Statistica Sinica Preprint No: SS-2021-0306	
Title	A Spectral-Based Framework for Hypothesis Testing in Populations of Networks
Manuscript ID	SS-2021-0306
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0306
Complete List of Authors	Li Chen, Nathaniel Josephs, Lizhen Lin, Jie Zhou and Eric D. Kolaczyk
Corresponding Authors	Nathaniel Josephs
E-mails	nathaniel.josephs@yale.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

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

Li Chen[1], Nathaniel Josephs[2], Lizhen Lin[3],

Jie Zhou[4], and Eric D. Kolaczyk[5]

Southwest Minzu University & Sichuan University, Yale University, The University of Notre Dame, Sichuan University, Boston University

Abstract: In this paper, 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. The asymptotic power guarantee of the test is also provided. The proper interplay between the number of networks and the number of nodes for each network is explored in characterizing the theoretical properties of the proposed testing statistic. Our test is applicable to both binary and weighted networks, operates under a very general framework where the networks are allowed to be large and sparse, and can be extended to multiple-sample testing. We provide an extensive simulation study to demonstrate the superior performance of our tests over the existing methods and apply our tests to three real datasets.

Key words and phrases: Populations of networks, hypothesis testing, 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. The statistical network analysis literature has been largely focused on proposing models and algorithms for analyzing a single network. However, the increasing prevalence of multiple network datasets, in which the network is the fundamental data object, along with the need to extract useful scientific information from them, have motivated the demand for developing statistical methods of inference for populations of networks.

For example, in the brain network data we examine in Section 5, one may be interested in testing whether the brain network structure from a group of individuals with schizophrenia is different 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 could provide a notion either of averaging networks or of clustering networks into different groups (Mukherjee et al., 2017). All of these cases are inference tasks for one or two samples of network objects, both of which have been

recently explored in the literature.

In Ginestet et al. (2017), the authors consider two-sample testing for networks with applications to functional neuroimaging. Very recently, this work is extended in Kolaczyk et al. (2020) through a geometric and statistical framework for inference on populations of unlabeled networks 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, e.g., Arroyo Relión et al. (2019) and Josephs et al. (2020), with the former considering network classification in neuroimaging and the latter employing Bayesian methods for classification, anomaly detection, and survival analysis.

Herein, 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 two random dot product networks ($m = 2$) defined on different vertex sets are generated from the model or not. Ghoshdastidar et al. (2020) study two-sample problems from a minimax perspective on testing whether two samples of binary networks of n

nodes are generated from the same link probability matrix against an alternative that says the two link probability matrices are ρ apart with respect to some matrix norm. Their work focuses on the 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) further apply the same test statistic and prove it converges to a normal distribution asymptotically. Recently, Yuan and Wen (2021) modifies the test statistic in Ghoshdastidar and von Luxburg (2018) and proposes 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 apply a two-sample, high-dimensional mean test. This strategy has been widely studied in the literature (Chen and Qin, 2010; Cai et al., 2014; Xu et al., 2016). Although this approach is model-free, one drawback is the potential loss of information in the conversion process, which essentially ignores the interconnectedness that defines network data. We return to this discussion in Section 4.

Compared to much of the work in the literature, such as Ginestet et al. (2017), in which the number of nodes is fixed, we consider a general framework which 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 utilize 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. Since we also want to understand the limiting behavior as the sample size increases, we explore the proper interplay 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 testing statistics. These statistics are conceptually simple and computational friendly and we provide an extensive simulation study under various models to demonstrate the superior performance of our test over the existing methods. In almost all the cases examined in our study, 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 datasets based on both weighted and binary networks.

The idea of applying a spectral method based on random matrix theory to network data is a natural one, as the network data (e.g., the adjacency or Laplacian matrix) can be naturally viewed as a random matrix. Our method is motivated by Dong et al. (2020), which proposes a spectral-based

hypothesis test for testing the community structure within a single network. In that work, the authors prove 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 community structure in a single network versus the null Erdös-Rényi model. In our work, we extend the statistic to be able to test the difference between arbitrary network models that 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 et al. (2020) and Chen et al. (2021). Compared to those two works, our spectral-based test has asymptotic standard normal distribution with much faster convergence rate under the null compared to the slow convergence of a test that has a Tracy-Widom law. Furthermore, our testing statistics require much milder conditions for the theoretical performance guarantees: an error estimate of the link probability estimates with $o_p(1)$ is needed compared to an error condition of $o_p(n^{-2/3})$ required in Chen et al. (2020).

The remainder of the paper is organized as follows. In Section 2, we describe our proposed spectral-based testing statistic and derive its asymptotic null distribution as well as an asymptotic power result. We extend

our test for weighted networks and multiple-sample testing in Section 3. Results of extensive simulation studies are reported in Section 4 and analysis on three real network datasets are given in Section 5. We conclude in Section 6 with a few final remarks and possible future directions of this work.

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. We assume one observes 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, i.e.

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

for $k = 1, 2, \ldots, m_1,\ i, j = 1, 2, \ldots, n$. Similarly, one observes 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, i.e., $A^{(k)}_{u,ii} = 0$ 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 or not, which is equivalent to testing

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

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

2.1 New spectral test for binary networks

Given two samples of networks $\{A^{(k)}_1\}_{k=1}^{m_1}$ and $\{A^{(k)}_2\}_{k=1}^{m_2}$ sampled from the link probability matrices P_1 and P_2, respectively, we introduce the normalized matrix with elements as follows:

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

where $\bar{A}_u$ is the sample average of 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

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

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 et al. (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 $\theta = \frac{1}{\sqrt{15}} Tr(Z^3)$, 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, the overview of the argument is as follows. First, one can see that 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 the remainder of the proof proceeds in two steps. 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. Then, the mean and the variance are obtained from Dong et al. (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 could obtain a p-value using $2P\big(\theta > |\theta_{obs}^{(Q=1)}|\big)$ as in Bickel and Sarkar (2016) and Dong et al. (2020), where $\theta_{obs}^{(Q=1)}$ is the sample test statistic and θ follows the null distribution of the testing statistic. In this case, though, 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

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

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 $\mathcal{N}(0, 1)$.

to reduce the noise induced by B. The analogous p-value estimate combining these Monte Carlo test statistics would be $2P\big(\theta > |\bar{\theta}_{obs}|\big)$, where $\bar{\theta}_{obs} = \frac{1}{Q}\sum_{q=1}^{Q} \theta_{obs}^{(q)}$.

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

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

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 some appropriate estimates of P_1 and P_2 with the hope that the plug-in estimator for the test statistic retains 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

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

The resulting test statistic is thus

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

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 asymp-*

totic distribution of the scaled test statistic $\hat{\theta} = \frac{1}{\sqrt{15}} Tr(\hat{Z}^3)$*:*

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

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 in the right 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 specifically becomes a network test through the estimation of the link probability matrices. Here, we propose three methods that satisfy the conditions in Theorem 2, which require the sample sizes of the observed networks to 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 the (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 also confirmed empirically by our extensive simulation studies in which SPE-AVG typically yields inferior performance compared to the next methods we present.

Another possible average estimator of $P_{u,ij}$ is based on the stochastic block model (SBM). The key idea is to approximate any 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 by community algorithms such as the method proposed in Ng et al. (2002). After the membership vector has been estimated, we can simply approximate $P_{u,ij}$ by the sample mean of all the entries in the submatrix over all $A_u^{(k)}$, $k = 1, 2, \ldots, m_u$, restricted to the corresponding block consisting of the communities of i and j. We refer to this test method based on SBM as SPE-SBM. Assuming the true community number is K_u, then the estimation error satisfies $\max_{i,j} |\hat{P}_{u,ij} - P_{u,ij}| = o_p(K_u m_u^{-1/2} n^{-1} \log n)$. It can be seen that 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 method of estimation based on the modified neighborhood smoothing (MNBS) proposed in Zhao et al. (2019). The idea is to perform neighborhood smoothing to the matrix $\bar{A}$, which is the weighted average of m networks and the smoothing procedure is applied to a shrunken neighborhood size. This results in a better bias-variance tradeoff 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 the adjacent matrix of a single network, and with a shrunken neighborhood size. This has the effect of reducing the variance due to the multiple networks available in the each sample. From Lemma 9.3 in Zhao et al. (2019), the size of 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. We also note that SPE-MNBS puts no structure conditions on the networks. Therefore, we expect SPE-MNBS 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 $i \neq j$,*

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

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 of the following conditions are satisfied,

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

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

then

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

The details of the proof are 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). The conditions (i) and (ii) above 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 large enough. For a better understanding of this, consider the case when $P_{1,ij} \geq P_{2,ij}$ for all i and j, i.e. $Z''_{ij} \geq 0$, 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 instead of the method of network moments or 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 et al. (2021); Zhang and Xia (2020); Bhattacharya et al. (2020).

3. Extending our test to other settings

In this section, we extend our test to be used for weighted networks, as well as for multiple samples in a manner analogous to 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, i.e., $A_{u,ii}^{(k)} = 0$ 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 0 since $A_{u,ii}^{(k)} = 0$. 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) by $\Sigma_{u,ij}$ and $\hat{\Sigma}_{u,ij}$, respectively. Just as in Section 2.3, estimates $\hat{\Sigma}_{u,ij}$ can be obtained using various methods, which will be 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

$$H_0 : F_1 = F_2 \text{ against } H_1 : F_1 \neq F_2 \ . \tag{3.10}$$

We define the normalized matrix Z as

$$Z_{ij} = \begin{cases} \frac{\bar{A}_{1,ij} - \bar{A}_{2,ij}}{\sqrt{n\left(\frac{1}{m_1}\Sigma_{1,ij} + \frac{1}{m_2}\Sigma_{2,ij}\right)}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases}, \tag{3.11}$$

where B is defined the same as in (2.4). Then the asymptotic distribution of $\theta = \frac{1}{\sqrt{15}}\mathrm{Tr}(Z^3)$ 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 (3.11). Assume 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 $\theta = \frac{1}{\sqrt{15}} Tr(Z^3)$, we have*

$$\theta \xrightarrow{d} \mathcal{N}(0, 1) \quad as \quad n \to \infty \ . \tag{3.12}$$

The proof is omitted since 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 of (3.10), 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, while 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, the covariance matrices Σ_1 and Σ_2 need to be estimated, with some conditions assumed 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 (3.11) can be written with entries as

$$\hat{Z}_{ij} = \begin{cases} \frac{\bar{A}_{1,ij} - \bar{A}_{2,ij}}{\sqrt{n\left(\frac{1}{m_1}\hat{\Sigma}_{1,ij} + \frac{1}{m_2}\hat{\Sigma}_{2,ij}\right)}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases} . \tag{3.13}$$

Therefore, our test statistic is

$$\hat{\theta} = \frac{1}{\sqrt{15}} \mathrm{Tr}(\hat{Z}^3) \ . \tag{3.14}$$

Then we have the following limiting law.

Theorem 5. *Under the two-sample framework of weighted networks, let $\hat{Z}$ be given in* (3.13)*. 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 $\hat{\theta} = \frac{1}{\sqrt{15}} Tr(\hat{Z}^3)$:*

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

The proof is similar to that of Theorem 2, so we only include the key

differences in the supplementary material, as the remainder of the proof can be completed straightforwardly.

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

$$\max_{i,j} |\hat{\Sigma}_{u,ij} - \Sigma_{u,ij}| = o_p(m_u^{-1/2} \log m_u) \ . \tag{3.15}$$

The proof of (3.15) can be found 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-size cases.

The second estimate of $\Sigma_{u,ij}$ is obtained similarly as 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. With a similar argument as the proof in the supplementary material, one has $\max_{i,j} |\hat{\Sigma}_{u,ij} - \Sigma_{u,ij}| = o_p(K_u m_u^{-1/2} n^{-1} \log n)$. 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 $i \neq j$,*

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

Define S_a and S_b as in Theorem 3 based on the above Z''. If either of the following conditions are satisfied,

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

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

then

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

Again, the proof is omitted as 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 one observes the symmetric binary adjacency matrices $A_s^{(1)}, \ldots, A_s^{(m_s)}$ that are

generated from a symmetric link probability matrix P_s, i.e.

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

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

$$H_0 : P_1 = P_2 = \cdots = P_S \ \text{ against } \ H_1 : P_s \text{ are not all equal } . \tag{3.16}$$

This is analogous to one-way ANOVA.

We define the pairwise normalized matrices with elements as follows:

$$Z_{ij}^{(s)} = \begin{cases} \dfrac{\bar{A}_{s,ij} - \bar{A}_{ij}}{\sqrt{n\left(\left(\frac{1}{m_s} - \frac{2}{m}\right) P_{s,ij}(1-P_{s,ij}) + \frac{1}{m^2}\sum_{s=1}^{S} m_s P_{s,ij}(1-P_{s,ij})\right)}} & \text{if } i \neq j \\ B_{ij} & \text{if } i = j \end{cases} , \tag{3.17}$$

where $\bar{A}_s$ is the sample average of 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 $\theta^{(s)} = \frac{1}{\sqrt{15}}\mathrm{Tr}\big((Z^{(s)})^3\big)$, 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

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

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

$$\theta \equiv \sum_{s=1}^{S} \big(\theta^{(s)}\big)^2 \approx \Gamma\Big(\frac{S}{u}, u\Big) \quad \text{as} \quad n \to \infty \ , \tag{3.18}$$

where the scale parameter u is given by

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

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

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 plug in these estimates to $Z^{(s)}$ in (3.17). Furthermore, although the pairwise

correlations ρ_{qr} are not analytically tractable, they can easily be estimated 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 (3.18) is very accurate even for small m and n.

Moreover, with this setup, it is possible to follow the same development of Theorem 2 in proving convergence of the plug-in estimator $\hat{\theta}$ that uses the estimated link probability matrices and estimated pairwise correlations. Similarly, (3.17) can be extended to weighted networks as in Section 3.1.

4. Simulation studies

In this section, we illustrate the performance of our proposed tests through extensive simulation study. For binary networks, we evaluate three plug-in estimators for the link probability matrices – AVG, SBM, and MNBS – and compare the results to the test proposed in Ghoshdastidar and von Luxburg (2018), which 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 Chen and Qin (2010), which is based on the 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) for comparison because their results are asymptotic in the sample size with a fixed number of nodes and the authors note that their test is expected to decrease power in larger dimensions, i.e. 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 to $\alpha = 0.05$. We follow the procedure described in Algorithm 1 with $Q = 1$ and report the empirical significance level as the average rejection rate on 5000 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 5000 separate samples of networks with $Q > 1$.

With this design, we consider three types of random graph model for sampling binary networks. In the supplementary material, we also include additional simulations for i) weighted networks, ii) networks from an exponential random graph model that introduce edge dependencies, and iii) multiple-sample testing. The results are detailed in the remainder of this

section, but 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, we see that 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 to DFRO and VEC.

Finally, we note that in our simulations, VEC always rejected H_0 even for the null settings. Furthermore, it was too computationally expensive for networks with many nodes – e.g., 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 non-trivial networks.

4.1 Stochastic block model (SBM)

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

$$P_{\text{SBM}} = \begin{bmatrix} 0.5 + \varepsilon_1 & 0.25 \\ 0.25 & 0.5 \end{bmatrix}, \tag{4.19}$$

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

$$M(i) = I(1 \leq i \leq \lfloor n/3 \rfloor) + 2I(\lfloor n/3 \rfloor + 1 \leq i \leq n) \ ,$$

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, the same setting as above is considered, 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.

We see from the first row of Figure 1, where the networks are dense, that 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. We also observe that 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 ε_1 smaller, i.e. more similar underlying SBM structures. For DFRO, it has zero rejection rate under H_0 and increases to unit power more slowly than our proposed

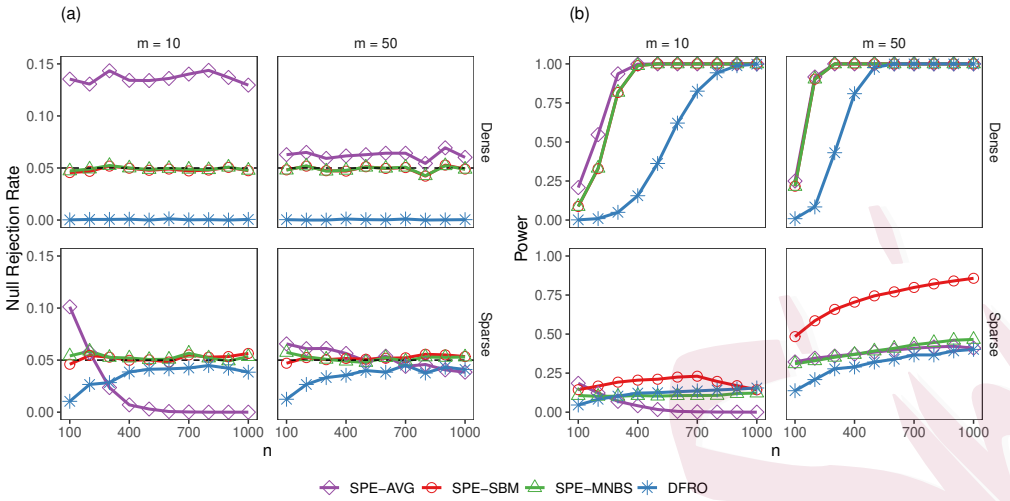

Figure 1: Simulation results for testing networks with an SBM structure for different network orders and sample sizes. The first and the second rows are for dense and sparse networks, respectively. (a) Null rejection rate. (b) Power under the alternative.

tests.

In the sparser settings on the bottom row, similar results hold for SPE-SBM and SPE-MNBS except the case of small $m = 10$, which is also hard for other methods. Moreover, DFRO performs more comparably to SPE-SBM and SPE-MNBS, while SPE-AVG suffers low rejection rate under H_0 with increasing n.

4.2 Graphon

In the second example, we focus on graphon structure which has found applications in hierarchical clustering (**?**) and link probability estimation

(Zhang et al., 2017). A gaphon f is defined as follows.

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

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

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

In our simulation, we consider a graphon structure from Zhang et al. (2017) in which

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

We generate $\{A_1^{(k)}\}_{k=1}^{m_1}$ from 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 $|S| = \lfloor n/10 \rfloor$, then generate $\{A_2^{(k)}\}_{k=1}^{m_2}$ from P_2 with $P_{2,ij} = P_{1,ij} - \varepsilon_2$, where

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

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.

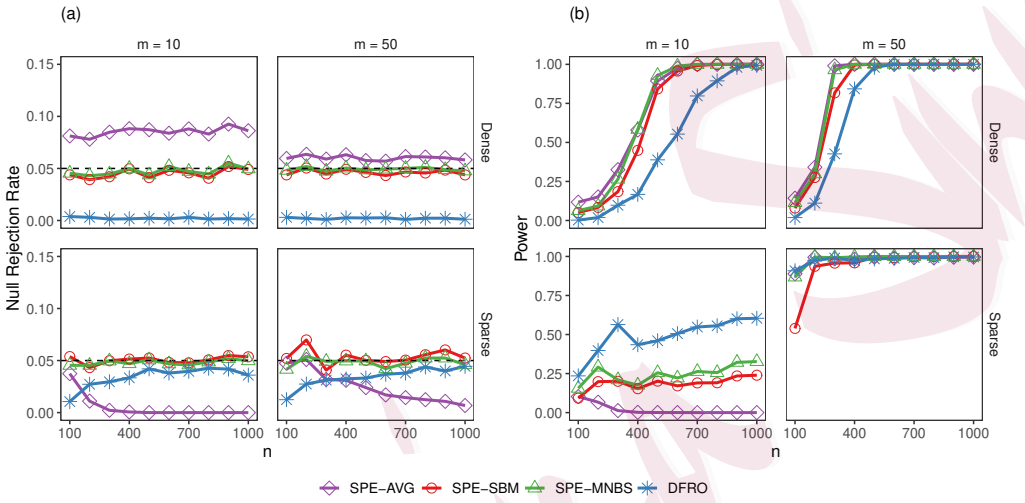

Figure 2: Simulation results for testing networks with a graphon structure. The first and the second rows are for dense and sparse networks, respectively. (a) Null rejection rate. (b) Power under the alternative.

We see from Figure 2 that SPE-MNBS exhibits superior performance over the other tests in terms of both null rejection rate and power except for the case of small $m = 10$ and sparse structure. We also observe that SPE-SBM shows a lower rejection rate than the nominal level in the dense case, which suggests that SPE-SBM is more sensitive to network topologies that deviate from an SBM. The behaviors of SPE-AVG and DFRO are similar

to those in the first example, as we continue to see subpar performance, especially for small m.

4.3 Correlated Erdös-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 from 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 $p_2 = p_1 = 0.9$, whereas $p_2 = 0.83$ under the alternative hypothesis. The results are shown in Figure 3.

We see that DFRO has consistently high power in the alternative setting for the entire range of n, which is only matched for our tests as n increases, with SPE-AVG outperforming both SPE-SBM and SPE-MNBS. However,

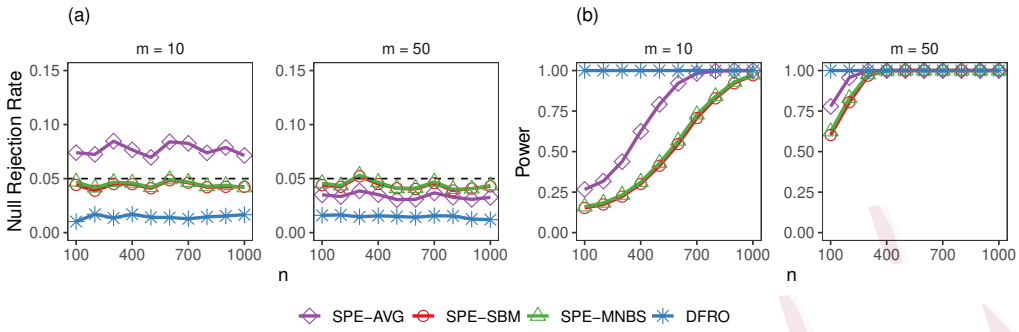

Figure 3: Simulation results for testing networks with a correlated ER structure. (a) Null rejection rate. (b) Power under the alternative.

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 on three real datasets representing three different settings of interest to the biological research community: StarPlus, COBRE, and MB datasets. 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 dataset

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 et al., 2017). A description of the datasets 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 is often done 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 from Yuan and Wen (2021), which we refer to as WRG. We test if the groups defined by their respective labels – picture/sentence, schizophrenic/control, preterm/term – are different. To do so, we specify a null hypothesis which says the underlying random distributions are equal against the alternative that says they are different. We refer to this as the "alternative setting" because the two samples are different with respect to their group label.

As is, we cannot directly apply WRG since it requires the same sample

sizes for both groups, which is not true of the COBRE or MB datasets. 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 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 reject rates of 1, 0.726, and 1, respectively. This is consistent with previous research on distinguishing the cognitive states of looking at a picture and a sentence (Mitchell et al., 2004; Wang et al., 2003; Mitchell et al., 2003). For the COBRE and MB datasets, 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 datasets, respectively.

Next, we perform an in silico experiment with the real data by sub-sampling 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 if the different groups are generated by different underlying distributions, e.g. one for schizophrenic and another for non-schizophrenic. Therefore, we want to check if the null rejection rate is close to the nominal level in an experiment where 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 datasets, respectively. For WRG, we test two subsamples of the two groups, both with half of the original sample size.

After 1000 random subsamples of the networks and $Q = 1$ for each subsample, for the StarPlus networks, SPE-SBM fails to reject the null hypothesis with reject rates of 0.006, which is expected since 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 are improved 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 performs better than WRG, the null rejection rates are still very inflated compared to the nominal $\alpha = 0.05$.

We speculate that this is happening because even within one class, there is a lot of variation. That is, a subsample of brain networks with schizophrenia may look very different from 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 could 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

As the results for the weighted tests showed inflated rejection rates in our simulated null setting, there is reason to believe 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 et al., 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 be 1 when the absolute values of the corresponding weights are larger than the threshold and 0 otherwise. This threshold relates directly to the density of the networks. Note that WRG is for weighted networks and is therefore excluded. With the same procedures 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.

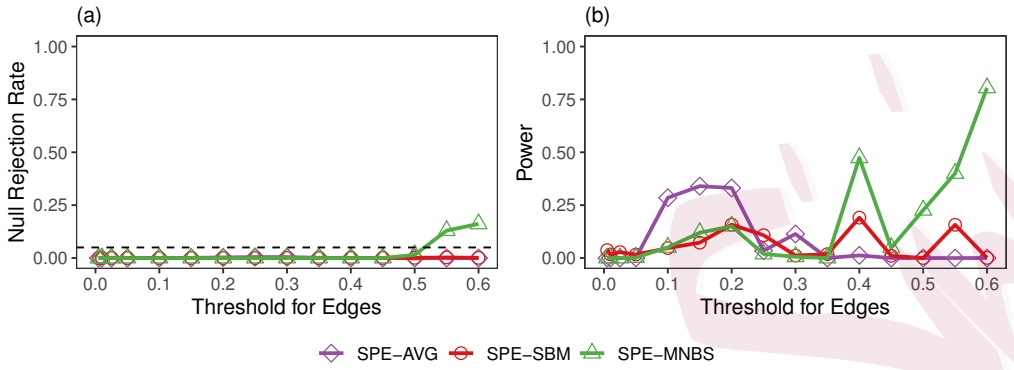

Figure 4: Figures (a) and (b) show the null rejection rate and power, respectively, for different thresholds for binarizing the StarPlus networks.

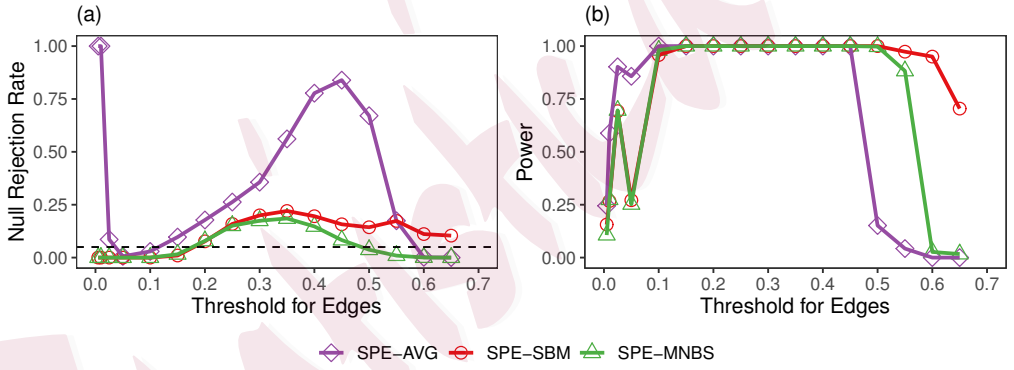

Figure 5: Figures (a) and (b) show the null rejection rate and power, respectively, for different thresholds for binarizing the COBRE networks.

The plots illustrate the tradeoff between the false positive rate in our null setting and the true positive rate in our alternative setting, which both are functions of the threshold. As the threshold for an edge increases, the

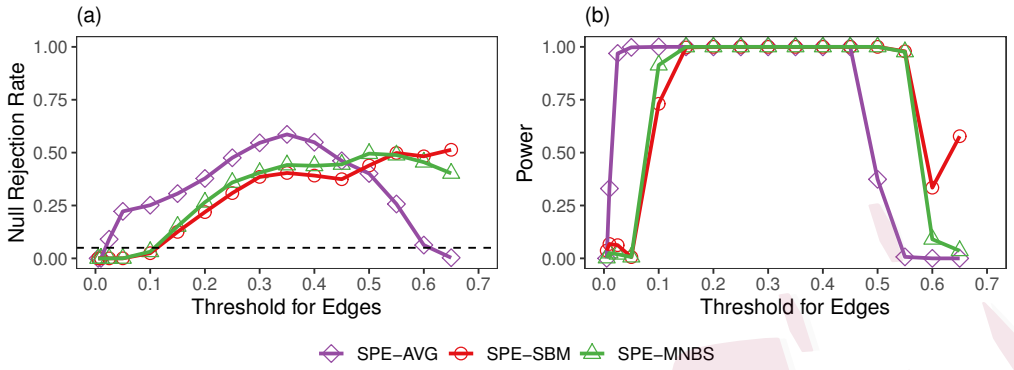

Figure 6: Figures (a) and (b) show the null rejection rate and power, respectively, for different thresholds for binarizing the MB networks.

network becomes more sparse, resulting in a higher rejection rate in our null setting. For thresholds above 0.6, some of the networks became 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 threshold could provide practitioners 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 networks directly. Moreover, we also see a low rejection rate in our null setting, especially for a threshold of 0.1, which seems to provide the best tradeoff. This suggests that the signal-to-noise ratio in the weights is too low, which can be mitigated through 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. Discussion

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

References

Arroyo Relión, J. D., D. Kessler, E. Levina, and S. F. Taylor (2019). Network classification with applications to brain connectomics. *The Annals of Applied Statistics 13*(3), 1648–1677.

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

Banerjee, D. and Z. Ma (2017). Optimal hypothesis testing for stochastic block models with

growing degrees. *arXiv:1705.05305*.

Bernstein, S. (1946). *The Theory of Probabilities*. Moscow, U.S.S.R.: Gastehizdat Publishing House. In Russian.

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

Bickel, P. J. and P. Sarkar (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78*(1), 253–273.

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

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

Chen, L., J. Zhou, and L. Lin (2021). Hypothesis testing for populations of networks. *Communications in Statistics-Theory and Methods*, 1–24.

Chen, S. X. and Y.-L. Qin (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics 38*(2), 808–835.

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

REFERENCES

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

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

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

Ghoshdastidar, D., M. Gutzeit, A. Carpentier, and U. V. Luxburg (2020). Two-sample hypothesis testing for inhomogeneous random graphs. *The Annals of Statistics 48*(4), 2208–2229.

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

Ginestet, C. E., A. P. Fournel, and A. Simmons (2014). Statistical network analysis for functional MRI: summary networks and group comparisons. *Frontiers in Computational Neuroscience 8*. Article 51.

Ginestet, C. E., J. Li, P. Balachandran, S. Rosenberg, E. D. Kolaczyk, et al. (2017). Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics 11*(2), 725–750.

Jin, J., Z. T. Ke, and S. Luo (2021). Optimal adaptivity of signed-polygon statistics for network testing. *49*(6), 3408–3433.

Josephs, N., L. Lin, S. Rosenberg, and E. D. Kolaczyk (2020). Bayesian classification, anomaly

detection, and survival analysis using network inputs with application to the microbiome. *arXiv:2004.04765*.

Kolaczyk, E. D., L. Lin, S. Rosenberg, J. Walters, and J. Xu (2020). Averages of unlabeled networks: Geometric characterization and asymptotic behavior. *The Annals of Statistics 48*(1), 514–538.

Layeghifard, M., D. M. Hwang, and D. S. Guttman (2017). Disentangling interactions in the microbiome: A network perspective. *Trends in Microbiology 25*(3), 217–228.

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

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

Mitchell, T. M., R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman (2004). Learning to decode cognitive states from brain images. *Machine Learning 57*(1-2), 145–175.

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

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

REFERENCES

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

Tang, M., A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics 26*(2), 344–354.

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

Xu, G., L. Lin, P. Wei, and W. Pan (2016). An adaptive two-sample test for high-dimensional means. *Biometrika 103*(3), 609–624.

Yuan, M. and Q. Wen (2021). A practical two-sample test for weighted random graphs. *Journal of Applied Statistics*, 1–17.

Zhang, Y., E. Levina, and J. Zhu (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika 104*(4), 771–783.

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

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