Goodness-of-Fit Tests for Inhomogeneous Random Graphs

Soham Dan [1] **Bhaswar B. Bhattacharya** [2]

Abstract

Hypothesis testing of random networks is an emerging area of modern research, especially in the high-dimensional regime, where the number of samples is smaller or comparable to the size of the graph. In this paper we consider the goodness-of-fit testing problem for large inhomogeneous random (IER) graphs, where given a (known) reference symmetric matrix $Q \in [0,1]^{n\times n}$ and m independent samples from an IER graph given by an unknown symmetric matrix $P \in [0,1]^{n\times n}$, the goal is to test the hypothesis $P = Q$ versus $||P-Q|| \geq \varepsilon$, where $||\cdot||$ is some specified norm on symmetric matrices. Building on recent related work on two-sample testing for IER graphs, we derive the optimal minimax sample complexities for the goodness-of-fit problem in various natural norms, such as the Frobenius norm and the operator norm. We also propose practical implementations of natural test statistics, using their asymptotic distributions and through the parametric bootstrap. We compare the performances of the different tests in simulations, and show that the proposed tests outperform the baseline tests across various natural random graphs models.

1. Introduction

With the ubiquitous presence networks in bioinformatics and social sciences, developing statistical methods for graph-valued data has become increasingly important. Although network analysis has been an area of active interest in statistics and machine learning, most classical approaches for graph testing are applicable in the relatively low-dimensional setting, where the population size (number of graphs) is larger than the size of the graphs (number of vertices). However, in the modern high-dimensional regime [19] the number of samples m could be potentially much smaller or comparable to the size of the graph n, for example, graphs may correspond to different sets of relations constructed for a set of actors [16]. As a result, theoretical understanding for testing random graphs is an area of emerging interesting. Most of the recent work has focussed primarily on detecting planted communities or sparse structures in the network [3, 4, 22, 26], and testing for network dependence [2, 8].

Here, we consider the problem of goodness-of-fit testing for network data, which involves developing statistical tests for assessing whether a given sample of networks fits a specified model. This problem has found many applications recently, for example, in assessing fit of protein-protein interaction (PPI) networks [9, 27] and in functional neuroimaging data [15]. In particular, Ospina-Forero et al. [27] developed a non-parametric procedure for testing whether network models used as backgrounds in community detection of social networks, such as the Erdős-Rényi model or the Chung-Lu model [6], can also be used to describe the appearance of subgraph counts in the Facebook networks in US universities. They also applied their method for assessing fit in PPI networks, which arise in many biological studies, such as the discovery of disease risk pathways and the investigation of genes undergoing age expression changes. Another application is in functional neuro-imaging data [15], where the vertices correspond to regions of interest in the brain, and an edge between two regions indicates functional connectivity, in the sense that the two regions interact together to achieve some higher-order function. Here, Ginestet et al. [15] considered the problem of testing whether the Laplacian matrix of the model generating a given sample of networks is equal to a reference Laplacian matrix, based on the asymptotics of the sample (Fréchet) mean, when the graph sizes are fixed and the sample sizes grow to infinity. Other goodness-of-fit tests for the β-model and the exponential random graph model (ERGM) are discussed in [7] and [23], respectively.

In this paper, we propose theoretically optimal and computationally efficient methods for network goodness-of-fit for inhomogeneous Erdős-Rényi (IER) random graph models. This is a general class of random graph model, which includes several popular network models, such as the Chung-Lu model [6], the β-model [5], random dot product graphs

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA, [2]Department of Statistics, University of Pennsylvania, Philadelphia, USA. Correspondence to: Soham Dan <sohamdan@seas.upenn.edu>, Bhaswar B. Bhattacharya <bhaswar@wharton.upenn.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[29], and stochastic block models [22]. Our proposed tests attain optimal sample complexities, under different matrix norms, and can be efficiently implemented, either by their asymptotic distributions or the parametric bootstrap. The tests also perform well in a range of simulation experiments, illustrating the broad applicability of the methods.

1.1. Problem Statement and Summary of Results

Given a symmetric matrix $P^{(n)} \in [0,1]^{n\times n}$ with zeroes on the diagonal, a graph G is said to be an *inhomogeneous Erdős-Rényi* (IER) random graph [1] with *edge probability* $P^{(n)} = ((p_{ij}))$, denoted as $G \sim \mathrm{IER}(P^{(n)})$, if its symmetric adjacency matrix $A(G) = ((a_{ij}(G))) \in \{0,1\}^{n\times n}$ have independent entries satisfying:

$$a_{ij}(G) \sim \mathrm{Ber}(p_{ij}) \text{ for all } i < j.$$

Recently, Tang et al. [29, 30] and Ghoshdastidar et al. [12, 13, 14] studied the problem of two-sample testing for this model which asks: Given two populations of random graphs, decide whether both populations are generated from the same distribution or not? This paper addresses the related problem of goodness-of-fit, where given independent graph samples $G_1, G_2, \ldots, G_m \sim \mathrm{IER}(P^{(n)})$, and a known matrix $Q^{(n)}$, the goal is to test the hypothesis $H_0 : P^{(n)} = Q^{(n)}$ versus

$$H_1 : ||P^{(n)} - Q^{(n)}|| \geq \varepsilon, \tag{1.1}$$

where $||\cdot||$ is some 'norm' on symmetric matrices. Here, we will consider the following three natural norms on a symmetric matrix $A = ((a_{ij})) \in [0,1]^{n\times n}$:

- *Frobenius Norm*: $||A||_F = \sqrt{\sum_{1\leq i,j\leq n} a_{ij}^2}$, which is the root of the sum of squares of all the entries in A.
- *Operator Norm*: $||A||_{\mathrm{op}} = \max\{|\lambda_n(A)|, |\lambda_1(A)|\}$, where $\lambda_n(A) \geq \lambda_{n-1}(A) \geq \cdots \geq \lambda_1(A)$ are the eigenvalues of A.
- *Zero Norm*: $||A||_0 = \sum_{1\leq i,j\leq n} \mathbf{1}\{a_{ij} \neq 0\}$, which is the number of non-zero entries in A. This is not really a norm, but is a popular way to quantity the sparsity of matrix.

Given i.i.d. samples $G_1, G_2, \ldots, G_m$ from $\mathrm{IER}(P)$, a test is a binary function $\phi : \boldsymbol{G}_m := (G_1, G_2, \ldots, G_m) \to \{0,1\}$, which is 0 when the test accepts H_0 and 1 otherwise. The worst-case risk of a test function ϕ for the testing problem (1.1) is defined as:

$$\mathcal{R}_m(Q^{(n)}, \phi, ||\cdot||) = \mathbb{P}_{Q^{(n)}}(\phi = 1) + \sup_{\substack{P^{(n)}: \\ ||P^{(n)} - Q^{(n)}|| \geq \varepsilon}} \mathbb{P}_{P^{(n)}}(\phi = 0), \tag{1.2}$$

which is the sum of the Type I error and the maximum possible Type II error rate of the test ϕ. (Hereafter, we will omit the dependence on the distance function $||\cdot||$ in (1.2) above, whenever it is clear from the context.) We are interested in the asymptotic regime where the risk (1.2) transitions from 0 to 1. This is formalized in the following definition:

Definition 1.1. Given $G_1, G_2, \ldots, G_m$ i.i.d. samples from $\mathrm{IER}(P^{(n)})$, where $m = m_n$ can depend on n, a sequence of test functions $\phi_{n,m}$ is said to be *asymptotically powerless* for (1.1), if there exists a sequence of symmetric matrices $Q^{(n)} \in [0,1]^{n\times n}$ such that $\lim_{n\to\infty} \mathcal{R}_m(Q^{(n)}, \phi_{m,n}) = 1$. On the other hand, a sequence of test functions $\phi_{n,m}$ is said to be *asymptotically powerful* for (1.1), if for all symmetric matrices $Q^{(n)} \in [0,1]^{n\times n}$, $\lim_{n\to\infty} \mathcal{R}_m(Q^{(n)}, \phi_{n,m}) = 0$.

The main focus of this paper is to derive optimality results for goodness-of-fit testing in IER graphs for the various norms described above, and complement these results with implementable tests based on asymptotic properties and the bootstrap. The following is summary of the results obtained in this paper:

- We show that the *optimal sample complexity* for testing separation as in (1.1), for both the Frobenius (Theorem 2.2) and the operator norm (Theorem 2.3), is n/ε^2. This means that there is a (computationally efficient) test which is asymptotical powerful for (1.1) when the sample size $m \gg n/\varepsilon^2$, and all tests asymptotically powerless when the sample size $m \ll n/\varepsilon^2$. We also show that testing for any separation is impossible in the zero norm (Theorem 2.1).

- Next, we derive the asymptotic null distribution and statistical consistency (against a large class of alternatives) of a natural goodness-of-fit test, based on a sample estimate of the Frobenius norm, which can be used to efficiently calibrate the test statistic for moderate to large sized networks (Section 3). This test statistic, however, fails to work when there is only one sample ($m = 1$), in which case, we propose a test based on the operator norm of the sample adjacency matrix. We also discuss how the method of parametric bootstrap [11] can be used to approximately calibrate any test statistic, and compare the performance of the different tests in finite sample simulations (Section 4). Our experiments show that the proposed asymptotic Frobenius norm based test accurately approximates the null distribution and has good power for a wide class of alternatives for moderate sized networks, even when the sample size m is very small.

1.2. Organization

The rest of the paper is organized as follows: The optimal sample complexities for testing in the norms described above are derived in Section 2. The asymptotic null distribution, consistency, and details of the bootstrap are discussed in Section 3. The performance of the different tests in finite sample simulations are described in Section 4. Proofs of the theorems and additional simulations are given in the appendix.

2. Minimax Sample Complexities

In this section, we obtain the optimal sample complexities for goodness-of-fit testing under the three norms described above. We begin by recalling some standard asymptotic notation. For two nonnegative sequences $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n \lesssim b_n$ means $a_n = O(b_n)$; $a_n \sim b_n$ means $a_n = (1+o(1))b_n$; $a_n \asymp b_n$ means $a_n \lesssim b_n \lesssim a_n$; $a_n \ll b_n$ means $a_n = o(b_n)$; and $a_n \gg b_n$ means $b_n = o(a_n)$.

We start with an impossibility theorem about the zero-norm. The following theorem, which is proved in Appendix 1, shows that testing in zero norm is impossible, for any symmetric matrix $Q^{(n)} \in [0,1]^{n\times n}$ and any $\varepsilon > 0$. This is because, we can increase the zero norm between two matrices to their maximum possible value (which is $n(n-1)$), by making arbitrarily small perturbations to the elements of the matrices.

Theorem 2.1. *For the testing problem (1.1) under the zero norm $||\cdot||_0$, all tests are asymptotically powerless for any sequence of symmetric matrices $Q^{(n)} \in [0,1]^{n\times n}$ and any $\varepsilon > 0$.*

Next, we consider testing in the Frobenius norm. In this case our test is based on the following statistic:

$$T_{m,n} := \sum_{1\leq i<j\leq n} L_{m,n}^{(i,j)} \cdot R_{m,n}^{(i,j)}, \tag{2.1}$$

where

- $L_{m,n}^{(i,j)} = \sum_{s\leq \frac{m}{2}} (a_{ij}(G_s) - q_{ij})$ and $R_{m,n}^{(i,j)} = \sum_{s>\frac{m}{2}} (a_{ij}(G_s) - q_{ij})$,
- $A(G_s) = ((a_{ij}(G_s)))$ is the adjacency matrix of the graph G_s, and
- $Q^{(n)} = ((q_{ij}))$ is the reference edge-probability matrix.

This statistic is a natural modification of the two-sample statistic introduced in [12, 14]. Note that $\mathbb{E}_{P^{(n)}}(T_{m,n}) = \frac{m^2}{8}||P^{(n)} - Q^{(n)}||_F^2$, that is, $\frac{1}{m^2}T_{m,n}$ is an unbiased estimate of $\frac{1}{8}||P^{(n)} - Q^{(n)}||_F^2$. The following theorem shows that this statistic attains the optimal sample complexity for testing in the Frobenius norm for IER graphs. The proof is given in Appendix 2.

Theorem 2.2. *For the testing problem (1.1) under the Frobenius norm $||\cdot||_F$, the following hold:*

(a) The test $\phi_{n,m} = \mathbf{1}\{T_{m,n} \geq \frac{1}{16}m^2\varepsilon^2\}$, where $T_{m,n}$ is as in (2.1) above, is asymptotically powerful for (1.1), whenever $m \gg n/\varepsilon^2$.

(b) On the other hand, all tests are asymptotically powerless for (1.1), whenever $m \ll n/\varepsilon^2$.

Remark 2.1. In fact, from our analysis of the $T_{m,n}$ statistic (in Appendix 2) we can obtain an upper bound on the sample complexity that depends on Q. In particular, the proof shows that the sample complexity for testing in Frobenius norm is greater than $\max\{1/\varepsilon^2, \min\{||Q||_F/\varepsilon^2, ||J-Q||_F/\varepsilon^2\}\}$, where J is the matrix with 1 in every off diagonal entry and 0 in every diagonal entry. (Note the symmetry in Q and $J-Q$ in the sample complexity bound above, which is due to the symmetry in the problem arising from observing the graph or its complement.) Depending on the structure of Q, this bound can be better than the worst-case n/ε^2 sample complexity reported above. For instance, when Q is the adjacency matrix of an Erdős-Rényi random graph with edge probabilities $q \in (0, 1/2)$, this simplifies to $\max\{1/\varepsilon^2, nq/\varepsilon^2\}$, which improves upon n/ε^2, for $q \ll 1$. We expect this upper bound to be tight for a wide range of parameters, as is evident from the lower bound calculations.

Finally, we consider testing separation in the operator norm. Here, the inequality $||P^{(n)} - Q^{(n)}||_F \geq ||P^{(n)} - Q^{(n)}||_{\text{op}}$ immediately implies the test $\phi_{n,m}$ in Theorem 2.2 above, will be asymptotically powerful for testing separation in the operator norm as well, whenever $m \gg n/\varepsilon^2$. The following theorem (proved in Appendix 3) shows that n/ε^2 samples are also necessary in this case.

Theorem 2.3. *For the testing problem (1.1) under the operator norm $||\cdot||_{\text{op}}$, the following hold:*

(a) The test $\phi_{n,m} = \mathbf{1}\{T_{m,n} \geq \frac{1}{16}m^2\varepsilon^2\}$ is asymptotically powerful for (1.1), whenever $m \gg n/\varepsilon^2$.

(b) On the other hand, all tests are asymptotically powerless for (1.1), whenever $m \ll n/\varepsilon^2$.

3. Asymptotic Distribution and the Parametric Bootstrap

In this section we discuss various practical goodness-of-fit tests for IER graphs, based either on asymptotic distributions or the parametric bootstrap [11]. To begin with note the classical low-dimensional large sample size regime in

the context of network testing, corresponds to a sample $G_1, G_2, \ldots, G_m$ of graphs on n vertices, where the size n of the graphs is fixed, but the sample size $m \to \infty$. In this case, natural goodness-of-fit tests can be obtained by using χ^2-type statistics discussed in [12, 15]. The asymptotic properties of these tests follow from classical theory, however, as in common for tests tailored for low-dimensional problems, these test perform poorly, in the high-dimensional regime, where the size of the graph n is much large than the sample size m.

In the following we will discuss different implementations of goodness-of-fit tests which are powerful when the sample size is much smaller or comparable to the size of the graph. In Section 3.1 we derive asymptotic null distribution of the statistic $T_{m,n}$ (recall (2.1)), using which an efficient test can be constructed when the number of samples $m > 1$. We also discuss a spectral test based on the (properly normalized) adjacency matrix of the observed graph, for the case $m = 1$. In Section 3.2 we describe the method of parametric bootstrap, a well-known resampling technique [11] for calibrating the null distribution of any test statistic.

3.1. Asymptotic Properties

We begin with the asymptotic distribution of $T_{m,n}$. The following theorem shows that $T_{m,n}$ (appropriately scaled) converges in distribution to standard normal under H_0. Using this we can construct a test with probability of Type I error converging to α (*asymptotically level* α), which is also consistent (probability of Type II error converging to zero), under a separation condition in terms of $||P^{(n)} - Q^{(n)}||_F^2$. The asymptotics is in $n \to \infty$, where $m > 1$ is also allowed to depend on n. We denote by z_α the $(1-\alpha)$-th quantile of the standard normal $N(0,1)$.

Theorem 3.1. *Suppose $\{Q^{(n)}\}_{n\geq 1}$ be a sequence of edge-probability matrices with entries bounded away from 1, such that $\lim_{n\to\infty} ||Q^{(n)}||_F = \infty$. Then under the null hypothesis H_0,*

$$Z_{m,n} := \frac{T_{m,n}}{\sqrt{\frac{m^2}{8}\sum_{1\leq i<j\leq n} q_{ij}(1-q_{ij})}} \xrightarrow{D} N(0,1),$$

where $T_{m,n}$ is as defined in (2.1). As a consequence, the test which rejects H_0 when $|Z_{m,n}| > z_{\alpha/2}$ is asymptotically level α, that is,

$$\lim_{n\to\infty} \mathbb{P}_{H_0}(|Z_{m,n}| > z_{\alpha/2}) = \alpha. \tag{3.1}$$

On the other hand, if $\{P^{(n)}\}_{n\geq 1}$ is such that $||P^{(n)} - Q^{(n)}||_F^2 \gg \frac{1}{m}||Q^{(n)}||_F$, then

$$\lim_{n\to\infty} \mathbb{P}_{P^{(n)}}(|Z_{m,n}| > z_{\alpha/2}) = 1, \tag{3.2}$$

that is, the power of the test converges to 1.

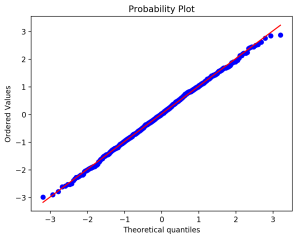


Figure 1. (a)

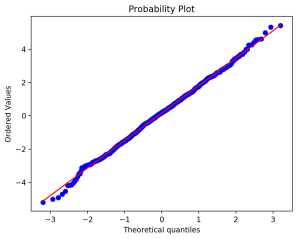


Figure 2. (b)

Figure 3. The quantile-quantile (QQ) plots of $Z_{m,n}$, for sequence of $m = 4$ graphs on $n = 100$ vertices each generated from (a) the Erdős-Rényi model $\mathrm{ER}(100, 0.5)$, and (b) the planted bisection model $\mathrm{PB}(100, 0.9, 0.1)$.

The proof of the theorem is given in Appendix 4. This result parallels the CLT for the related two-sample statistic derived in [12]. However, unlike in the two-sample case where the corresponding test is conservative (that is, the limiting Type I error is less than equals to α), for the goodness-of-fit problem the test which rejects H_0 when $|Z_{m,n}| > z_{\alpha/2}$ has asymptotic Type I error exactly α (see (3.1)). This is because in the goodness-of-fit problem, we know the variance of $T_{m,n}$ under the null, and, hence, the standardized statistic $Z_{m,n}$ can be directly computed from the data. Moreover, since the test based on $T_{m,n}$ unbiasedly estimates $||P^{(n)} - Q^{(n)}||_F^2$, it is natural to expect consistency when this separation becomes large (as in (3.2)).

Figure 3 shows the quantile-quantile (QQ) plots of the statistic $Z_{m,n}$, under the null, for the Erdős-Rényi model and the planted bisection model:

- *Erdős-Rényi Model*: This is a special case of the IER model, where all the edge interconnection probabilities are the equal. More formally, given $q \in [0,1]$, we denote this model by $\mathrm{ER}(n,q)$, that is, a random graph on n vertices where each edge is present or absent independently with probability q. This is one of the most fundamental models for random networks, and has been extensively studied over the last few decades [20].

- *Planted Bisection Model*: This is a special case of the well-known stochastic block model [25], in which the nodes are divided into two equal-sized communities and then edges are added randomly in a way that depends on the community membership. More formally, the planted bisection model is an IER graph on n vertices where the edge-probability edge $Q = ((q_{ij}))$ has

the following form 2-block structure:

$$q_{ij} := \begin{cases} a & \text{if } 1 \le i \ne j \le \frac{n}{2} \text{ or } \frac{n}{2} < i \ne j \le n, \\ b & \text{if } 1 \le i \le \frac{n}{2} \text{ and } \frac{n}{2} < j \le n \text{ or} \\ & \quad \frac{n}{2} < i \le n \text{ and } 1 \le j \le \frac{n}{2} \\ 0 & \text{if } i = j, \end{cases}$$

where $a, b \in [0, 1]$. Given $a, b \in [0, 1]$, we denote a random graph on n vertices from this model as $\mathrm{PB}(n, a, b)$.

To validate the asymptotic results in Theorem 3.1, we simulate the null distribution of $Z_{m,n}$ for the two models described above, for a collection of $m = 4$ graphs on $n = 100$ vertices, and compare the empirical quantiles of $Z_{m,n}$ (over 1000 iterations) with the predicted theoretical quantiles of $N(0, 1)$. The plots show that the asymptotics in Theorem 3.1 give accurate approximations for moderate-size graphs ($n = 100$) sample size as small as 4. We investigate the power of this test for various alternatives in Section 4 and in Appendix 5.

Remark 3.1. (The case $m = 1$.) When the data consists of only a single graph, the statistic (2.1) becomes degenerate, and it no longer unbiasedly estimates $||P^{(n)} - Q^{(n)}||_F^2$. In this case, we propose a test based on the spectral norm of the (scaled) observed adjacency of the graph. To this end, suppose $G_1 \sim \mathrm{IER}(P^{(n)})$ and consider the scaled adjacency matrix $W_n := ((w_{ij}))$, where

$$w_{ij} := \frac{a_{ij}(G_1) - q_{ij}}{\sqrt{(n-1)q_{ij}(1-q_{ij})}}, \tag{3.3}$$

for $1 \le i, j \le n$, where $Q^{(n)} = ((q_{ij}))$ is the reference edge-probability matrix under the null H_0. Note that, under the null H_0, W_n is a symmetric random matrix, whose entries above the diagonal are independent with mean zero and variance $\frac{1}{n-1}$. Hence, by directly applying well-known results from random matrix theory [10, 21] we can get the limiting null distribution of the largest and the smallest eigenvalues of W_n. In particular, $\lambda_1(W_n)$ and $\lambda_n(W_n)$ has the same limiting distribution as $\lambda_1(D_n)$ and $\lambda_n(D_n)$, where D_n is a symmetric random matrix with zero diagonal, whose entries above the diagonal are i.i.d. normal with mean zero and variance $\frac{1}{n-1}$ [10]. This combined with results of Lee and Yin [21] about $\lambda_1(D_n)$ and $\lambda_n(D_n)$ implies

$$n^{\frac{2}{3}}(\lambda_1(W_n) - 2) \xrightarrow{D} TW_1$$

and

$$n^{\frac{2}{3}}(-\lambda_n(W_n) - 2) \xrightarrow{D} TW_1,$$

where TW_1 is Tracy-Widom law for orthogonal ensembles [31]. Then, recalling that $||W_n||_{\mathrm{op}} = \max\{|\lambda_1(W_n)|, |\lambda_n(W_n)|\}$, and a union bound, implies that under the null,

$$\limsup_{n\to\infty} \mathbb{P}_{H_0}\left(n^{\frac{2}{3}}(||W_n||_{\mathrm{op}} - 2) \ge \tau_{\alpha/2}\right) \le \alpha, \tag{3.4}$$

where $\tau_{\alpha/2}$ is the $(1 - \frac{\alpha}{2})$-th quantile of the TW_1 distribution.

3.2. The Parametric Bootstrap

The parametric bootstrap is a well-known resampling technique [11], which can be used to approximate the null distribution of any test statistic for the goodness-of-fit problem, by repeatedly sampling from the null model. The details of the algorithm are given below in Algorithm 1.

Algorithm 1 Bootstrapping a Test Statistic

Input : Samples $G_1, \ldots, G_m$ from a IER model, the null edge-probability matrix Q, a test statistic $S_{m,n}$, a significance Level $\alpha \in (0, 1)$, and $B \ge 1$ (the number of bootstrap repetitions).

1: **for** $b = 1$ to B **do**
2: Draw m samples $G_1^{(b)}, \ldots, G_m^{(b)}$ from the $\mathrm{IER}(Q)$ model.
3: Compute the test statistic $S_{m,n}^{(b)} = S_{m,n}(G_1^{(b)}, \ldots, G_m^{(b)})$.
4: Denote by $L_{m,n}$ and $U_{m,n}$ the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ empirical quantiles of $\{S_{m,n}^{(1)}, S_{m,n}^{(2)}, \ldots, S_{m,n}^{(B)}\}$, respectively.
5: **end for**

Output: Reject H_0 if $S_{m,n}(G_1, \ldots, G_m) \notin [L_{m,n}, U_{m,n}]$. Otherwise accept H_0.

Note that sampling a $\mathrm{IER}(Q)$ random graph on n vertices takes $O(n^2)$ time, therefore, the algorithm above can be easily implemented for moderate sized networks. In Section 4, we compare the asymptotic tests described above with their bootstrap counterparts. We will refer to the bootstrapped analogue of $T_{m,n}$ as the `Bootstrapped Frobenius Test`. We will also consider a bootstrap test based on the statistic

$$\left|\left|\sum_{s=1}^{m} A(G_s) - mQ^{(n)}\right|\right|_{\mathrm{op}}, \tag{3.5}$$

where $G_1, G_2, \ldots, G_m$ are i.i.d. $\mathrm{IER}(Q^{(n)})$, which we will refer to as the `Bootstrapped Operator-Norm Test`.[1] In addition, we will consider the bootstrap versions of the following two baseline tests:

- *The Edge Test*: Given $G_1, G_2, \ldots, G_m$ i.i.d. samples from $\mathrm{IER}(Q^{(n)})$, the edge test rejects the

[1] Incidentally, it can be shown by an application of matrix concentration inequalities that the test based on (3.5) is asymptotic powerful for detecting ε separation in the operator norm, whenever $m \gg n \log n/\varepsilon^2$. One might also expect to remove the factor of $\log n$ to match the lower bound in Theorem 2.3, using a moment-method based argument, similar to that in [14], where the analogous two-sample problem was studied.

null for large/small values of the sum of the total number of edges in the observed sample, that is, $\sum_{s=1}^{m}\sum_{1\le i<j\le n} a_{ij}(G_s)$. This test is especially powerful, when the total number edges in the alternative model is significantly different from the null model, for example, perturbations in the Erdős-Rényi model (see Section 4), but is powerless for alternatives with similar number of edges as the null. We will refer to the bootstrapped version of this test as the `Bootstrapped Edge Test`.

- *The Cycle Test*: This is another natural test for goodness-of-fit when $m = 1$, based on the trace of the scaled adjacency matrix W_n (as defined in (3.3)). Note that $\mathrm{tr}(W_n^g) = \sum_{i=1}^{n}\lambda_i^g(W_n)$, where $\lambda_1(W_n) \ge \lambda_2(W_n) \ge \cdots \lambda_n(W_n)$ are the eigenvalues of W_n. Note that $\mathrm{tr}(W_n^g)^{\frac{1}{g}}$ is the g-th norm of the eigenvalues of W_n, which in the $g \to \infty$ limit gives $\max_{1\le i\le n}|\lambda_i(W_n)| = ||W_n||_{\mathrm{op}}$, and, hence, can be thought of as a (moment-based) approximation to the operator norm. We will refer to the bootstrapped version of this test as the `Bootstrapped Cycle Test`.

4. Numerical Results

In section we compare the power of the tests described above for 3 random graph models: the Erdős-Rényi model, the planted bisection model, and the β-model. The Erdős-Rényi model and the planted bisection model are described above in Section 3.1. The β-model is another popular IER model [5, 17, 26, 28], where the edge-probability matrix $Q = ((q_{ij}))$ is given by

$$q_{ij} = \frac{e^{\beta_i+\beta_j}}{1+e^{\beta_i+\beta_j}},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$ is n-dimensional parameter vector. Given $\boldsymbol{\beta} \in \mathbb{R}^n$, we denote a random graph on n vertices from this model as $\mathscr{B}(n, \boldsymbol{\beta})$. The β-model is a simple version of a collection of exponential models actively in use for analyzing network data (for example, it includes as a special case the stochastic block model), and is a close analogue to the Bradley-Terry model for rankings [18].

Here, we study the power of the different tests as a function of increasing separation, keeping the sample size and the size of the graph fixed (Section 4.1). (Additional simulations illustrating the dependence on the size of the graph and the sample size are given in Appendix 5.) We also compare the performance of the different tests, in graphon-based IER (Section 4.2) and sparser networks (Section 4.3).

4.1. Dependence on Separation

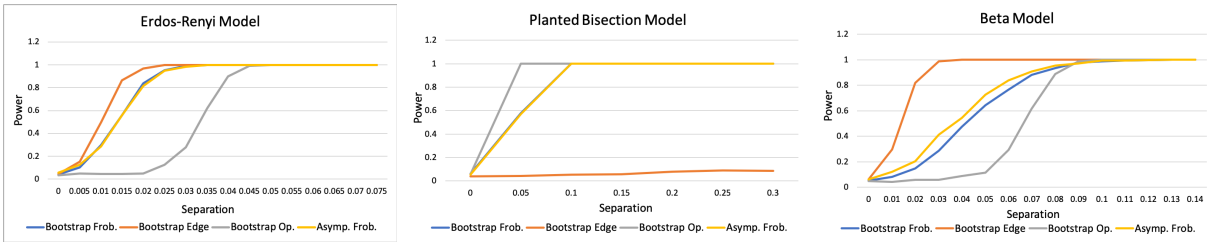
For our simulations, we fix the size of the graph $n = 100$, the sample size $m = 4$, and a reference edge-probability matrix $Q^{(n)}$ (which corresponds to the null), and consider samples G_1, G_2, G_3, G_4 i.i.d from $\mathrm{IER}(P^{(n)})$, where $P^{(n)}$ is a certain perturbation of the $Q^{(n)}$. The figures below show the empirical power of the tests over 1000 iterations (calibrated either using the asymptotic distribution or the parametric bootstrap at level $\alpha = 0.05$) as the perturbation parameter increases. We consider the following three scenarios:

- In Figure 4(a) the reference matrix $Q^{(n)}$ corresponds to $\mathrm{ER}(100, \frac{1}{2})$ and the matrix $P^{(n)}$ corresponds to $\mathrm{ER}(100, \frac{1}{2}+\Delta)$. Figure 4(a) shows the empirical power of the tests as a function of increasing Δ. As expected, the `Bootstrapped Edge Test` has the highest power here, since uniformly increasing the edge-probabilities, increases the expected number of edges, making this the most powerful test in this case. The `Asymptotic Frobenius Test` based on $Z_{m,n}$ (recall Theorem 3.1) and `Bootstrapped Frobenius Test` also perform very well, with power converging to 1 around $\Delta = 0.025$. On the other hand, `Bootstrapped Operator-Norm Test` has power converging to 1 much slowly, in this case.

- In Figure 4(b) the reference matrix $Q^{(n)}$ corresponds to the planted bisection model $\mathrm{PB}(100, 0.6, 0.4)$, and the alternative edge-probability matrix $P^{(n)}$ corresponds to $\mathrm{PB}(100, 0.6+\Delta, 0.4-\Delta)$. Figure 4(b) shows the empirical power of the tests as a function of increasing Δ. Here, the `Bootstrapped Operator-Norm Test` has the highest power. Also, as before, the `Asymptotic Frobenius Test` and `Bootstrapped Frobenius Test` perform similarly and very well, with power converging to 1 around $\Delta = 0.01$. However, unlike in the Erdős-Rényi case, the `Bootstrapped Edge Test` is powerless because the perturbations considered in the planted bisection model do not change the expected number of edges the graph.

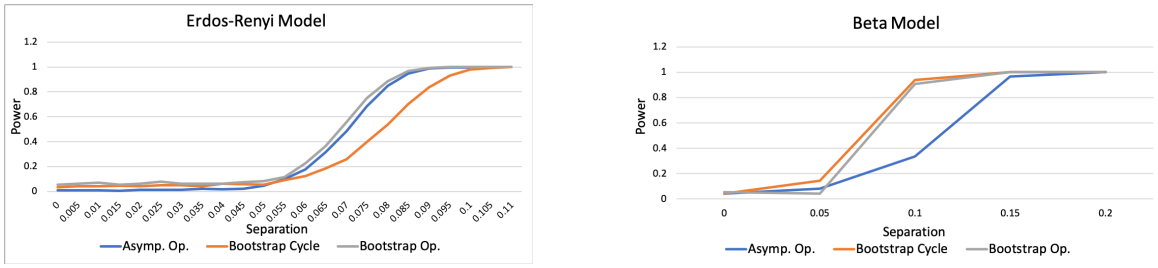
- In Figure 4(c) the reference matrix $Q^{(n)}$ corresponds to the β-model $\mathscr{B}(100, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is chosen uniformly from the surface of ball in $\mathbb{R}^{100}$ with radius 20, and $P^{(n)}$ corresponds to the β-model $\mathscr{B}(100, \boldsymbol{\beta}+\Delta)$. Figure 4(c) shows the empirical power of the tests as a function of increasing Δ. Here, the `Bootstrapped Edge Test` has the highest power. Again the `Asymptotic Frobenius Test` and `Bootstrapped Frobenius Test` have similar performance with power converging to 1 around $\Delta = 0.09$.

Overall we see that our `Asymptotic Frobenius Test` and `Bootstrapped Frobenius Test` have



(a) (b) (c)

Figure 4. Empirical power of the different tests as a function of increasing separation for $m = 4$ graphs of size $n = 100$ in (a) the Erdős-Rényi model, (b) the planted bisection model and (c) the β-model.



(a) (b)

Figure 5. Empirical power of the different tests as a function of increasing separation for graphs of size $n = 100$, based on a single ($m = 1$) graph from the (a) the Erdős-Rényi model, and (b) the β-model.

very good power across the 3 different IER models, illustrating their robustness and consistency against a large class of alternatives (as proved in Theorem 3.1). In practice, when the graph size is small we recommend using the `Bootstrapped Frobenius Test`, because it can accurately approximate the null distribution. For moderate to large size graphs, where the asymptotic approximation becomes accurate, it is computationally much more efficient to use the `Asymptotic Frobenius Test`.

We now study the power of the different tests when the number of samples $m = 1$. Here, we have a single sample G_1 of size $n = 100$ from the IER($P^{(n)}$) model, where $P^{(n)}$ is certain perturbation of the reference matrix $Q^{(n)}$. Note that, in this case, the `Asymptotic Frobenius Test` is no longer applicable, because the sum over the indices in (2.1) is vacuous. However, the `Asymptotic Operator-Norm Test` and the bootstrap tests are still applicable. We consider the following two cases:

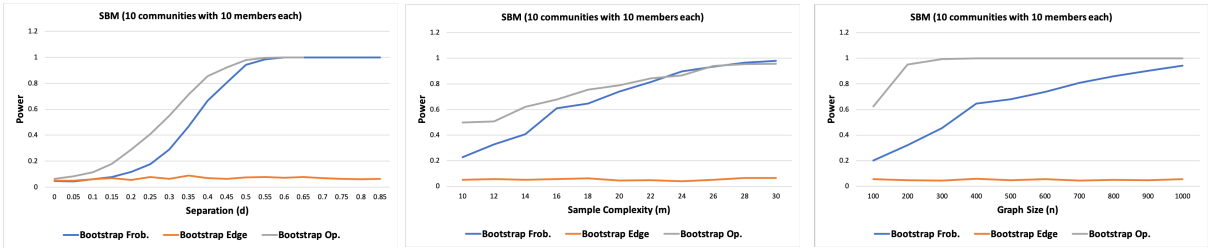
- In Figure 5(a) shows the empirical power (out of 100 iterations) as a function of Δ in the Erdős-Rényi model, when the reference matrix corresponds to ER($100, \frac{1}{2}$) and the matrix $P^{(n)}$ is chosen as ER($100, \frac{1}{2} + \Delta$). Here, the `Bootstrapped Operator-Norm Test` and the `Asymptotic Operator-Norm Test` (as in (3.4)) have the highest power. The `Bootstrapped Cycle Test`, on the other hand has power converging to 1 much slowly.
- In Figure 5(b) shows the empirical power (out of 100 iterations) as a function of Δ, in the β-model, where the reference matrix $Q^{(n)}$ corresponds to the β-model $\mathscr{B}(100, \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is chosen uniformly from the surface of ball in $\mathbb{R}^{100}$ with radius 20, and $P^{(n)}$ corresponds to the β-model $\mathscr{B}(100, \boldsymbol{\beta} + \Delta)$. Here, the `Bootstrapped Operator-Norm Test` and the `Bootstrapped Cycle Test` are more powerful, compared to the `Asymptotic Operator-Norm Test`.

In Appendix 5, we have additional simulations which compares the power of the different tests by varying the size of the graph n and the sample size m.

4.2. Graphon-Based IERs

Another natural way to generate a more general class of IERs is through graphons. A graphon is a measurable function $W : [0,1]^2 \to [0,1]$ that satisfy $W(x,y) = W(y,x)$, for all $x, y \in [0,1]$. These appear as limits of large dense graphs [24], and has found many applications in statistics, computer science, and and related areas. Given a graphon $W : [0,1]^2 \to [0,1]$, one can generate a IER with the edge probability matrix $P^{(n)}(a,b) = W(\frac{a}{n}, \frac{b}{n})$. This is a very general framework which includes all dense IERs (including the examples considered above).

In Table 1 we consider the power of the `Bootstrapped Frobenius Test`, `Boostrapped Operator`



(a) (b) (c)

Figure 6. Empirical power of the different tests in a sparse 10-block SBM: (a) $n = 100$, $m = 4$, varying separation d, (b) $n = 100$, $d = 0.3$, varying m, and (c) $m = 4$, $d = 0.3$, varying n.

m	Frob.	Op.	Edge
1	0.10	1	0.07
2	0.12	1	0.1
3	0.27	1	0.19
4	1	1	0.28
5	1	1	0.35

(a)

n	Frob.	Op.	Edge
100	1	1	0.23
200	1	1	0.26
300	1	1	0.26
400	1	1	0.23
500	1	1	0.21

(b)

Table 1. Power of graphon-based IER models for the Bootstrapped Frobenius, Operator, and Edge tests, for (a) $n = 100$ and varying m, and (b) $m = 4$ and varying n.

Test, and the Boostrapped Edge Test where the reference matrix $Q^{(n)}$ corresponds to $\mathrm{ER}(n, \frac{1}{4})$ and

$$P^{(n)}(a, b) = \frac{ab}{n^2}, \quad \text{for } 1 \le a \ne b \le n,$$

that is, the graphon $W(x, y) = xy$. In Table 1(a) the graph size is fixed at $n = 100$ and the sample size m varies, and in Table 1(b) the sample size is fixed at $m = 4$, while the graph size varies. Note that in both the models the expected number of edges is asymptotically $\frac{n^2}{8}$, hence the edge-test has very low power here. On the other hand, the proposed Frobenius and Operator norm tests perform very well, illustrating yet again the robustness of these methods.

4.3. Sparser Networks

In this section we illustrate the performance of the proposed tests in the sparse regime, where the edge connection probabilities can depend on the size n of the graph. To this end, we chose the reference matrix $Q^{(n)}$ corresponding to a 10-block stochastic block model (SBM), where the inter-community edge probabilities are set to 0 and the intra-community edge probabilities is set to $\frac{1}{\sqrt{n}}$, while for $P^{(n)}$ we keep the inter-community edge probability as 0, but change the intra-community edge probability to $\frac{1+d}{\sqrt{n}}$ for half of the communities, and $\frac{1-d}{\sqrt{n}}$ for the other half of the communities (see Figure 7). In Figure 6(a) the empirical power of the different tests for graphs of size $n = 100$, sample size $m = 4$, and varying d is shown. Figure 6(b) fixes $d = 0.3$, graph size $n = 100$, and varies the sample size m, while Figure 6(c) fixes $d = 0.3$, sample size $m = 4$, and varies the graph size n from 100 to 1000. Here, as expected, the edge test is powerless, because the expected number of edges in $Q^{(n)}$ and $P^{(n)}$ are the same. On the other hand, the proposed tests have power converging to 1, illustrating the usefulness of these methods for sparser IER graphs.

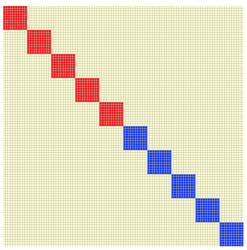


Figure 7. The structure of the adjacency matrix of a 10 block SBM.

5. Conclusion and Future Work

In this paper, motivated by the problem of determining whether a sample of real-world networks, fits a reference model, we developed goodness-of-fit tests for inhomogeneous random graphs (IERs). We proposed tests attaining minimax optimal sample complexities, for testing under different matrix norms, such as the Frobenius norm and the Operator norm. We also proposed practical implementations of the tests, using the asymptotic distribution and the parametric bootstrap. The proposed tests outperform the baseline tests, in a wide range of simulation experiments, illustrating the broad applicability and robustness of these methods.

References

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

[2] G. Bresler and D. Nagaraj, Optimal single sample tests for structured versus unstructured network data, *Conference on Learning Theory (COLT)*, 1657–1690, 2018.

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

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

[5] S. Chatterjee, P. Diaconis, and A. Sly, Random graphs with a given degree sequence, *Annals of Applied Probability*, Vol. 21 (4), 1400–1435, 2011.

[6] F. Chung and L. Lu, Connected components in random graphs with given expected degree sequences, *Annals of Combinatorics*, Vol. 6, 125–145, 2002.

[7] V. Csiszár, P. Hussami, J. Komlós, T. F. Móri, L. Rejtõ, and G. Tusnády, *Algorithms*, Testing goodness of fit of random graph models, Vol. 5 (4), 629–635, 2012.

[8] C. Daskalakis, N. Dikkala, and G. Kamath, Testing Ising models, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1989–2007, 2018.

[9] A. Elliott, E. Leicht, A. Whitmore, G. Reinert, and F. Reed-Tsochas, A nonparametric significance test for sampled networks. *Bioinformatics*, 34(1), 64-71, 2017.

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

[11] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1994.

[12] D. Ghoshdastidar and U. von Luxburg, Practical methods for graph two-sample testing, *Neural Information Processing Systems (NeurIPS)*, 3019-3028, 2018.

[13] D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg, Two-sample tests for large random graphs using network statistics, *Conference on Learning Theory (COLT)*, 954–977, 2017

[14] D. Ghoshdastidar, M. Gutzeit, A. Carpentier, and U. von Luxburg, Two-sample hypothesis testing for inhomogeneous random graphs, *Annals of Statistics*, to appear, 2020.

[15] 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*, Vol. 11(2), 725–750, 2017.

[16] R. A. Hanneman and M. and Riddle, *Introduction to social network methods*, 2005. (http://faculty.ucr.edu/~hanneman/nettext/)

[17] P. Holland and S. Leinhardt, An exponential family of probability distributions for directed graphs, *Journal of the American Statistical Association*, Vol. 76, 33–65, 1981.

[18] D. R. Hunter, MM algorithms for generalized Bradley–Terry models, *Annals of Statistics*, Vol. 32, 384–406, 2004.

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

[20] S. Janson, T. Luczak, and A. Rucinski, *Random Graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, 2000.

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

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

[23] Y. Li and K. C. Carriere. Assessing goodness of fit of exponential random graph models, *International Journal of Statistics and Probability*, Vol. 2 (4), 64–74, 2013.

[24] L. Lovász, *Large networks and graph limits*, AMS, Providence, RI, 2012.

[25] E. Mossel, J. Neeman, and A. Sly, Consistency thresholds for the planted bisection model, *Electronic Journal of Probability*, Vol. 21, 1–24, 2016.

[26] R. Mukherjee, S. Mukherjee, S. Sen, Detection thresholds for the beta model in sparse graphs, *Annals of Statistics*, Vol. 46, 1288–1317, 2018.

[27] L. Ospina-Forero, C.M. Deane, and G. Reinert, Assessment of model fit via network comparison methods based on subgraph counts, *Journal of Complex Networks*, Vol. 2, 226–253, 2019.

[28] J. Park and M. E. J. Neuman, Statistical mechanics of networks, *Phys. Rev. E (3)*, Vol. 70 066117, 13, 2004.

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

[30] 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*, Vol. 26 (2), 344–354, 2016.

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