rate and power, respectively, for different thresholds for binarizing the COBRE networks.

same procedures as in Section 5.1, the results are given in Figures 4–6. The dashed lines for the null rejection rate in these figures all indicate the nominal level of 0.05.

The plots illustrate the trade-off between the false positive rate in our null setting and the true positive rate in our alternative setting, which are both functions of the threshold. As the threshold for an edge increases, the network becomes more sparse, resulting in a higher rejection rate in our null setting. For thresholds above 0.6, some of the networks become too sparse, even resulting in some null graphs. On the other hand, for a low threshold, there is less power to detect a difference in our alternative setting. Such curves as a function of the threshold could provide practitioners with a way to understand the signal-to-noise ratio of their edge weights.

For the COBRE and MB networks, we have high power for a wide range of threshold values, which is consistent with our findings using the weighted

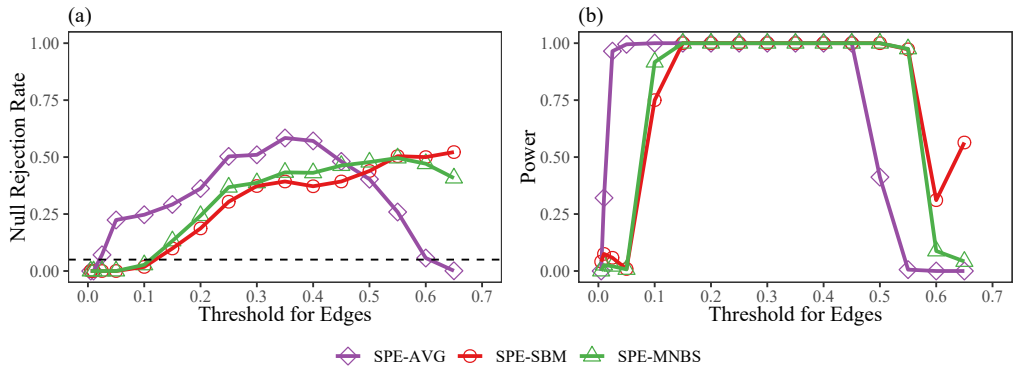

Figure 6. Figures (a) and (b) show the null rejection rate and power, respectively, for different thresholds for binarizing the MB networks.

networks directly. Moreover, we find a low rejection rate in our null setting, especially for a threshold of 0.1, which seems to provide the best trade-off. This suggests that the signal-to-noise ratio in the weights is too low, which can be mitigated by using thresholding. For the StarPlus networks, a threshold between 0.1 and 0.2 seems to provide the best balance between signal and noise for SPE-AVG, whereas 0.4 is better for SPE-SBM and SPE-MNBS.

6. Conclusion

In this work, we have proposed new spectral-based statistics for hypothesis testing of populations of networks that applies to both binary and weighted networks under a very general framework. The test statistics are simple, computationally friendly, and supported theoretically by our derivations of the limiting null distribution and the asymptotic power guarantees. We have demonstrated our method using a simulation study and a real-data analysis. In future work, we will focus on spectral-based methods for studying inference problems for networks with additional constraints or structures, such as directed networks.

Supplementary Material

The online Supplementary Material contains two additional simulations (for weighted networks and multiple-sample testing), a description of the three datasets in the applications, background on spectral theory, and the proofs for the results in the main text.

Acknowkedgments

The work of Li Chen was supported by the China Scholarship Council under Grant 201806240032 and the Fundamental Research Funds for the Central Universities, Southwest Minzu University under grant 2021NQNCZ02. Lizhen Lin acknowledges the generous support from NSF grants IIS 1663870, DMS Career 1654579 and a DARPA grant N66001-17-1-4041. The work of Jie Zhou was supported in part by the National Natural Science Foundation of China under grant 11871357 and the Sichuan Science and Technology Program under grant 2019YJ0122.

References

Arroyo Relión, J. D., Kessler, D., Levina, E. and Taylor, S. F. (2019). Network classification with applications to brain connectomics. *The Annals of Applied Statistics* **13**, 1648–1677.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer, New York; London.

Banerjee, D. and Ma, Z. (2017). Optimal hypothesis testing for stochastic block models with growing degrees. *arXiv:1705.05305*.

Bernstein, S. (1946). *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow.

Bhattacharya, B. B., Das, S. and Mukherjee, S. (2020). Motif estimation via subgraph sampling: The fourth moment phenomenon. *arXiv:2011.03026*.

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

Cai, T. T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**, 349–372.

Chen, L., Lin, L. and Zhou, J. (2020). A hypothesis testing for large weighted networks with applications to functional neuroimaging data. *IEEE Access* **8**, 191815–191825.

Chen, L., Zhou, J. and Lin, L. (2021). Hypothesis testing for populations of networks. *Communications in Statistics-Theory and Methods* **52**, 3661–3684.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

Dong, Z., Wang, S. and Liu, Q. (2020). Spectral based hypothesis testing for community detection in complex networks. *Information Sciences* **512**, 1360–1371.

Eldridge, J., Belkin, M. and Wang, Y. (2016). Graphons, mergeons, and so on! *Advances in Neural Information Processing Systems* **29**.

Ferrari, A. (2019). A note on sum and difference of correlated chi-squared variables. *arXiv: 1906.09982*.

Gao, C. and Lafferty, J. (2017). Testing for global network structure using small subgraph statistics. *arXiv:1710.00862*.

Garrison, K. A., Scheinost, D., Finn, E. S., Shen, X. and Constable, R. T. (2015). The (in)stability of functional brain network measures across thresholds. *NeuroImage* **118**, 651–661.

Ghoshdastidar, D., Gutzeit, M., Carpentier, A. and Luxburg, U. V. (2020). Two-sample hypothesis testing for inhomogeneous random graphs. *The Annals of Statistics* **48**, 2208–2229.

Ghoshdastidar, D. and von Luxburg, U. (2018). Practical methods for graph two-sample testing. *arXiv:1811.12752*.

Ginestet, C. E., Fournel, A. P. and Simmons, A. (2014). Statistical network analysis for functional MRI: Summary networks and group comparisons. *Frontiers in Computational Neuroscience* **8**.

Ginestet, C. E., Li, J., Balachandran, P., Rosenberg, S. and Kolaczyk, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics* **11**, 725–750.

Jin, J., Ke, Z. T. and Luo, S. (2021). Optimal adaptivity of signed-polygon statistics for network testing **49**, 3408–3433.

Josephs, N., Lin, L., Rosenberg, S. and Kolaczyk, E. D. (2020). Bayesian classification, anomaly detection, and survival analysis using network inputs with application to the microbiome. *arXiv:2004.04765*.

Kolaczyk, E. D., Lin, L., Rosenberg, S., Walters, J. and Xu, J. (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics* **48**, 514–538.

Layeghifard, M., Hwang, D. M. and Guttman, D. S. (2017). Disentangling interactions in the microbiome: A network perspective. *Trends in Microbiology* **25**, 217–228.

Lovász, L. (2012). *Large Networks and Graph Limits*. American Mathematical Society, Providence.

Mitchell, T. M., Hutchinson, R., Just, M. A., Niculescu, R. S., Pereira, F. and Wang, X. (2003). Classifying instantaneous cognitive states from fMRI data. In *AMIA Annual Symposium Proceedings*, 465–469.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. et al. (2004). Learning to decode cognitive states from brain images. *Machine Learning* **57**, 145–175.

Mukherjee, S. S., Sarkar, P. and Lin, L. (2017). On clustering network-valued data. In *Advances in Neural Information Processing Systems*, 7071–7081. Long Beach.

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* **14**, 849–856. Cambridge.

Pedarsani, P. and Grossglauser, M. (2011). On the privacy of anonymized networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1235–1243.

Tang, M., Athreya, A., Sussman, D. L., Lyzinski, V. and Priebe, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics* **26**, 344–354.

Wang, X., Hutchinson, R. and Mitchell, T. M. (2003). Training fMRI classifiers to detect cognitive states across multiple human subjects. In *Advances in Neural Information Processing Systems* **16**, 709–716.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103**, 609–624.

Yuan, M. and Wen, Q. (2021). A practical two-sample test for weighted random graphs. *Journal of Applied Statistics* **50**, 495–511.

Zhang, Y., Levina, E. and Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika* **104**, 771–783.

Zhang, Y. and Xia, D. (2020). Edgeworth expansions for network moments. *arXiv:2004.06615*.

Zhao, Z., Chen, L. and Lin, L. (2019). Change-point detection in dynamic networks via graphon estimation. *arXiv:1908.01823*.

Li Chen
Department of Mathematics, Southwest Minzu University, Chengdu, China.
E-mail: lchen@swun.edu.cn
Nathaniel Josephs
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
E-mail: nathaniel.josephs@ncsu.edu
Lizhen Lin
Department of Mathematics, University of Maryland, College Park, MD 20742, USA.
E-mail: lizhen01@umd.edu
Jie Zhou
College of Mathematics, Sichuan University, Chengdu, China.
E-mail: jzhou@scu.edu.cn
Eric D. Kolaczyk
Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada.
E-mail: eric.kolaczyk@mcgill.ca

(Received September 2021; accepted April 2022)