# Practical Methods for Graph Two-Sample Testing

**Debarghya Ghoshdastidar**
Department of Computer Science
University of Tübingen
ghoshdas@informatik.uni-tuebingen.de

**Ulrike von Luxburg**
Department of Computer Science
University of Tübingen
Max Planck Institute for Intelligence Systems
luxburg@informatik.uni-tuebingen.de

## Abstract

Hypothesis testing for graphs has been an important tool in applied research fields for more than two decades, and still remains a challenging problem as one often needs to draw inference from few replicates of large graphs. Recent studies in statistics and learning theory have provided some theoretical insights about such high-dimensional graph testing problems, but the practicality of the developed theoretical methods remains an open question.

In this paper, we consider the problem of two-sample testing of large graphs. We demonstrate the practical merits and limitations of existing theoretical tests and their bootstrapped variants. We also propose two new tests based on asymptotic distributions. We show that these tests are computationally less expensive and, in some cases, more reliable than the existing methods.

## 1 Introduction

Hypothesis testing is one of the most commonly encountered statistical problems that naturally arises in nearly all scientific disciplines. With the widespread use of networks in bioinformatics, social sciences and other fields since the turn of the century, it was obvious that the hypothesis testing of graphs would soon become a key statistical tool in studies based on network analysis. The problem of testing for differences in networks arises quite naturally in various situations. For instance, Bassett et al. (2008) study the differences in anatomical brain networks of schizophrenic patients and healthy individuals, whereas Zhang et al. (2009) test for statistically significant topological changes in gene regulatory networks arising from two different treatments of breast cancer. As Clarke et al. (2008) and Hyduke et al. (2013) point out, the statistical challenge associated with network testing is the curse of dimensionality as one needs to test large graphs based on few independent samples. Ginestet et al. (2014) show that complications can also arise due to the widespread use of multiple testing principles that rely on performing independent tests for every edge.

Although network analysis has been a primary research topic in statistics and machine learning, theoretical developments related to testing random graphs have been rather limited until recent times. Property testing of graphs has been well studied in computer science (Goldreich et al., 1998), but probably the earliest instances of the theory of random graph testing are the works on community detection, which use hypothesis testing to detect if a network has planted communities or to determine the number of communities in a block model (Arias-Castro and Verzelen, 2014, Bickel and Sarkar, 2016, Lei, 2016). In the present work, we are interested in the more general and practically important problem of two-sample testing: *Given two populations of random graphs, decide whether both populations are generated from the same distribution or not.* While there have been machine learning approaches to quantify similarities between graphs for the purpose of classification, clustering etc. (Borgwardt et al., 2005, Shervashidze et al., 2011), the use of graph distances for the purpose of hypothesis testing is more recent (Ginestet et al., 2017). Most approaches for graph testing based on classical two-sample tests are applicable in the relatively low-dimensional setting, where the

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

population size (number of graphs) is larger than the size of the graphs (number of vertices). However, Hyduke et al. (2013) note that this scenario does not always apply because the number of samples could be potentially much smaller — for instance, one may need to test between two large regulatory networks (that is, population size is one). Such scenarios can be better tackled from a perspective of high-dimensional statistics as shown in Tang et al. (2016), Ghoshdastidar et al. (2017a), where the authors study two-sample testing for specific classes of random graphs with particular focus on the small population size.

In this work, we focus on the framework of the graph two-sample problem considered in Tang et al. (2016), Ginestet et al. (2017), Ghoshdastidar et al. (2017a), where all graphs are defined on a common set of vertices. Assume that the number of vertices in each graph is $n$, and the sample size of either population is $m$. One can consider the two-sample problem in three different regimes: **(i)** $m$ is large; **(ii)** $m > 1$, but much smaller than $n$; and **(iii)** $m = 1$. The first setting is the simplest one, and practical tests are known in this case (Gretton et al., 2012, Ginestet et al., 2017). However, there exist many application domains where already the availability of only a small population of graphs is a challenge, and large populations are completely out of bounds. The latter two cases of small $m > 1$ and $m = 1$ have been studied in Ghoshdastidar et al. (2017a) and Tang et al. (2016), where theoretical tests based on concentration inequalities have been developed and practical bootstrapped variants of the tests have been suggested. The contribution of the present work is three-fold:

1. For the cases of $m > 1$ and $m = 1$, we propose new tests that are based on asymptotic null distributions under certain model assumptions and we prove their statistical consistency (Sections 4 and 5 respectively). The proposed tests are devoid of bootstrapping, and hence, computationally faster than existing bootstrapped tests for small $m$. Detailed descriptions of the tests are provided in the supplementary material.
2. We compare the practical merits and limitations of existing tests with the proposed tests (Section 6 and supplementary). We show that the proposed tests are more powerful and reliable than existing methods in some situations.
3. Our aim is also to make the existing and proposed tests more accessible for applied research. We provide Matlab implementations of the tests in the supplementary material.

The present work is focused on the assumption that all networks are defined over the same set of vertices. This may seem restrictive in some application areas, but it is commonly encountered in other areas such as brain network analysis or molecular interaction networks, where vertices correspond to well-defined regions of the brain or protein structures. Few works study the case where graphs do not have vertex correspondences in context of clustering (Mukherjee et al., 2017) and testing (Ghoshdastidar et al., 2017b, Tang et al., 2017). But, theoretical guarantees are only known for specific choices of network functions (triangle counts or graph spectra), or under the assumption of an underlying embedding of the vertices.

**Notation.** We use the asymptotic notation $o_n(\cdot)$ and $\omega_n(\cdot)$, where the asymptotics are with respect to the number of vertices $n$. We say $x = o_n(y)$ and $y = \omega_n(x)$ when $\lim_{n\to\infty} \frac{x}{y} = 0$. We denote the matrix Frobenius norm by $\|\cdot\|_F$ and the spectral norm or largest singular value by $\|\cdot\|_2$.

## 2 Problem Statement

We consider the following framework of two-sample setting. Let $V$ be a set of $n$ vertices. Let $G_1, \ldots, G_m$ and $H_1, \ldots, H_m$ be two populations of undirected unweighted graphs defined on the common vertex set $V$, where each population consists of independent and identically distributed samples. The two-sample hypothesis testing problem is as follows:

*Test whether $(G_i)_{i=1,\ldots,m}$ and $(H_i)_{i=1,\ldots,m}$ are generated from the same random model or not.*

There exist a plethora of nonparametric tests that are provably consistent for $m \to \infty$. In particular, kernel based tests (Gretton et al., 2012) are known to be suitable for two-sample problems in large dimensions. These tests, in conjunction with graph kernels (Shervashidze et al., 2011, Kondor and Pan, 2016) or distances (Mukherjee et al., 2017), may be used to derive consistent procedures for testing between two large populations of graphs. Such principles are applicable even under a more general framework without vertex correspondence (see Gretton et al., 2012). However, given graphs

on a common vertex set, the most natural approach is to construct tests based on the graph adjacency matrix or the graph Laplacian (Ginestet et al., 2017). To be precise, one may view each undirected graph on $n$ vertices as a $\binom{n}{2}$-dimensional vector and use classical two-sample tests based on the $\chi^2$ or $T^2$ statistics (Anderson, 1984). Unfortunately, such tests require an estimate of the $\binom{n}{2} \times \binom{n}{2}$-dimensional sample covariance matrix, which cannot be accurately obtained from a moderate sample size. For instance, Ginestet et al. (2017) need regularisation of the covariance estimate even for moderate sized problems ($n = 40, m = 100$), and it is unknown whether such methods work for brain networks obtained from a single-lab experimental setup ($m < 20$). For $m \ll n$, it is indeed hard to prove consistency results under the general two-sample framework described above since the correlation among the edges can be arbitrary. Hence, we develop our theory for random graphs with independent edges. Tang et al. (2016) show that tests derived for such graphs are also useful in practice.

We assume that the graphs are generated from the inhomogeneous Erdős-Rényi (IER) model (Bollobas et al., 2007). This model has been considered in the work of Ghoshdastidar et al. (2017a) and subsumes other models studied in the context of graph testing such as dot product graphs (Tang et al., 2016) and stochastic block models (Lei, 2016). Given a symmetric matrix $P \in [0,1]^{n \times n}$ with zero diagonal, a graph $G$ is said to be an IER graph with population adjacency $P$, denoted as $G \sim \text{IER}(P)$, if its symmetric adjacency matrix $A_G \in \{0,1\}^{n \times n}$ satisfies:

$$(A_G)_{ij} \sim \text{Bernoulli}(P_{ij}) \text{ for all } i < j, \quad \text{and} \quad \{(A_G)_{ij} : i < j\} \text{ are mutually independent.}$$

For any $n$, we state the two-sample problem as follows. Let $P^{(n)}, Q^{(n)} \in [0,1]^{n \times n}$ be two symmetric matrices. Given $G_1, \ldots, G_m \sim_{\text{iid}} \text{IER}\left(P^{(n)}\right)$ and $H_1, \ldots, H_m \sim_{\text{iid}} \text{IER}\left(Q^{(n)}\right)$, test the hypotheses

$$\mathcal{H}_0 : P^{(n)} = Q^{(n)} \quad \text{against} \quad \mathcal{H}_1 : P^{(n)} \neq Q^{(n)}. \tag{1}$$

Our theoretical results in subsequent sections will often be in the asymptotic case as $n \to \infty$. For this, we assume that there are two sequences of models $\left(P^{(n)}\right)_{n \geq 1}$ and $\left(Q^{(n)}\right)_{n \geq 1}$, and the sequences are identical under the null hypothesis $\mathcal{H}_0$. We derive asymptotic powers of the proposed tests assuming certain separation rates under the alternative hypothesis.

## 3 Testing large population of graphs $(m \to \infty)$

Before proceeding to the case of small population size, we discuss a baseline approach that is designed for the large $m$ regime ($m \to \infty$). The following discussion provides a $\chi^2$-type test statistic for networks, which is a simplification of Ginestet et al. (2017) under the IER assumption. Given the adjacency matrices $A_{G_1}, \ldots, A_{G_m}$ and $A_{H_1}, \ldots, A_{H_m}$, consider the test statistic

$$T_{\chi^2} = \sum_{i<j} \frac{\left((\overline{A}_G)_{ij} - (\overline{A}_H)_{ij}\right)^2}{\frac{1}{m(m-1)} \sum_{k=1}^{m} \left((A_{G_k})_{ij} - (\overline{A}_G)_{ij}\right)^2 + \frac{1}{m(m-1)} \sum_{k=1}^{m} \left((A_{H_k})_{ij} - (\overline{A}_H)_{ij}\right)^2}, \tag{2}$$

where $(\overline{A}_G)_{ij} = \frac{1}{m} \sum_{k=1}^{m} (A_{G_k})_{ij}$. It is easy to see that under $\mathcal{H}_0$, $T_{\chi^2} \to \chi^2\left(\frac{n(n-1)}{2}\right)$ in distribution as $m \to \infty$ for any fixed $n$. This suggests a $\chi^2$-type test similar to Ginestet et al. (2017). However, like any classical test, no performance guarantee can be given for small $m$ and our numerical results show that such a test is powerless for small $m$ and sparse graphs. Hence, in the rest of the paper, we consider tests that are powerful even for small $m$.

## 4 Testing small populations of large graphs $(m > 1)$

The case of small $m > 1$ for IER graphs was first studied from a theoretical perspective in Ghoshdastidar et al. (2017a), and the authors also show that, under a minimax testing framework, the testing problem is quite different for $m = 1$ and $m > 1$. From a practical perspective, small $m > 1$ is a common situation in neural imaging with only few subjects. The case of $m = 2$ is also interesting for testing between two individuals based on test-retest diffusion MRI data, where two scans are collected from each subject with a separation of multiple weeks (Landman et al., 2011).

Under the assumption of IER models described in Section 2 and given the adjacency matrices $A_{G_1}, \ldots, A_{G_m}$ and $A_{H_1}, \ldots, A_{H_m}$, Ghoshdastidar et al. (2017a) propose test statistics based on

estimates of the distances $\left\|P^{(n)} - Q^{(n)}\right\|_2$ and $\left\|P^{(n)} - Q^{(n)}\right\|_F$ up to certain normalisation factors that account for sparsity of the graphs. They consider the following two test statistics

$$T_{spec} = \frac{\left\|\sum_{k=1}^{m} A_{G_k} - A_{H_k}\right\|_2}{\sqrt{\max_{1\le i\le n}\sum_{j=1}^{n}\sum_{k=1}^{m}(A_{G_k})_{ij} + (A_{H_k})_{ij}}}, \quad \text{and} \tag{3}$$

$$T_{fro} = \frac{\sum_{i<j}\left(\sum_{k\le m/2}(A_{G_k})_{ij} - (A_{H_k})_{ij}\right)\left(\sum_{k>m/2}(A_{G_k})_{ij} - (A_{H_k})_{ij}\right)}{\sqrt{\sum_{i<j}\left(\sum_{k\le m/2}(A_{G_k})_{ij} + (A_{H_k})_{ij}\right)\left(\sum_{k>m/2}(A_{G_k})_{ij} + (A_{H_k})_{ij}\right)}}. \tag{4}$$

Subsequently, theoretical tests are constructed based on concentration inequalities: one can show that with high probability, the test statistics are smaller than some specified threshold under the null hypothesis, but they exceed the same threshold if the separation between $P^{(n)}$ and $Q^{(n)}$ is large enough. In practice, however, the authors note that the theoretical thresholds are too large to be exceeded for moderate $n$, and recommend estimation of the threshold through bootstrapping. Each bootstrap sample is generated by randomly partitioning the entire population $G_1, \ldots, G_m, H_1, \ldots, H_m$ into two parts, and $T_{spec}$ or $T_{fro}$ are computed based on this random partition. This procedure provides an approximation of the statistic under the null model. We refer to these tests as `Boot-Spectral` and `Boot-Frobenius`, and show their limitations for small $m$ via simulations. Detailed descriptions of these tests are included in Appendix B in the supplementary.

We now propose a test based on the asymptotic behaviour of $T_{fro}$ in (4) as $n \to \infty$. We state the asymptotic behaviour in the following result.

**Theorem 1 (Asymptotic test based on $T_{fro}$).** *In the two-sample framework of Section 2, assume that $P^{(n)}, Q^{(n)}$ have entries bounded away from 1, and satisfy $\max\left\{\left\|P^{(n)}\right\|_F, \left\|Q^{(n)}\right\|_F\right\} = \omega_n(1)$. Under the null hypothesis, $\lim_{n\to\infty} T_{fro}$ is dominated by a standard normal random variable, and hence, for any $\alpha \in (0,1)$,*

$$\mathbb{P}\big(T_{fro} \notin [-t_\alpha, t_\alpha]\big) \le \alpha + o_n(1), \tag{5}$$

*where $t_\alpha = \Phi^{-1}(1-\frac{\alpha}{2})$ is the $\frac{\alpha}{2}$ upper quantile of the standard normal distribution.*

*On the other hand, if $\left\|P^{(n)} - Q^{(n)}\right\|_F^2 = \omega_n\left(\frac{1}{m}\max\left\{\left\|P^{(n)}\right\|_F, \left\|Q^{(n)}\right\|_F\right\}\right)$, then*

$$\mathbb{P}\big(T_{fro} \in [-t_\alpha, t_\alpha]\big) = o_n(1). \tag{6}$$

The proof, given in Appendix A, is based on the use of the Berry-Esseen theorem (Berry, 1941). Using Theorem 1, we propose an $\alpha$-level test based on asymptotic normal dominance of $T_{fro}$.

**Proposed Test** `Asymp-Normal`: *Reject the null hypothesis if $|T_{fro}| > t_\alpha$.*

A detailed description of this test is given in Appendix B. The assumption $\left\|P^{(n)}\right\|_F, \left\|Q^{(n)}\right\|_F = \omega_n(1)$ is not restrictive since it is quite similar to assuming that the number of edges is super-linear in $n$, that is, the graphs are not too sparse. We note that unlike the $\chi^2$-test of Section 2, here the asymptotics are for $n \to \infty$ instead of $m \to \infty$, and hence, the behaviour under null hypothesis may not improve for larger $m$. The asymptotic unit power of the `Asymp-Normal` test, as shown in Theorem 1, is proved under a separation condition, which is not surprising since we have access to only a finite number of graphs. The result also shows that for large $m$, smaller separations can be detected by the proposed test.

**Remark 2 (Computational effort).** Note that the computational complexity for computing the test statistics in (3) and (4) is *linear in the total number of edges in the entire population*. However, the bootstrap tests require computation of the test statistic multiple times (equal to number of bootstrap samples $b$; we use $b = 200$ in our experiments). On the other hand, the proposed test compute the statistic once, and is much faster ($\sim$200 times). Moreover, if the graphs are too large to be stored in memory, bootstrapping requires multiple passes over the data, while the proposed test requires only a single pass.

## 5 Testing difference between two large graphs $(m = 1)$

The case of $m = 1$ is perhaps the most interesting from theoretical perspective: the objective is to detect whether two large graphs $G$ and $H$ are identically distributed or not. This finds application in detecting differences in regulatory networks (Zhang et al., 2009) or comparing brain networks of individuals (Tang et al., 2016). Although the concentration based test using $T_{spec}$ is applicable even for $m = 1$ (Ghoshdastidar et al., 2017a), bootstrapping based on label permutation is infeasible for $m = 1$ since there is no scope of permuting labels with unit population size. Tang et al. (2016), however, propose a concentration based test in this case and suggest a bootstrapping based on low rank assumption of the population adjacency. Tang et al. (2016) study the two-sample problem for random dot product graphs, which are IER graphs with low rank population adjacency matrices (ignoring the effect of zero diagonal). This class includes the stochastic block model, where the rank equals the number of communities. Let $G \sim \text{IER}\left(P^{(n)}\right)$ and $H \sim \text{IER}\left(Q^{(n)}\right)$, and assume that $P^{(n)}$ and $Q^{(n)}$ are of rank $r$. One defines the adjacency spectral embedding (ASE) of graph $G$ as $X_G = U_G \Sigma_G^{1/2}$, where $\Sigma_G \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing $r$ largest singular values of $A_G$ and $U_G \in \mathbb{R}^{n\times r}$ is the matrix of corresponding left singular vectors. Tang et al. (2016) propose the test statistic

$$T_{ASE} = \min\left\{\|X_G - X_H W\|_F : W \in \mathbb{R}^{r\times r}, WW^T = I\right\}, \tag{7}$$

where the rank $r$ is assumed to be known. The rotation matrix $W$ aligns the ASE of the two graphs. Tang et al. (2016) theoretically analyse a concentration based test, where the null hypothesis is rejected if $T_{ASE}$ crosses a suitably chosen threshold. In practice, they suggest the following bootstrapping to determine the threshold (Algorithm 1 in Tang et al., 2016). One may approximate $P^{(n)}$ by the estimated population adjacency (EPA) $\widehat{P} = X_G X_G^T$. More random dot product graphs can be simulated from $\widehat{P}$, and a bootstrapped threshold can be obtained by computing $T_{ASE}$ for pairs of graphs generated from $\widehat{P}$. Instead of the $T_{ASE}$ statistic, one may also use a statistic based on EPA as

$$T_{EPA} = \left\|\widehat{P} - \widehat{Q}\right\|_F. \tag{8}$$

This statistic has been used as distance measure in the context of graph clustering (Mukherjee et al., 2017). We refer to the tests based on the statistics in (7) and (8), and the above bootstrapping procedure by `Boot-ASE` and `Boot-EPA` (see Appendix B for detailed descriptions). We find that the latter performs better, but both tests work under the condition that the population adjacency is of low rank, and the rank is precisely known. Our numerical results demonstrate the limitations of these tests when the rank is not correctly known.

Alternatively, we propose a test based on the asymptotic distribution of eigenvalues that is not restricted to graphs with low rank population adjacencies. Given $G \sim \text{IER}\left(P^{(n)}\right)$ and $H \sim \text{IER}\left(Q^{(n)}\right)$, consider the matrix $C \in \mathbb{R}^{n\times n}$ with zero diagonal and for $i \neq j$,

$$C_{ij} = \frac{(A_G)_{ij} - (A_H)_{ij}}{\sqrt{(n-1)\left(P_{ij}^{(n)}\left(1 - P_{ij}^{(n)}\right) + Q_{ij}^{(n)}\left(1 - Q_{ij}^{(n)}\right)\right)}}. \tag{9}$$

We assume that the entries of $P^{(n)}$ and $Q^{(n)}$ are not arbitrarily close to 1, and define $C_{ij} = 0$ when $C_{ij} = \frac{0}{0}$. We show that the extreme eigenvalues of $C$ asymptotically follow the Tracy-Widom law, which characterises the distribution of the largest eigenvalues of matrices with independent standard normal entries (Tracy and Widom, 1996). Subsequently, we show that $\|C\|_2$ is a useful test statistic.

**Theorem 3 (Asymptotic test based on $\|C\|_2$).** *Consider the above setting of two-sample testing, and let $C$ be as defined in* (9)*. Let $\lambda_1(C)$ and $\lambda_n(C)$ be the largest and smallest eigenvalues of $C$.*

*Under the null hypothesis, that is, if $P^{(n)} = Q^{(n)}$ for all $n$, then*

$$n^{2/3}\big(\lambda_1(C) - 2\big) \to TW_1 \quad \textit{and} \quad n^{2/3}\big(-\lambda_n(C) - 2\big) \to TW_1$$

*in distribution as $n \to \infty$, where $TW_1$ is the Tracy-Widom law for orthogonal ensembles. Hence,*

$$\mathbb{P}\left(n^{2/3}(\|C\|_2 - 2) > \tau_\alpha\right) \leq \alpha + o_n(1), \tag{10}$$

*for any $\alpha \in (0, 1)$, where $\tau_\alpha$ is the $\frac{\alpha}{2}$ upper quantile of the $TW_1$ distribution.*

*On the other hand, if $P^{(n)}$ and $Q^{(n)}$ are such that $\|\mathbb{E}[C]\|_2 \geq 4 + \omega_n(n^{-2/3})$, then*

$$\mathbb{P}\left(n^{2/3}(\|C\|_2 - 2) \leq \tau_\alpha\right) = o_n(1). \tag{11}$$

The proof, given in Appendix A, relies on results on the spectrum of random matrices (Erdős et al., 2012, Lee and Yin, 2014), and have been previously used for the special case of determining the number of communities in a block model (Bickel and Sarkar, 2016, Lei, 2016). If the graphs are assumed to be block models, then asymptotic power can be proved under more precise conditions on difference in population adjacencies $P^{(n)} - Q^{(n)}$ (see Appendix A.3). From a practical perspective, $C$ cannot be computed since $P^{(n)}$ and $Q^{(n)}$ are unknown. Still, one may approximate them by relying on a weaker version of Szemerédi's regularity lemma, which implies that large graphs can be approximated by stochastic block models with possibly large number of blocks (Lovász, 2012). To this end, we propose to estimate $P^{(n)}$ from $A_G$ as follows. We use a community detection algorithm, such as normalised spectral clustering (Ng et al., 2002), to find $r$ communities in $G$ ($r$ is a parameter for the test). Subsequently $P^{(n)}$ is approximated by a block matrix $\widetilde{P}$ such that if $i, j$ lie in communities $V_1, V_2$ respectively, then $\widetilde{P}_{ij}$ is the mean of the sub-matrix of $A_G$ restricted to $V_1 \times V_2$. Similarly one can also compute $\widetilde{Q}$ from $A_H$. Hence, we propose a Tracy-Widom test statistic as

$$T_{TW} = n^{2/3}\left(\left\|\widetilde{C}\right\|_2 - 2\right), \tag{12}$$

$$\text{where} \quad \widetilde{C}_{ij} = \frac{(A_G)_{ij} - (A_H)_{ij}}{\sqrt{(n-1)\left(\widetilde{P}_{ij}\left(1-\widetilde{P}_{ij}\right) + \widetilde{Q}_{ij}\left(1-\widetilde{Q}_{ij}\right)\right)}} \quad \text{for all } i \neq j$$

and the diagonal is zero. The proposed $\alpha$-level test based on $T_{TW}$ and Theorem 3 is the following.

**Proposed Test** `Asymp-TW`: *Reject the null hypothesis if $T_{TW} > \tau_\alpha$.*

A detailed description of the test, as used in our implementations, is given in Appendix B. We note that unlike bootstrap tests based on $T_{ASE}$ or $T_{EPA}$, the proposed test uses the number of communities (or rank) $r$ only for approximation of $P^{(n)}, Q^{(n)}$, and the power of the test is not sensitive to the choice of $r$. In addition, the computational benefit of a distribution based test over bootstrap tests, as noted in Remark 2, is also applicable in this case.

## 6 Numerical results

In this section, we empirically compare the merits and limitations of the tests discussed in the paper. We present our numerical results in three groups: (i) results for random graphs for $m > 1$, (ii) results for random graphs for $m = 1$, and (iii) results for testing real networks. For $m > 1$, we consider four tests. `Boot-Spectral` and `Boot-Frobenius` are the bootstrap tests based on $T_{spec}$ (3) and $T_{fro}$ (4), respectively. `Asymp-Chi2` is the $\chi^2$-type test based on $T_{\chi^2}$ (2), which is suited for the large $m$ setting, and finally, the proposed test `Asymp-Normal` is based on the normal dominance of $T_{fro}$ as $n \to \infty$ as shown in Theorem 1. For $m = 1$, we consider three tests. `Boot-ASE` and `Boot-EPA` are the bootstrap tests based on $T_{ASE}$ (7) and $T_{EPA}$ (8), respectively. `Asymp-TW` is the proposed test based on $T_{TW}$ (12) and Theorem 3. Appendices B and C in the supplementary contain descriptions of all tests and additional numerical results. Matlab codes are provided in the supplementary.[1]

### 6.1 Comparative study on random graphs for $m > 1$

For this study, we generate graphs from stochastic block models with 2 communities as considered in Tang et al. (2016). We define $P^{(n)}$ and $Q^{(n)}$ as follows. The vertex set of size $n$ is partitioned into two communities, each of size $n/2$. In $P^{(n)}$, edges occur independently with probability $p$ within each community, and with probability $q$ between two communities. $Q^{(n)}$ has the same block structure as $P^{(n)}$, but edges occur with probability $(p + \epsilon)$ within each community. Under the null hypothesis $\epsilon = 0$ and hence $Q^{(n)} = P^{(n)}$, whereas under the alternative hypothesis, we set $\epsilon > 0$.

[1] Also available at: `https://github.com/gdebarghya/Network-TwoSampleTesting`.

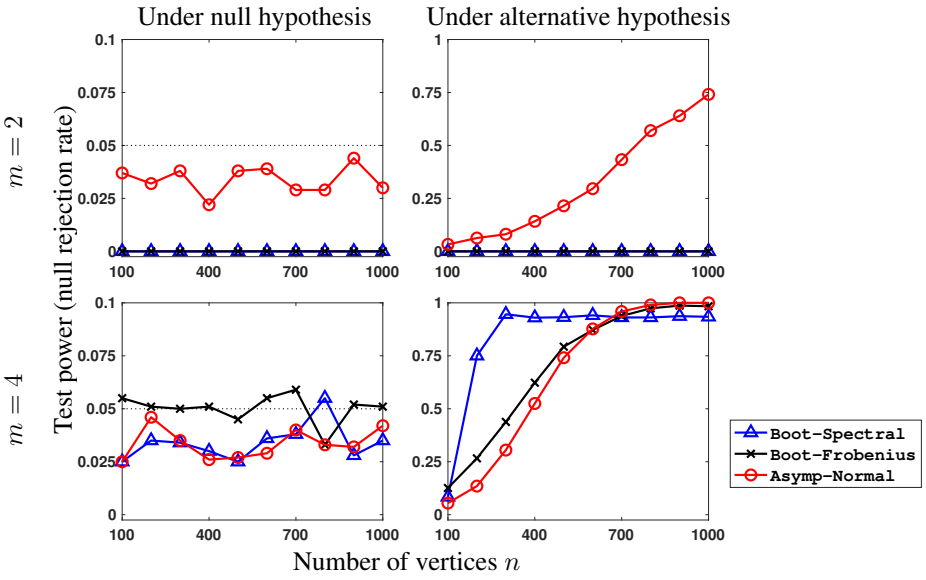


Figure 1: Power of different tests for increasing number of vertices $n$, and for $m = 2, 4$. The dotted line for case of null hypothesis corresponds to the significance level of 5%.

In our first experiment, we study the performance of different tests for varying $m$ and $n$. We let $n$ grow from 100 to 1000 in steps of 100, and set $p = 0.1$ and $q = 0.05$. We set $\epsilon = 0$ and 0.04 for null and alternative hypotheses, respectively. We use two values of population size, $m \in \{2, 4\}$, and fix the significance level at $\alpha = 5\%$. Figure 1 shows the rate of rejecting the null hypothesis (test power) computed from 1000 independent runs of the experiment. Under the null model, the test power should be smaller than $\alpha = 5\%$, whereas under the alternative model, a high test power (close to 1) is desirable. We see that for $m = 2$, only `Asymp-Normal` has power while the bootstrap tests have zero rejection rate. This is not surprising as bootstrapping is impossible for $m = 2$. For $m = 4$, `Boot-Frobenius` has a behaviour similar to `Asymp-Normal` although the latter is computationally much faster. `Boot-Spectral` achieves a higher power for small $n$ but cannot achieve unit power. `Asymp-Chi2` has an erratic behaviour for small $m$, and hence, we study it for larger sample size in Figure 3 (in Appendix C). As is expected, `Asymp-Chi2` has desired performance only for $m \gg n$.

We also study the effect of edge sparsity on the performance of the tests. For this, we consider the above setting, but scale the edge probabilities by a factor of $\rho$, where $\rho = 1$ is exactly same as the above setting while larger $\rho$ corresponds to denser graphs. Figure 4 in the appendix shows the results in this case, where we fix $n = 500$ and vary $\rho \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$ and $m \in \{2, 4, 6, 8, 10\}$. We again find that `Asymp-Normal` and `Boot-Frobenius` have similar trends for $m \geq 4$. All tests perform better for dense graphs, but `Boot-Spectral` may be preferred for sparse graphs when $m \geq 6$.

### 6.2 Comparative study on random graphs for $m = 1$

We conduct similar experiments for the case of $m = 1$. Recall that bootstrap tests for $m = 1$ work under the assumption that the population adjacencies are of low rank. This holds in above considered setting of block models, where the rank is 2. We first demonstrate the effect of knowledge of true rank on the test power. We use $r \in \{2, 4\}$ to specify the rank parameter for bootstrap tests, and also as the number of blocks used for community detection step of `Asymp-TW`. Figure 2 shows the power of the tests for the above setting with $\rho = 1$ and growing $n$. We find that when $r = 2$, that is, true rank is known, both bootstrap tests perform well under alternative hypothesis, and outperform `Asymp-TW`, although `Boot-ASE` has a high type-I error rate. However, when an over-estimate of rank is used ($r = 4$), both bootstrap tests break down — `Boot-EPA` always rejects while `Boot-ASE` always accepts — but the performance of `Asymp-TW` is robust to this parameter change.

We also study the effect of sparsity by varying $\rho$ (see Figure 5 in Appendix C). We only consider the case $r = 2$. We find that all tests perform better in dense regime, and the rejection rate of `Asymp-TW` under null is below 5% even for small graphs. However, the performance of both `Boot-ASE` and

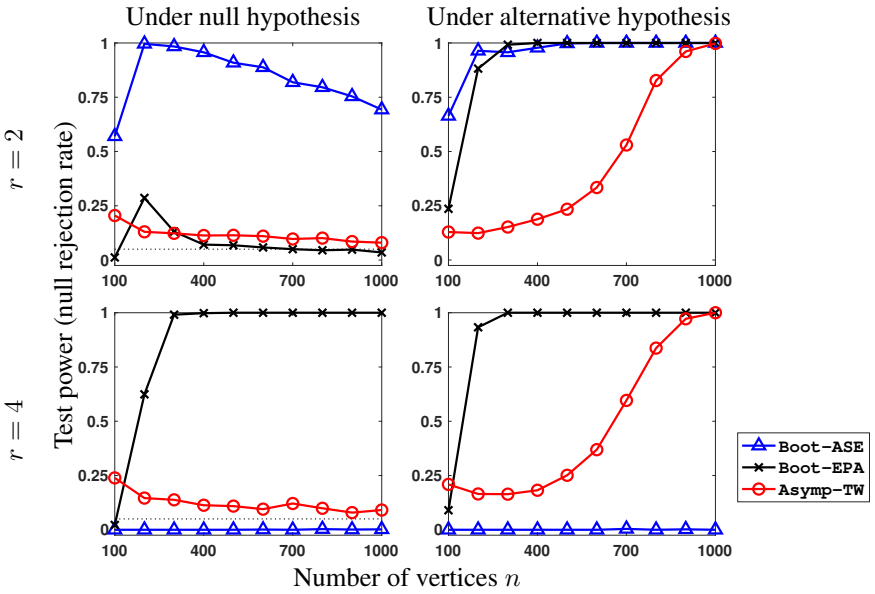


Figure 2: Power of different tests with increase number of vertices $n$, and for rank parameter $r = 2, 4$. The dotted line under null hypothesis corresponds to the significance level of 5%.

`Asymp-TW` are poor if the graphs are too sparse. Hence, `Boot-EPA` may be preferable for sparse graphs, but only if the rank is correctly known.

### 6.3 Qualitative results for testing real networks

We use the proposed asymptotic tests to analyse two real datasets. These experiments demonstrate that the proposed tests are applicable beyond the setting of IER graphs. In the first setup, we consider moderate sized graphs ($n = 178$) constructed by thresholding autocorrelation matrices of EEG recordings (Andrzejak et al., 2001, Dua and Taniskidou, 2017). The network construction is described Appendix C.2. Each group of networks corresponds to either epileptic seizure activity or four other resting states. In Tables 1–4 in Appendix C, we report the test powers and p-values for `Asymp-Normal` and `Asymp-TW`. We find that, except for one pair of resting states, networks for different groups can be distinguished by both tests. Further observations and discussions are also provided in the appendix.

We also study networks corresponding to peering information of autonomous systems, that is, graphs defined on the routers comprising the Internet with the edges representing *who-talks-to-whom* (Leskovec et al., 2005, Leskovec and Krevl, 2014). The information for $n = 11806$ systems was collected once a week for nine consecutive weeks, and two networks are available for each date based on two sets of information ($m = 2$). We run `Asymp-Normal` test for every pair of dates and report the p-values in Table 5 (Appendix C.3). It is interesting to observe that as the interval between two dates increase, the p-values decrease at an exponential rate, that is, the networks differ drastically according to our tests. We also conduct semi-synthetic experiments by randomly perturbing the networks, and study the performance of `Asymp-Normal` and `Asymp-TW` as the perturbations increase (see Figures 6–7). Since the networks are large and sparse, we perform the community detection step of `Asymp-TW` using BigClam (Yang and Leskovec, 2013) instead of spectral clustering. We infer that the limitation of `Asymp-TW` in sparse regime (observed in Figure 5) could possibly be caused by poor performance of standard spectral clustering in sparse regime.

## 7 Concluding remarks

In this work, we consider the two-sample testing problem for undirected unweighted graphs defined on a common vertex set. This problem finds application in various domains, and is often challenging due to unavailability of large number of samples (small $m$). We study the practicality of existing

theoretical tests, and propose two new tests based on asymptotics for large graphs (Thereoms 1 and 3). We perform numerical comparison of various tests, and also provide their Matlab implementations. In the $m > 1$ case, we find that `Boot-Spectral` is effective for $m \geq 6$, but `Asymp-Normal` is recommended for smaller $m$ since it is more reliable and requires less computation. For $m = 1$, we recommend `Asymp-TW` due to robustness to the rank parameter and computational advantage. For large sparse graphs, `Asymp-TW` should be used with a robust community detection step (BigClam).

One can certainly extend some of these tests to more general frameworks of graph testing. For instance, *directed graphs* can be tackled by modifying $T_{fro}$ such that the summation is over all $i, j$ and Theorem 1 would hold even in this case. For *weighted graphs*, Theorem 3 can be used if one modifies $C$ (9) by normalising with variance of $(A_G)_{ij} - (A_H)_{ij}$. Subsequently, these variances can be approximated again through block modelling. For $m > 1$, we believe that *unequal population sizes* can be handled by rescaling the matrices appropriately, but we have not verified this.

**Acknowledgements**

This work is supported by the German Research Foundation (Research Unit 1735) and the Institutional Strategy of the University of Tübingen (DFG, ZUK 63).

## References

T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley and Sons, 1984.

R. G. Andrzejak, K. Lehnertz, C. Rieke, F. Mormann, P. David, and C. E. Elger. Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64:061907, 2001.

E. Arias-Castro and N. Verzelen. Community detection in dense random networks. *Annals of Statistics*, 42(3):940–969, 2014.

D. S. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg. Hierarchical organization of human cortical networks in health and schizophrenia. *The Journal of Neuroscience*, 28(37):9239–9248, 2008.

A. C. Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941.

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

B. Bollobas, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 31(1):3–122, 2007.

K. M. Borgwardt, C. S. Ong, S. Schönauer, S. V. Vishwanathan, A. J. Smola, and H. P. Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(Suppl 1):i47–56, 2005.

F. Bornemann. On the numerical evaluation of distributions in random matrix theory. *Markov Processes and Related Fields*, 16:803–866, 2010.

R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: Implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8:37–49, 2008.

D. Dua and K. Taniskidou. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2017.

L. Erdős, H.-T. Yau, and J. Yin. Rigidity of eigenvalues of generalized Wigner matrices. *Advances in Mathematics*, 229(3):1435–1515, 2012.

D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg. Two-sample hypothesis testing for inhomogeneous random graphs. arXiv preprint (arXiv:1707.00833), 2017a.

D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg. Two-sample tests for large random graphs using network statistics. In *Conference on Learning Theory (COLT)*, 2017b.

C. E. Ginestet, A. P. Fournel, and A. Simmons. Statistical network analysis for functional MRI: Summary networks and group comparisons. *Frontiers in computational neuroscience*, 8(51): 10.3389/fncom.2014.00051, 2014.

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

O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, 1998.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–733, 2012.

D. R. Hyduke, N. E. Lewis, and B. Palsson. Analysis of omics data with genome-scale models of metabolism. *Molecular BioSystems*, 9(2):167–174, 2013.

R. Kondor and H. Pan. The multiscale Laplacian graph kernel. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

B. A. Landman, A. J. Huang, A. Gifford, D. S. Vikram, I. A. Lim, J. A. Farrell, J. A. Bogovic, J. Hua, M. Chen, S. Jarso, S. A. Smith, S. Joel, S. Mori, J. J. Pekar, P. B. Barker, J. L. Prince, and P. C. van Zijl. Multi-parametric neuroimaging reproducibility: A 3-T resource study. *Neuroimage*, 54(4): 2854–2866, 2011.

J. O. Lee and J. Yin. A necessary and sufficient condition for edge universality of Wigner matrices. *Duke Mathematical Journal*, 163(1):117–173, 2014.

J. Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1):401–424, 2016.

J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

L. Lovász. *Large networks and graph limits*. American Mathematical Society, 2012.

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

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

N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A semiparametric two-sample hypothesis testing problem for random graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, 2016.

M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23:1599–1630, 2017.

C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177:727–754, 1996.

J. Yang and J. Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM)*, pages 587–596, 2013.

B. Zhang, H. Li, R. B. Riggins, M. Zhan, J. Xuan, Z. Zhang, E. P. Hoffman, R. Clarke, and Y. Wang. Differential dependency network analysis to identify condition-specific topological changes in biological networks. *Bioinformatics*, 25(4):526–532, 2009.

**Supplementary material for the paper**

# Practical Methods for Graph Two-Sample Testing

published in NIPS-2018

Authors: Debarghya Ghoshdastidar, Ulrike von Luxburg
(University of Tübingen, MPI-IS Tübingen)

Here, we provide additional details such as proofs, description of tests, additional numerical results and discussions. Section A provides proofs for the theorems stated in the paper along with a corollary of Theorem 3. Section B provides detailed descriptions of all tests considered in our implementations, both existing tests as well as proposed ones. Section C provides additional numerical results, which we have referred to in the paper.

## A Proofs for results

In this section, we present the proofs for Theorems 1 and 3, which provide the theoretical foundations for the proposed tests `Asymp-Normal` and `Asymp-TW`, respectively.

### A.1 Proof of Theorem 1

For convenience, we assume $m$ is even. The extension to odd $m$ is straightforward. We also write $P, Q$ instead of $P^{(n)}, Q^{(n)}$ and define

$$\widehat{\mu}_{ij} = \left( \sum_{k \leq m/2} (A_{G_k})_{ij} - (A_{H_k})_{ij} \right) \left( \sum_{k > m/2} (A_{G_k})_{ij} - (A_{H_k})_{ij} \right),$$

$$\widehat{s}^2_{ij} = \left( \sum_{k \leq m/2} (A_{G_k})_{ij} + (A_{H_k})_{ij} \right) \left( \sum_{k > m/2} (A_{G_k})_{ij} + (A_{H_k})_{ij} \right),$$

$$\widehat{\mu} = \sum_{i<j} \widehat{\mu}_{ij}, \qquad \text{and} \qquad \widehat{s} = \sqrt{\sum_{i<j} \widehat{s}^2_{ij}}.$$

Also let $\mu = \mathbb{E}[\widehat{\mu}] = \frac{m^2}{8}\|P - Q\|_F^2$, $s^2 = \mathbb{E}[\widehat{s}^2] = \frac{m^2}{8}\|P + Q\|_F^2$, and $\sigma^2 = \sum_{i<j} \mathrm{Var}(\widehat{\mu}_{ij})$.

Under the null hypothesis, that is $P = Q$, $\{\widehat{\mu}_{ij} : i < j\}$ are centred mutually independent random variables, and hence, due to the central limit theorem, we can claim that $\frac{\widehat{\mu}}{\sigma}$ converges to a standard normal random variable as $n \to \infty$. The rate of convergence is given by the Berry-Esseen theorem (Berry, 1941) as

$$\sup_x \left| F_{\widehat{\mu}/\sigma}(x) - \Phi(x) \right| \leq \frac{10}{\sigma^3} \sum_{i<j} \mathbb{E}\left[ |\widehat{\mu}_{ij}|^3 \right],$$

where $F_{\widehat{\mu}/\sigma}(\cdot)$ is the distribution function for $\frac{\widehat{\mu}}{\sigma}$. Recall our assumption that the entries are bounded away from 1. Let $\max_{ij} P_{ij} \leq 1 - \delta$ for some $\delta > 0$. Observe that $\widehat{\mu}_{ij}$ is product of two i.i.d. random variables, where each of them is a difference of two binomials. Hence, under $\mathcal{H}_0$, we can compute

$$\sigma^2 = \sum_{i<j} \left( \frac{m}{2} 2P_{ij}(1 - P_{ij}) \right)^2 \geq \frac{m^2\delta^2}{2} \|P\|_F^2,$$

and by using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\left[ |\widehat{\mu}_{ij}|^3 \right] &\leq \sqrt{\mathbb{E}\left[\widehat{\mu}^2_{ij}\right] \mathbb{E}\left[\widehat{\mu}^4_{ij}\right]} \\
&= mP_{ij}(1 - P_{ij}) \left( mP_{ij}(1 - P_{ij})^3 + \frac{m}{2}\left(\frac{m}{2} - 1\right) 4P^2_{ij}(1 - P_{ij})^2 \right) \\
&\leq m^2 P^2_{ij} + m^3 P^3_{ij} \leq 2m^3 P^2_{ij}.
\end{aligned}$$

Hence, the Berry-Esseen bound can be written as

$$\sup_x \left|F_{\widehat{\mu}/\sigma}(x) - \Phi(x)\right| \leq 20\sqrt{2}\frac{m^3\|P\|_F^2}{m^3\delta^3\|P\|_F^3} = o_n(1)$$

since $\|P\|_F = \omega_n(1)$. We now compute the probability of type-I error in the following way:

$$\mathbb{P}\big(T_{fro} \notin [-t_\alpha, t_\alpha]\big) = \mathbb{P}\left(\frac{|\widehat{\mu}|}{\widehat{s}} > t_\alpha\right) \leq \mathbb{P}\left(\frac{|\widehat{\mu}|}{\sigma} > (1-\epsilon)t_\alpha\right) + \mathbb{P}\left(\widehat{s}^2 < (1-\epsilon)^2\sigma^2\right) \tag{13}$$

for any $\epsilon \in (0, \frac{1}{2})$. Using the Berry-Esseen bound, we bound the first term as

$$\begin{aligned}
\mathbb{P}\left(\frac{|\widehat{\mu}|}{\sigma} > (1-\epsilon)t_\alpha\right) &= 2\big(1 - \Phi((1-\epsilon)t_\alpha)\big) + 2\left|F_{\widehat{\mu}/\sigma}((1-\epsilon)t_\alpha) - \Phi((1-\epsilon)t_\alpha)\right| \\
&= \alpha + 2\big(\Phi(t_\alpha) - \Phi((1-\epsilon)t_\alpha)\big) + o_n(1) \\
&\leq \alpha + \epsilon t_\alpha\sqrt{\frac{2}{\pi}}\exp\left(-\frac{t_\alpha^2}{8}\right) + o_n(1)
\end{aligned}$$

where we use $\epsilon \leq \frac{1}{2}$ in the last step. Taking $\epsilon = \|P\|_F^{-1/2}$ leads to a bound $\alpha + o_n(1)$.

We now deal with the second term in (13). Observe that $\sigma^2 \leq \frac{m^2}{2}\|P\|_F^2 \leq s^2$. Hence, we have

$$\begin{aligned}
\mathbb{P}\left(\widehat{s}^2 < (1-\epsilon)^2\sigma^2\right) &\leq \mathbb{P}\left(\widehat{s}^2 < (1-\epsilon)s^2\right) \\
&= \mathbb{P}\left(s^2 - \widehat{s}^2 > \epsilon s^2\right) \leq \frac{\mathrm{Var}(\widehat{s}^2)}{\epsilon^2 s^4}
\end{aligned}$$

by the Chebyshev inequality. We can compute the variance term for any $P, Q$ as

$$\begin{aligned}
&\mathrm{Var}(\widehat{s}^2) \\
&= \sum_{i<j}\frac{m^2}{4}\left(P_{ij}(1-P_{ij}) + Q_{ij}(1-Q_{ij})\right)^2 + \frac{m^3}{4}(P_{ij}+Q_{ij})^2\left(P_{ij}(1-P_{ij}) + Q_{ij}(1-Q_{ij})\right)
\end{aligned} \tag{14}$$

In particular, under $\mathcal{H}_0$, $\mathrm{Var}(\widehat{s}^2) \leq 2m^3\|P\|_F^2$. Using this, the Chebyshev bound is smaller than $\frac{4}{m\epsilon^2\|P\|_F^2} = o_n(1)$ for $\epsilon = \|P\|_F^{-1/2}$. Hence, we obtained the claimed type-I error bound.

For the type-II error rate, we consider the stated separation condition in the form $\frac{m\|P-Q\|_F^2}{\|P+Q\|_F} = \omega_n(1)$. We can bound the error probability as

$$\mathbb{P}\big(T_{fro} \in [-t_\alpha, t_\alpha]\big) \leq \mathbb{P}\left(\frac{|\widehat{\mu}|}{s} \leq 2t_\alpha\right) + \mathbb{P}\left(\widehat{s}^2 \geq 4s^2\right).$$

For the second term, we use the Chebyshev inequality as above to show that the probability is $o_n(1)$ since $\|P+Q\|_F = \omega_n(1)$. For the first term, observe that we have $\frac{\mu}{s} = \omega_n(1)$ under the separation condition, and hence for any fixed $\alpha$, we have $2t_\alpha \leq \frac{\mu}{2s}$ for large enough $n$. So,

$$\mathbb{P}\left(\frac{|\widehat{\mu}|}{s} \leq 2\tau_\alpha\right) \leq \mathbb{P}\left(\frac{\widehat{\mu}}{s} \leq \frac{\mu}{2s}\right) \leq \frac{4\mathrm{Var}(\widehat{\mu})}{\mu^2}.$$

One can compute $\mathrm{Var}(\widehat{\mu})$ similar to (14) to obtain

$$\begin{aligned}
\mathrm{Var}(\widehat{\mu}) &\leq \sum_{i<j}\frac{m^2}{4}(P_{ij}+Q_{ij})^2 + \frac{m^3}{4}(P_{ij}-Q_{ij})^2(P_{ij}+Q_{ij}) \\
&\leq \frac{m^2}{8}\|P+Q\|_F^2 + \frac{m^3}{8}\|P-Q\|_F^2\|P+Q\|_F,
\end{aligned}$$

where the second inequality follows from use of the Cauchy-Schwarz inequality followed by the observation that $\ell_4$-norm is smaller than $\ell_2$-norm. Hence, the error probability is bounded as

$$\mathbb{P}\big(T_{fro} \in [-t_\alpha, t_\alpha]\big) \leq 32\frac{m^2\|P+Q\|_F^2 + m^3\|P-Q\|_F^2\|P+Q\|_F}{m^4\|P-Q\|_F^4} + o_n(1) = o_n(1)$$

under the assumed separation. Hence, the claim.

### A.2 Proof of Theorem 3

We first derive the asymptotic distribution under the null hypothesis. This part is similar to the proof of Lemma A.1 in Lei (2016). Observe that under $\mathcal{H}_0$, $C$ in (9) is a symmetric random matrix, whose entries above the diagonal are independent with mean zero and variance $\frac{1}{n-1}$. Now, let $D$ be a symmetric random matrix with zero diagonal, whose entries above the diagonal are i.i.d. normal with mean zero and variance $\frac{1}{n-1}$. Due to the results of Erdős et al. (2012), we know that $\lambda_1(C)$ and $\lambda_1(D)$ have the same limiting distribution. Lee and Yin (2014) show that $n^{2/3}(\lambda_1(D) - 2) \to TW_1$ as $n \to \infty$, and hence the same conclusion holds for $n^{2/3}(\lambda_1(C) - 2)$. The corresponding result for $-\lambda_n(C)$ can be proved by considering the matrix $-C$. Based on this asymptotic result, we have

$$\mathbb{P}\left(n^{2/3}(\lambda_1(C) - 2) > \tau_\alpha\right) = \frac{\alpha}{2} + o_n(1), \text{ and}$$
$$\mathbb{P}\left(n^{2/3}(-\lambda_n(C) - 2) > \tau_\alpha\right) = \frac{\alpha}{2} + o_n(1),$$

where $\tau_\alpha$ is the $\frac{\alpha}{2}$ upper quantile of the $TW_1$ distribution. Since, $\|C\|_2 = \max\{\lambda_1(C), -\lambda_n(C)\}$, an union bound leads to the stated conclusion under the null hypothesis.

Under the alternative hypothesis, one can see that $\mathbb{E}[C]$ is a re-scaled version of $P - Q$ with each entry being scaled by normalising term of $\sqrt{(n-1)(P_{ij}(1-P_{ij}) + Q_{ij}(1-Q_{ij}))}$ (we drop the superscript $n$ for convenience). Under the stated separation condition on $\|\mathbb{E}[C]\|_2$, it is easy to see that $n^{2/3}(\|C\|_2 - 2) \to \infty$ with high probability. So, the probability of the test statistic being smaller than $\tau_\alpha$ is $o_n(1)$. To be precise, we decompose $C$ as $C = \mathbb{E}[C] + (C - \mathbb{E}[C])$, and using Weyl's inequality, we can write

$$\|C\|_2 \geq \|\mathbb{E}[C]\|_2 - \|C - \mathbb{E}[C]\|_2 \geq \|\mathbb{E}[C]\|_2 - \left(2 + n^{-2/3}\tau_\beta\right)$$

with probability at most $\beta + o_n(1)$. The second inequality follows by noting that $(C - \mathbb{E}[C])$ is a mean zero matrix whose spectral norm can be bounded using the arguments stated under the null hypothesis. Hence, $n^{2/3}(\|C\|_2 - 2) \geq n^{2/3}(\|\mathbb{E}[C]\|_2 - 4) - \tau_\beta$ with probability $\beta + o_n(1)$. We set $\tau_\beta = n^{2/3}(\|\mathbb{E}[C]\|_2 - 4) - \tau_\alpha$, and observe that $\tau_\beta = \omega_n(1)$, that is $\beta = o_n(1)$, if $\|\mathbb{E}[C]\|_2 \geq 4 + \omega_n(n^{-2/3})$.

### A.3 Theorem 3 for stochastic block models

We state the following corollary, which provides an understanding of the condition on $\mathbb{E}[C]$ in Theorem 3 under a block model assumption.

**Corollary 4.** *Assume that $P^{(n)}, Q^{(n)}$ correspond to stochastic block models with at most $r_n$ communities, and let $\rho_n = \max_{ij}\left\{P^{(n)}_{ij}, Q^{(n)}_{ij}\right\}$. If $\left\|P^{(n)} - Q^{(n)}\right\|_F^2 = \omega_n\left(nr_n^2\rho_n\right)$, then*

$$\mathbb{P}\left(n^{2/3}(\|C\|_2 - 2) \leq \tau_\alpha\right) = o_n(1). \tag{15}$$

One can observe that if $r_n$ is bounded by a constant and all entries of $P^{(n)}, Q^{(n)}$ are of the same order (same as $\rho_n$), then the above separation condition is similar to the one stated in Theorem 1.

*Proof.* The claim would follow if we show that under the stated separation, the condition on $\mathbb{E}[C]$ used in Theorem 3 holds. In fact, we show that in the present case, $\|\mathbb{E}[C]\|_2 = \omega_n(1)$. For convenience, we simply write $P, Q$ and define $R_{ij} = \sqrt{(n-1)(P_{ij}(1-P_{ij}) + Q_{ij}(1-Q_{ij}))} \leq \sqrt{2n\rho_n}$. Note that

$$\mathbb{E}[C_{ij}] = \frac{P_{ij} - Q_{ij}}{R_{ij}},$$

and hence, $\mathbb{E}[C]$ has a block structure with at most $r_n^2$ blocks (ignoring that the diagonal is zero). Thus, there is a diagonal matrix $\Lambda$ such that $\Lambda + \mathbb{E}[C]$ has rank at most $r_n^2$. Note that the diagonal entries of $\Lambda$ are same as the diagonal blocks of $C$, and so, $\|\Lambda\|_2 \leq \max_{ij}\frac{|P_{ij} - Q_{ij}|}{R_{ij}} \leq 2\sqrt{\frac{\rho_n}{(n-1)(1-\rho_n)}} = o_n(1)$

assuming that $\rho_n$ is bounded away from 1. Hence, we can write

$$\|\mathbb{E}[C]\|_2 \geq \|\Lambda + \mathbb{E}[C]\|_2 - \|\Lambda\|_2 \geq \frac{1}{r_n}\|\Lambda + \mathbb{E}[C]\|_F - o_n(1)$$
$$\geq \frac{1}{r_n}\|\mathbb{E}[C]\|_F - o_n(1) \geq \frac{\|P-Q\|_F}{r_n\sqrt{2n\rho_n}} - o_n(1),$$

which is $\omega_n(1)$ under the stated condition. For the second inequality, we use the relation between spectral and Frobenius norms of a matrix with rank $r_n^2$. Finally, Theorem 3 leads to the result. □

# B Detailed description of tests

In this section, we describe all the tests discussed in this paper. First, we provide description of the asymptotic tests, which include the tests `Asymp-Normal` and `Asymp-TW` proposed in this paper, as well as the large-sample test `Asymp-Chi2`. We next describe the bootstrapped tests `Boot-Spectral` and `Boot-Frobenius`, which are based on approximating the null distribution by randomly permuting the group assignments of the graphs. Tang et al. (2016) provide an algorithmic description of `Boot-ASE`. For completeness, we include this description along with that of `Boot-EPA`, which also generates bootstrap samples based on a low rank approximation of population adjacency. Throughout this section, we refer to the null hypothesis $\mathcal{H}_0$ as the hypotheses that both graphs (or graph populations) have the same population adjacency.

## B.1 Asymptotic tests

We first describe the `Asymp-Normal` test below. In addition to accepting or rejecting the null hypothesis, we also present how to compute the *p-value*, which is defined as the probability that the null hypothesis is true. This is often useful to quantify the amount of dissimilarity between two populations. We use the standard rule of rejecting the null hypothesis when p-value is less than the prescribed significance level $\alpha$. Note that in `Asymp-Normal`, the p-value involves a factor of 2 to take into account both the upper and the lower tail probabilities.

---

**Test** `Asymp-Normal`

---

**Input:** Graphs $G_1, \ldots, G_m$ and $H_1, \ldots, H_m$ defined on a common vertex set $V$, where $m > 1$; Significance level $\alpha$.

1: Compute $T_{fro}$ as shown in (4).

2: p-value $= 2\big(1 - \Phi(-|T_{fro}|)\big)$, where $\Phi$ is the standard normal distribution function.

**Output:** Reject the null hypothesis if p-value $\leq \alpha$.

---

The `Asymp-Chi2` test is listed below. For convenience, we write $T_{\chi^2} = \sum_{i<j} \frac{\tilde{\mu}_{ij}^2}{\tilde{\sigma}_{ij}^2}$, where $\tilde{\mu}_{ij}^2$ and $\tilde{\sigma}_{ij}^2$ denote the numerator and denominator of each term in the summation (2). This notation corresponds to the fact that $\tilde{\mu}_{ij}$ is the sample mean difference for entry $(i,j)$, and $\tilde{\sigma}_{ij}^2$ is an estimate of the variance of $\tilde{\mu}_{ij}$. We note that for sparse graphs and small $m$, the summation in (2) may have terms of the form $\frac{0}{0}$. Hence, we sum only over the set of edges in $\mathcal{C}$ defined below.

---

**Test** `Asymp-Chi2`

---

**Input:** Graphs $G_1, \ldots, G_m$ and $H_1, \ldots, H_m$, where $m > 1$; Significance level $\alpha$.

1: Let $\mathcal{C} = \{(i,j) : i < j,\ \tilde{\mu}_{ij} \neq 0 \text{ or } \tilde{\sigma}_{ij} \neq 0\}$, where $\tilde{\mu}_{ij}$, $\tilde{\sigma}_{ij}$ are defined above.

2: Compute $T_{\chi^2}$ similar to (2), but sum only over $(i,j) \in \mathcal{C}$.

3: p-value $= 1 - F_{\chi^2}\left(T_{\chi^2}, \frac{n(n-1)}{2}\right)$, where $F_{\chi^2}(\cdot, \nu)$ is the $\chi^2$-distribution function with degree of freedom $\nu$.

**Output:** Reject the null hypothesis if p-value $\leq \alpha$.

---

We now described `Asymp-TW`, which is the proposed asymptotic test for testing between two given graphs $G$ and $H$ (that is, $m = 1$). As noted in the main paper, this test uses a block model approximation to compute the matrices $\widetilde{P}, \widetilde{Q}$. In the following description, we assume that a partition

of $V$ into $V_1, \ldots, V_r$ is provided as input to the test. For simplicity, we assume that the same partitioning is used for both graphs, but this is not a necessity. In our implementations, we use normalised spectral clustering (Ng et al., 2002) to compute the partition from the average of the two adjacency matrices. A minor difference here is that we use the dominant singular vectors of the normalised adjacency instead of the dominant eigenvectors. This modification is made since the networks could be either homophilic (communities are highly connected) or heterophilic (inter-community links are more frequent as in a bi-partite graph). We also provide an option to externally provide the communities. We use this feature for the real data from Stanford network collection, where we pre-compute the community structure using BigClam (Yang and Leskovec, 2013). From the test statistic $T_{TW}$, we compute the p-value by using available table of distribution function for Tracy-Widom law.[2] The factor of 2 is due to the fact that only the extreme eigenvalues are known to follow the $TW_1$ distribution, and hence, we need union bound for $\|\widetilde{C}\|_2 = \max\left\{\lambda_1(\widetilde{C}), -\lambda_n(\widetilde{C})\right\}$.

**Test** `Asymp-TW`

**Input:** Graphs $G, H$ defined on vertex set $V$; Partition of $V$ into $V_1, \ldots, V_r$; Significance level $\alpha$.

**for all** $V_k$ **do**
  **for all** $i, j \in V_k, i \neq j$ **do**
    Let $\widetilde{P}_{ij} = \frac{2}{|V_k|(|V_k|-1)} \sum_{i',j' \in V_k : i' < j'} (A_G)_{ij}$ and $\widetilde{Q}_{ij} = \frac{2}{|V_k|(|V_k|-1)} \sum_{i',j' \in V_k : i' < j'} (A_H)_{ij}$.
  **end for**
**end for**
**for all** $V_k, V_l, k \neq l$ **do**
  **for all** $i \in V_k, j \in V_l$ **do**
    Compute $\widetilde{P}_{ij} = \frac{1}{|V_k||V_l|} \sum_{i' \in V_k, j' \in V_l} (A_G)_{ij}$ and $\widetilde{Q}_{ij} = \frac{1}{|V_k||V_l|} \sum_{i' \in V_k, j' \in V_l} (A_H)_{ij}$.
  **end for**
**end for**
Compute $\widetilde{C}$ and $T_{TW}$ as in (12).
p-value $= 2\big(1 - F_{TW_1}(T_{TW})\big)$, where $F_{TW_1}$ is the distribution function for Tracy-Widom law.

**Output:** Reject the null hypothesis if p-value $\leq \alpha$.

### B.2 Bootstrap tests

We begin with the description of `Boot-Spectral` and `Boot-Frobenius`. We present both tests together since they follow the same bootstrapping procedure, and only differ in terms of the test statistic. The differences of `Boot-Frobenius` from `Boot-Spectral` are noted in parentheses.

**Test** `Boot-Spectral` **(or** `Boot-Frobenius`**)**

**Input:** Graphs $G_1, \ldots, G_m$ and $H_1, \ldots, H_m$, where $m > 1$; Significance level $\alpha$; Number of bootstraps $b$.

Let $T = T_{spec}$ as computed in (3) (or $T = T_{fro}$ in (4)).
**for** $i = 1$ **to** $b$ **do**
  Randomly split $\{G_1, \ldots, G_m, H_1, \ldots, H_m\}$ into two populations of equal size.
  Let $T_i$ be the spectral norm statistic (3) for this split (or Frobenius norm statistic (4)).
**end for**
p-value $= \frac{1}{b}\left(\left|\{i : T_i \geq T\}\right| + 0.5\right)$, where 0.5 is added for continuity correction.

**Output:** Reject the null hypothesis if p-value $\leq \alpha$.

Finally, we present the tests `Boot-ASE` and `Boot-EPA` based on adjacency spectral embedding (ASE) and estimated population adjacency (EPA), respectively. The differences of `Boot-EPA` from `Boot-ASE` are noted in parentheses. Note that these tests compute two approximations of the null distribution — one based on pairs of graphs generated from $\widehat{P}$, and other based on graph pairs generated from $\widehat{Q}$. The p-value is finally computed to ensure that the null is rejected only when the test statistic is in the upper $\alpha$-quantile for both approximate distributions.

[2]The table, based on Bornemann (2010), was obtained from `http://www.wisdom.weizmann.ac.il/~nadler/Wishart_Ratio_Trace/TW_ratio.html`. This limited table can provide $F_{TW_1}(\cdot) \leq 0.9998$.

**Test** `Boot-ASE` **(or** `Boot-EPA`**)**

**Input:** Graphs $G$ and $H$; Significance level $\alpha$; Number of bootstraps $b$.

Let $X_G$ and $\widehat{P}$ be the ASE and EPA for graph $G$, respectively (as described in Section 5).
Let $X_H$ and $\widehat{Q}$ be the ASE and EPA for graph $H$, respectively.
Compute $T = T_{ASE}$ as in (7) (or $T = T_{EPA}$ in (8)).
**for** $i = 1$ **to** $b$ **do**
    Generate $G_1, G_2 \sim_{\text{iid}} \text{IER}(\widehat{P})$.
    Let $T_i$ be the ASE statistic (7) between $G_1, G_2$ (or EPA statistic (8)).
**end for**
Compute $p = \frac{1}{b}\left(\left|\{i : T_i \geq T\}\right| + 0.5\right)$, where 0.5 is added for continuity correction.
**for** $i = 1$ **to** $b$ **do**
    Generate $H_1, H_2 \sim_{\text{iid}} \text{IER}(\widehat{Q})$.
    Let $T'_i$ be the ASE statistic (7) between $H_1, H_2$ (or EPA statistic (8)).
**end for**
Compute $p' = \frac{1}{b}\left(\left|\{i : T'_i \geq T\}\right| + 0.5\right)$.
p-value $= \max\{p, p'\}$.

**Output:** Reject the null hypothesis if p-value $\leq \alpha$.

# C Additional numerical results and discussions

Here, we provide additional results along with further details for the experiments with real data.

## C.1 Further simulations for random graphs

In this section, we present the figures related to experiments on block models, which we have referred to in the main paper. We have earlier noted that `Asymp-Chi2` has an erratic behaviour for small $m$. This is not surprising since the variance estimates used in (2) are not reliable for small $m$, particularly when the graphs are sparse. We demonstrate this effect even for slightly larger $m$ by comparing `Asymp-Chi2` and `Asymp-Normal` for $m \in \{10, 20, 50, 100, 200\}$. The graph sizes are kept relatively small $n \in \{50, 100, 150, 200\}$. The models are same as the ones used in the experiment of Figure 1.

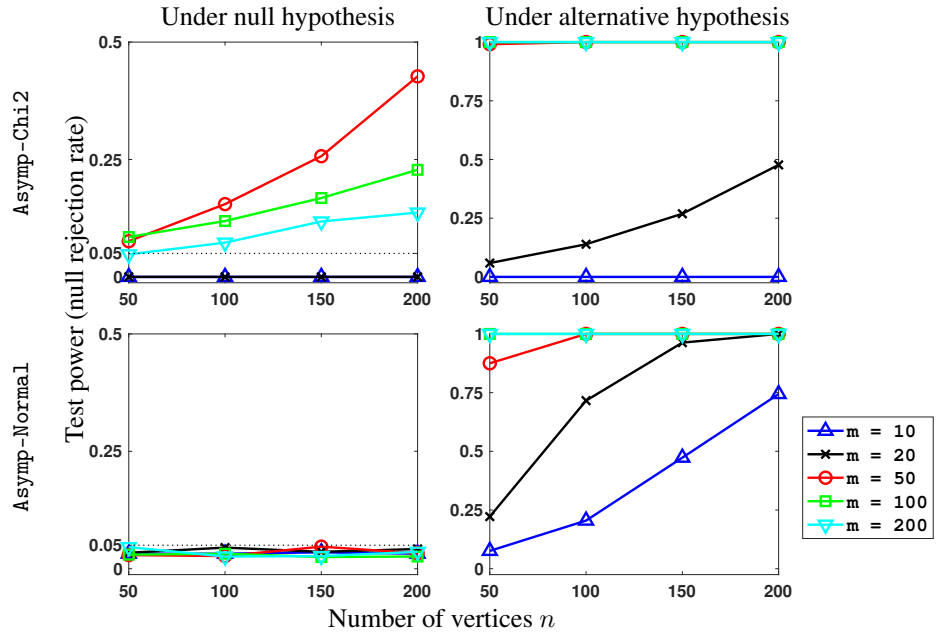


Figure 3: Power of the asymptotic tests for different values of graph size $n$ and population size $m$. Each row corresponds to a particular test.

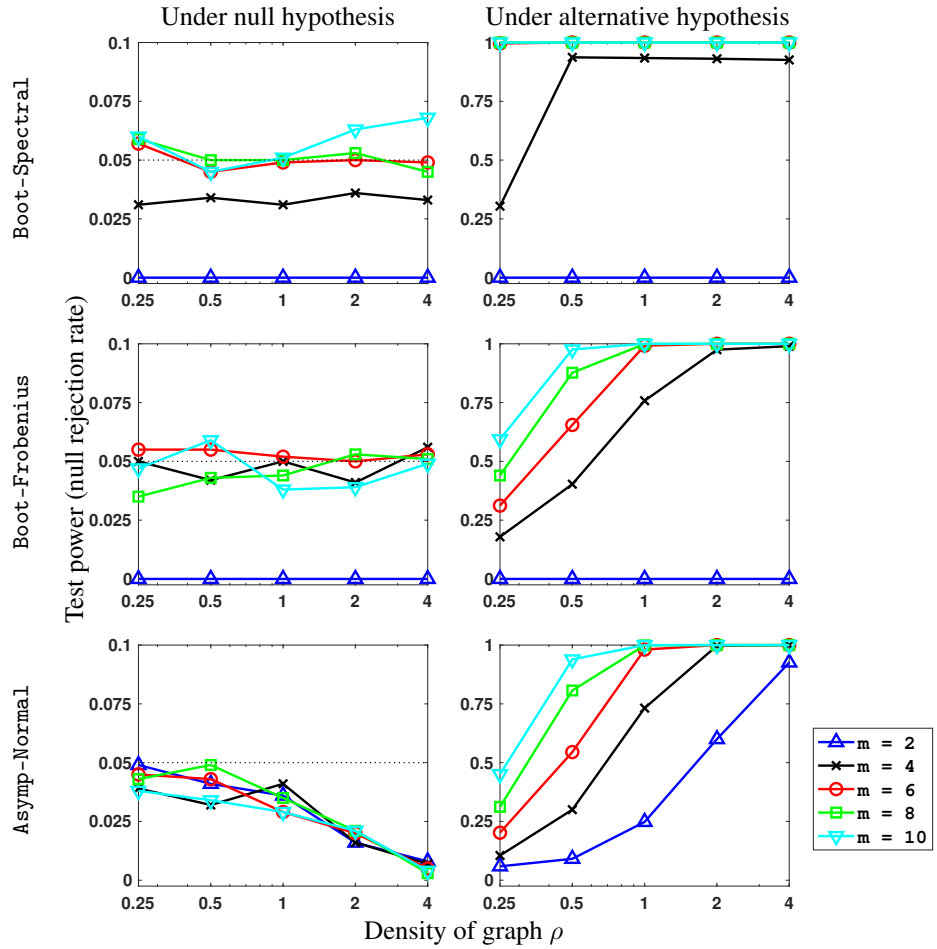


Figure 4: Power of different tests for varying levels of sparsity $\rho$ (larger $\rho$ implies denser graphs), and for different values of population size $m$. Each row corresponds to a particular test.

The result, plotted in Figure 3, reveals the undesirable behaviour of `Asymp-Chi2` as the test always has zero rejection rate for $m = 10$. For $m \geq 50$, the test power under alternative hypothesis is 1, but the rejection under null increases with $n$. In particular, rejection rate under null is less than significance level only for $m = 200$ and $n = 50$. Thus, `Asymp-Chi2` is reliable only for $m \gg n$. On the other hand, both Figures 1 and 3 confirm our theoretical observation that the behaviour of `Asymp-Normal` under $\mathcal{H}_0$ does not change with $m$, while its power under $\mathcal{H}_1$ improves for larger $m$.

Figure 4 corresponds to our study related to varying levels of graph sparsity. In this case, the models for $P^{(n)}$ and $Q^{(n)}$ are stochastic block models with same two communities. For $P^{(n)}$, within-class edge probability is $\rho p$ and across-class probability is $\rho q$. We define $Q^{(n)}$ such that the within-class edge probability is $\rho(p + \epsilon)$. Thus, this setting is identical to previous case of Figure 1 for $\rho = 1$. In Figure 4, we fix $n = 500$ and show the rejection rates of the tests for varying sample size $m$ and density $\rho$. The key conclusions are given in the main paper. Additionally, we note the effect of normal dominance in case of `Asymp-Normal`. Recall that $T_{fro}$ does not converge to the normal distribution, but it is dominated by a standard normal random variable. Thus, our threshold for rejection is actually higher than the $\frac{\alpha}{2}$-upper quantile of true asymptotic distribution of $T_{fro}$. This effect is pronounced for dense graphs, where the rejection rate under null is much smaller than the pre-fixed 5% level.

We present a similar study on the effect of sparsity in the case $m = 1$. The results in Figure 5 are based on the above setup, where we have $m = 1$ and vary the the graph size $n$ and the density parameter $\rho$. In this experiment, we use the true rank $r = 2$. This provides an advantage to the bootstrap tests since we observe in Figure 2 that these tests fail when approximation based on a different rank is used. We note that `Boot-ASE` has a high rejection rate under both $\mathcal{H}_0$ and $\mathcal{H}_1$. The rejection rate under $\mathcal{H}_0$ is

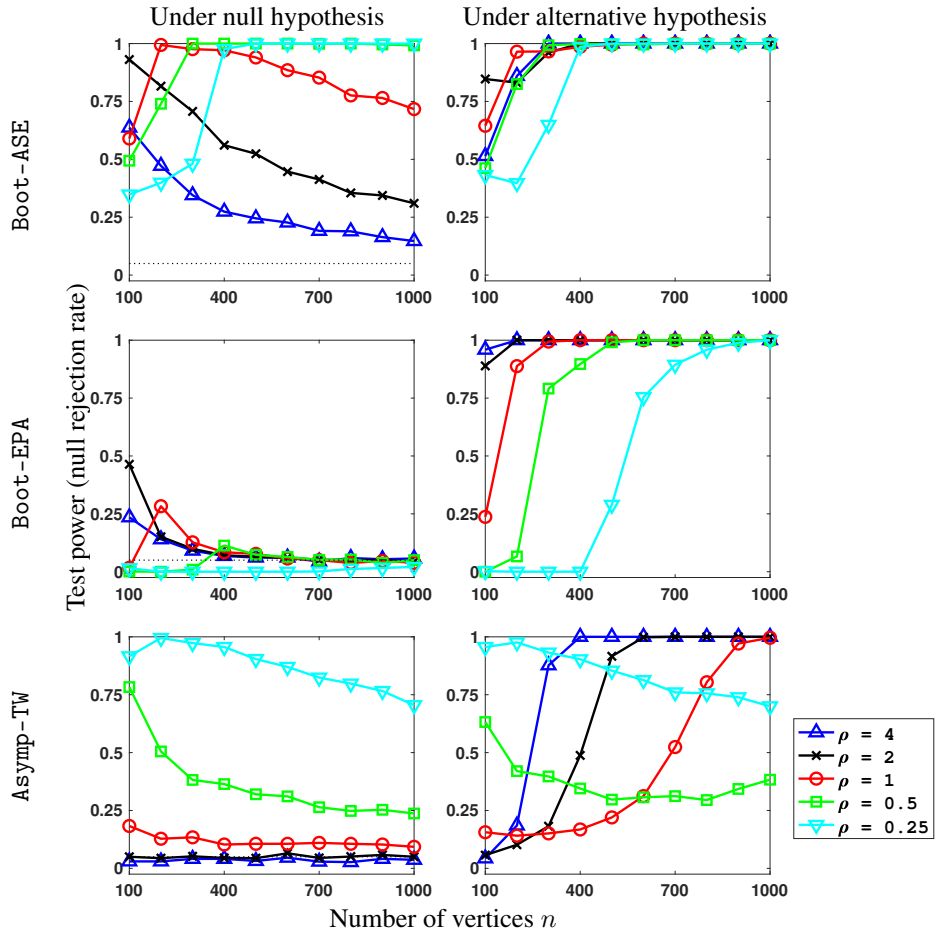


Figure 5: Power of different tests with increase number of vertices $n$, and for different levels of sparsity $\rho$. Each row corresponds to a particular test.

smaller for dense graphs, but still above the desired 5% level. For sparse graphs $\rho < 1$, this test is not reliable. On the other hand, `Boot-EPA` performs quite well for both sparse and dense graphs although it uses the same bootstrapping principle. Hence, we may conclude that the test statistic $T_{EPA}$, which was previously not used in the testing literature, is a more useful test statistic. The asymptotic test `Asymp-TW` works well for dense graphs $\rho \geq 1$, but is not reliable in the sparse regime. There can be two potential reasons for this: (i) the approximation of normalisation terms in (9) using $\widetilde{P}$ and $\widetilde{Q}$ is poor in the sparse regime; or (ii) the use of standard spectral clustering for community detection fails in sparse graphs. We believe that the latter reason is more probable since, in a later experiment with sparse real networks, we observe desirable performance from `Asymp-TW` when the community detection is done using BigClam (Yang and Leskovec, 2013).

### C.2 Experiments with EEG recordings of epileptic seizure

In this section, we describe our experiments with networks constructed from EEG recordings of patients with epileptic seizure (Andrzejak et al., 2001). We obtained the data from Dua and Taniskidou (2017), where each EEG recording is divided into several one-second snapshots containing 178 time points ($n = 178$). There are a total of 11500 snapshots available that are classified into five groups:
**Group-1:** Recording of seizure activity;
**Group-2:** Recording of an area with tumour;
**Group-3:** Recording of a healthy brain area;
**Group-4:** Recording of patient with eyes open;
**Group-5:** Recording of patient with eyes closed.

Table 1: Power of `Asymp-Normal` for EEG correlation networks.

| | **G1.1** | **G1.2** | **G2.1** | **G2.2** | **G3.1** | **G3.2** | **G4.1** | **G4.2** | **G5.1** | **G5.2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **G1.1** | 0 | 0.011 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **G1.2** | 0.011 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **G2.1** | 1 | 1 | 0 | 0.003 | 0.009 | 0.008 | 1 | 1 | 1 | 1 |
| **G2.2** | 1 | 1 | 0.003 | 0 | 0.009 | 0.005 | 1 | 1 | 1 | 1 |
| **G3.1** | 1 | 1 | 0.009 | 0.009 | 0 | 0 | 1 | 1 | 1 | 1 |
| **G3.2** | 1 | 1 | 0.008 | 0.005 | 0 | 0 | 1 | 1 | 1 | 1 |
| **G4.1** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| **G4.2** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| **G5.1** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.010 |
| **G5.2** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.010 | 0 |

In our experiments, we construct networks by thresholding the autocorrelation matrices of the EEG snapshots. The reason for considering such networks is due to their ubiquity in bioinformatics and neuroscience, where most networks are typically derived from correlations or covariances. Moreover, through this setup, we also establish that though the proposed tests are theoretically analysed for edge-independent graphs, they can also be used for other types of networks.

We randomly split each class into four parts of equal size, and compute autocorrelation matrices from the snapshots in each part. Unweighted graphs are obtained by retaining only the largest 10% of correlations (total of 20 graphs). For `Asymp-Normal` test, two graphs are needed for each population. Hence, for each class-$i$, we create two sub-groups **G$i$.1** and **G$i$.2**, each with two networks. We subsequently test between every pair of the 10 sub-groups — **G$i$.1** vs. **G$i$.2** is an instance of null hypothesis while every other pair is an instance of alternative hypothesis. For `Asymp-TW`, we only use the first graph in the sub-group for testing and use $r = 10$ communities for approximation. We run the above setup for 1000 independent trials (the randomness is induced by the splits of the classes during network construction) and report the powers of both tests in Tables 1 and 2.

Table 1 shows that for **G$i$.1** vs. **G$i$.2**, the null hypothesis is nearly always accepted by `Asymp-Normal` (rejection rate less than 1.1%). In other cases, the rejection is 100% except for **G2.**$x$ vs. **G3.**$x$ which shows that these two classes have identical behaviour. Table 2 shows that `Asymp-TW` arrives at mostly similar conclusions, but in several cases of alternative hypothesis the power can be much smaller than 1. This is not surprising since the problem is harder for $m = 1$. We note that the authors of the dataset also do not claim that the various rest states can be distinguished, and state that the data is

Table 2: Power of `Asymp-TW` for EEG correlation networks.

| | **G1.1** | **G1.2** | **G2.1** | **G2.2** | **G3.1** | **G3.2** | **G4.1** | **G4.2** | **G5.1** | **G5.2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **G1.1** | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **G1.2** | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **G2.1** | 1 | 1 | 0 | 0.002 | 0 | 0.001 | 1 | 1 | 0.243 | 0.260 |
| **G2.2** | 1 | 1 | 0.002 | 0 | 0 | 0.001 | 1 | 1 | 0.247 | 0.251 |
| **G3.1** | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0.234 | 0.245 |
| **G3.2** | 1 | 1 | 0.001 | 0.001 | 0 | 0 | 1 | 1 | 0.243 | 0.258 |
| **G4.1** | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.029 | 0.699 | 0.664 |
| **G4.2** | 1 | 1 | 1 | 1 | 1 | 1 | 0.029 | 0 | 0.647 | 0.667 |
| **G5.1** | 1 | 1 | 0.243 | 0.247 | 0.234 | 0.243 | 0.699 | 0.647 | 0 | 0.049 |
| **G5.2** | 1 | 1 | 0.260 | 0.251 | 0.245 | 0.258 | 0.664 | 0.667 | 0.049 | 0 |

Table 3: Negative logarithm of p-value (averaged over 1000 runs) computed by `Asymp-Normal` for EEG correlation networks.

| | **G1.1** | **G1.2** | **G2.1** | **G2.2** | **G3.1** | **G3.2** | **G4.1** | **G4.2** | **G5.1** | **G5.2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **G1.1** | 0 | 0.7 | 47.3 | 47.5 | 63.6 | 63.5 | 176.2 | 176.1 | 37.6 | 37.5 |
| **G1.2** | 0.7 | 0 | 47.4 | 47.7 | 63.7 | 63.6 | 176.5 | 176.5 | 37.9 | 37.8 |
| **G2.1** | 47.3 | 47.4 | 0 | 0.5 | 1.0 | 1.0 | 331.8 | 332.0 | 37.2 | 37.0 |
| **G2.2** | 47.5 | 47.7 | 0.5 | 0 | 1.0 | 1.0 | 331.6 | 331.9 | 37.1 | 37.1 |
| **G3.1** | 63.6 | 63.7 | 1.0 | 1.0 | 0 | 0.2 | 407.5 | 407.7 | 61.8 | 61.9 |
| **G3.2** | 63.5 | 63.6 | 1.0 | 1.0 | 0.2 | 0 | 407.3 | 407.6 | 61.6 | 62.0 |
| **G4.1** | 176.2 | 176.5 | 331.8 | 331.6 | 407.5 | 407.3 | 0 | 0.3 | 45.7 | 45.3 |
| **G4.2** | 176.1 | 176.5 | 332.0 | 331.9 | 407.7 | 407.6 | 0.3 | 0 | 45.8 | 45.4 |
| **G5.1** | 37.6 | 37.9 | 37.2 | 37.1 | 61.8 | 61.6 | 45.7 | 45.8 | 0 | 0.6 |
| **G5.2** | 37.5 | 37.8 | 37.0 | 37.1 | 61.9 | 62.0 | 45.3 | 45.4 | 0.6 | 0 |

often used for binary setting of Group-1 (seizure) against other rest states. To this end, both tests clearly show that Group-1 is significantly different from all other groups (100% rejection).

A surprising observation from Table 2 ($m = 1$) is that the rejection rate is 100% within Group-1 (**G1.1** vs. **G1.2**), whereas this is not the case for Table 1 ($m = 2$). This agrees with the conclusion of Ghoshdastidar et al. (2017a) that the graph testing problem is fundamentally different for $m = 1$ and $m > 1$. Our intuition is that the networks for seizure activity are significantly different from each other, and hence, rejected by `Asymp-TW`. When we group them ($m > 1$), the fundamental question is whether two groups are identically distributed or not, and hence, the variance within each group is also taken into account. Hence, `Asymp-Normal` detects that **G1.1** and **G1.2** are identical when both graphs in each group are considered.

Although Tables 1 and 2 show that the different groups are typically rejected, they do not clearly show the degree of dissimilarity between two groups. The dissimilarity can quantified in terms of the p-value obtained from the tests. While p-value $\leq 5\%$ leads to rejection, we find that in many cases, the p-value is exponentially small. In Tables 3 and 4, we show the negative logarithm of p-value, that is $-\ln(\text{p-value})$, obtained from `Asymp-Normal` and `Asymp-TW`, respectively. The reported value is the average over 1000 independent runs. We note that the 5% significance level corresponds to $-\ln(\text{p-value}) \approx 3$, and hence, values larger than 3 correspond to rejection. Table 3 shows that this quantity can be as high as 400, and in particular, it shows that Group-4 is most dissimilar from other groups. The results of Table 4 are less conclusive since the maximum reported dissimilarity is only

Table 4: Negative logarithm of p-value (averaged over 1000 runs) computed by `Asymp-TW` for EEG correlation networks.

| | **G1.1** | **G1.2** | **G2.1** | **G2.2** | **G3.1** | **G3.2** | **G4.1** | **G4.2** | **G5.1** | **G5.2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **G1.1** | 0 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 |
| **G1.2** | 7.727 | 0 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 |
| **G2.1** | 7.727 | 7.727 | 0 | 0.017 | 0.003 | 0.008 | 7.727 | 7.727 | 1.791 | 1.924 |
| **G2.2** | 7.727 | 7.727 | 0.017 | 0 | 0 | 0.009 | 7.727 | 7.727 | 1.841 | 1.928 |
| **G3.1** | 7.727 | 7.727 | 0.003 | 0 | 0 | 0 | 7.727 | 7.727 | 1.718 | 1.780 |
| **G3.2** | 7.727 | 7.727 | 0.008 | 0.009 | 0 | 0 | 7.727 | 7.727 | 1.823 | 1.889 |
| **G4.1** | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 0 | 0.195 | 5.149 | 4.950 |
| **G4.2** | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 7.727 | 0.195 | 0 | 4.821 | 4.952 |
| **G5.1** | 7.727 | 7.727 | 1.791 | 1.841 | 1.718 | 1.823 | 5.149 | 4.821 | 0 | 0.366 |
| **G5.2** | 7.727 | 7.727 | 1.924 | 1.928 | 1.780 | 1.889 | 4.950 | 4.952 | 0.366 | 0 |

7.727. This is caused by our use of a pre-computed table for the Tracy-Widom distribution that does not return values arbitrarily close to 1 (see Appendix B for a discussion provided along with the description of the test). However, Table 4 still shows that the networks in Group-5 are relatively less different from those in Groups-2, 3 and 4.

### C.3 Experiments with autonomous systems peering networks

Our second experiment with real networks is based on a collection of networks obtained from the Stanford large network collection (Leskovec and Krevl, 2014). The networks are defined on the set of autonomous systems, which is the technical term for groups of routers that comprise the Internet. The edges correspond to communication between two autonomous systems. The first set of networks, called Oregon-1, are created from data collected by *Oregon route-views* between March 31, 2001 and May 26, 2001 once per week. This set contains 9 networks, one date per week. The second set of networks, called Oregon-2, are based on data collected on the same dates, but the peering information is inferred from a combination of *Oregon route-views, Looking glass, and Routing registry*.

All the networks are defined on a set of $n = 11806$ distinct vertices (autonomous systems), but none of the networks include all vertices, that is, every graph has few isolated vertices. The networks are also quite sparse with the number of edges varying between 22000 to 33000. We view the network collection from the following perspective. For each date, we observe two networks (one from each set) that can be considered as a population of size 2 ($m = 2$). Different dates correspond to different models for the networks, and we test for the similarity across different classes. To this end, we perform `Asymp-Normal` to detect differences, and report $-\ln(\text{p-value})$ for every test in Table 5. It is not surprising to find that the test rejects the null hypothesis at 5% significance for every pair of dates, that is, $-\ln(\text{p-value}) > 3$. The interesting observation is that $-\ln(\text{p-value})$ monotonically increases as the interval between two dates becomes larger, that is, the networks vary significantly over time. This observation is also in conjunction with the findings of Leskovec et al. (2005), where a more qualitative analysis was made based on number of edges and average node degree. We do not report corresponding results for `Asymp-TW` since our current implementation can provide a maximum $-\ln(\text{p-value})$ of at most 7.727, and hence, does not provide any additional information.

We next perform semi-synthetic experiments with Oregon network dataset. We first consider the case of $m = 2$, where we use `Asymp-Normal`. For every pair of networks, we randomly select $k = 118$ vertices (1% of vertex set), and replace the sub-graph by an Erdős-Rényi (ER) graph with edge probability $p$. On Figure 6 (left panel), we show how $-\ln(\text{p-value})$ varies as the edge density of the ER graph increases from $p = 0.2$ to $0.4$, where each line corresponds to one date (one pair of networks) and the results are averaged over 100 runs. We find that $-\ln(\text{p-value})$ increases linearly with $p$, that is, p-value decreases exponentially. The trend is almost similar for every network pair. We also study the effect of adding sparse ER graphs in Figure 6 (right panel). Here we plant an ER graph on a random subset of $k$ vertices, where $k$ varies from 1% to 2% of total number of vertices. However, the planted ER graphs are sparse with $p = \frac{20}{k}$, that is they have constant average degree. We observe a slightly super-linear increase of $-\ln(\text{p-value})$ in this case.

Table 5: Negative logarithm of p-value obtained by `Asymp-TW` for every pair of dates in the Oregon network dataset.

| | Mar 31 | Apr 7 | Apr 14 | Apr 21 | Apr 28 | May 5 | May 12 | May 19 | May 26 |
|---|---|---|---|---|---|---|---|---|---|
| Mar 31 | 0 | 13.7 | 25.0 | 36.4 | 59.6 | 77.4 | 96.8 | 106.2 | 135.0 |
| Apr 7 | 13.7 | 0 | 6.5 | 15.2 | 31.0 | 45.7 | 61.1 | 69.7 | 93.4 |
| Apr 14 | 25.0 | 6.5 | 0 | 6.0 | 17.9 | 29.6 | 42.5 | 50.2 | 71.4 |
| Apr 21 | 36.4 | 15.2 | 6.0 | 0 | 8.5 | 17.2 | 27.6 | 34.9 | 54.7 |
| Apr 28 | 59.6 | 31.0 | 17.9 | 8.5 | 0 | 5.3 | 12.8 | 22.6 | 45.7 |
| May 5 | 77.4 | 45.7 | 29.6 | 17.2 | 5.3 | 0 | 4.8 | 13.0 | 31.2 |
| May 12 | 96.8 | 61.1 | 42.5 | 27.6 | 12.8 | 4.8 | 0 | 4.7 | 18.3 |
| May 19 | 106.2 | 69.7 | 50.2 | 34.9 | 22.6 | 13.0 | 4.7 | 0 | 5.6 |
| May 26 | 135.1 | 93.4 | 71.4 | 54.7 | 45.7 | 31.2 | 18.3 | 5.6 | 0 |

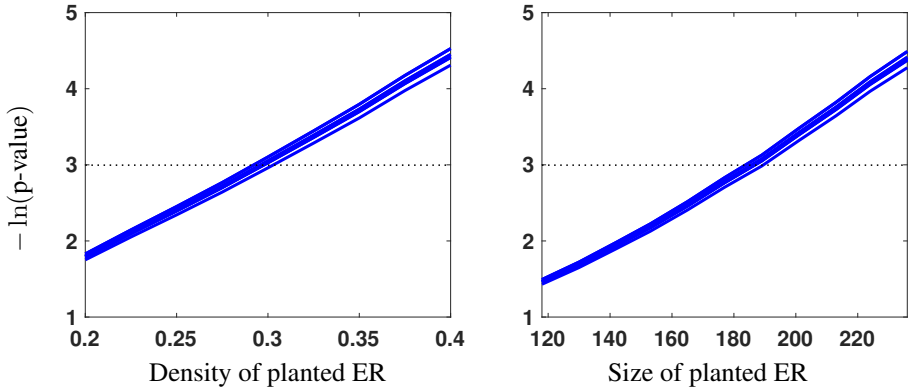


Figure 6: Variation of $-\ln(\text{p-value})$ for `Asymp-Normal` when Erdős-Rényi subgraphs are planted into the network. Each line corresponds to one of the 9 pairs. The dotted line corresponds to 5% significance level. **(Left)** Subgraph size is 1% of network size, and edge probability is varied. **(Right)** Subgraph size is varied from 1-2% of network size, and edge probability is decreased.

Finally, we consider a semi-synthetic experiment with $m = 1$, where we use `Asymp-TW`. For each of the 18 networks, we randomly select #e pairs of vertices and toggle their connection, that is, if an edge is present then we remove it, or the reverse. We vary #e from 0 to 300 in steps of 25. Figure 7 reports the values for $-\ln(\text{p-value})$ (averaged over 100 runs) for each network. We present the results in two panels corresponding to the two datasets Oregon-1 and Oregon-2. Surprisingly, we find that $-\ln(\text{p-value})$ rapidly increases with #e although the number of perturbed edges are much smaller than the total of $\binom{11806}{2}$ possible pairs. We also find that the networks in each collection have a similar trend, and the Oregon-2 networks show a slightly smaller value than Oregon-1. This is possibly because the Oregon-2 networks are more dense than their Oregon-1 counterparts.

We conclude our discussion with some implementation details for `Asymp-TW` in this setup related to the community detection step. Since the networks are large and sparse, standard spectral clustering fails to return reasonable communities. Hence, we use BigClam (Yang and Leskovec, 2013), which is suitable for finding a large number of communities in a large network. The method returns multiple community assignments for some vertices and does not make any assignments for few. We use BigClam to find an initial set of 50 overlapping communities from the union of all graphs, and then resolve cases of overlap or no-assignments by assigning vertices to communities with which they have maximum connection. These pre-computed communities are used for the purpose of approximation in `Asymp-TW` test. The above results in Figure 7 show that the use of `Asymp-TW` in conjunction BigClam provides reliable results, and hence, we believe that `Asymp-TW` is applicable even in the sparse regime provided that it is used with a reasonable community detection algorithm.

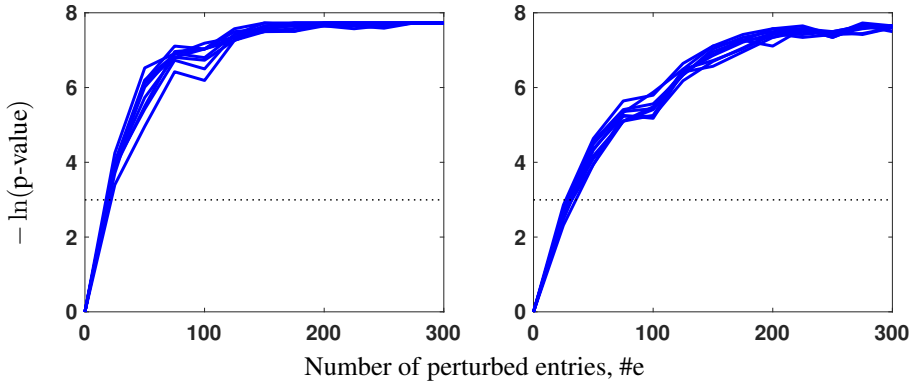


Figure 7: Variation of $-\ln(\text{p-value})$ for `Asymp-TW` when a random set of #e out of $\binom{n}{2}$ edges are inserted/deleted. The dotted line corresponds to 5% significance level. **(Left)** Each line corresponds to one of the 9 networks from Oregon-1 set. **(Right)** Each line is for a network from Oregon-2 set.