# A Stein Goodness-of-fit Test for Exponential Random Graph Models

**Wenkai Xu**
Gatsby Computational Neuroscience Unit
University College London

**Gesine Reinert**
Department of Statistics
Oxford University

## Abstract

We propose and analyse a novel nonparametric goodness-of-fit testing procedure for exchangeable exponential random graph models (ERGMs) when a single network realisation is observed. The test determines how likely it is that the observation is generated from a target unnormalised ERGM density. Our test statistics are derived from a kernel Stein discrepancy, a divergence constructed via Stein's method using functions in a reproducing kernel Hilbert space, combined with a discrete Stein operator for ERGMs. The test is a Monte Carlo test based on simulated networks from the target ERGM. We show theoretical properties for the testing procedure for a class of ERGMs. Simulation studies and real network applications are presented.

## 1 INTRODUCTION

Complex data from many application areas are often represented as networks, and probabilistic network models help to understand the expected behaviour of such networks. In social science, exponential random graph models (ERGMs) have been successfully employed for this task, see for example Wasserman and Faust [1994]. ERGMs can be viewed as exponential family models or energy-based models, and as typical for such models, statistical inference for ERGMs suffers from intractable normalisation constants. Monte Carlo methods for parameter estimations in ERGMs alleviate this issue [Snijders, 2002], and model diagnoses via maximum likelihood (MLE) and maximum pseudo-likelihood are developed [Morris et al., 2008]. Statistical properties with particular attention to the normalisation constant are studied in Chatterjee and Diaconis [2013]. For analysing distributions with intractable normalisation constants, Stein's method [Barbour and Chen, 2005; Chen et al., 2010] provides a promising approach [Chwialkowski et al., 2016; Liu and Wang, 2016; Bresler and Nagaraj, 2019]. In Reinert and Ross [2019], Stein's method is developed for ERGMs but not yet applied to goodness-of-fit tests.

Goodness-of-fit tests for random graph models address the problem of whether the proposed model generates the observed network(s), and play a key role in understanding and interpreting network structures in real-world applications. A main issue is that replicates are usually not available; the data are represented as only one network. Standard goodness-of-fit tests for ERGMs to date rely on Monte Carlo tests for particular summary statistics such as vertices degrees [Ouadah et al., 2020], motifs or subgraph counts [Bhattacharyya and Bickel, 2015; Ospina-Forero et al., 2019; Chen and Onnela, 2019], or spectral properties [Shore and Lubin, 2015]. The goodness-of-fit tests for ERGM in Hunter et al. [2008] or Schweinberger [2012] also assess the model assumptions via graphical assessments. Lospinoso and Snijders [2019] combines such statistics into a Mahalanobis-type distance which is assessed via Monte Carlo tests. The consistency of type 1 error and the power of the test have not yet been systematically investigated.

Nonparametric goodness-of-fit tests based on Stein operators [Gorham and Mackey, 2015; Ley et al., 2017] and functions in a reproducing kernel Hilbert space (RKHS) [Berlinet and Thomas, 2004] for data with replicates build on a kernel Stein discrepancy (KSD) that utilises the strength of a Stein operator to treat unnormalised models and optimises over test functions in a rich enough RKHS to best distinguish the data from the model distributions. Such schemes are consistent and have high test power in various scenarios, including multivariate distributions [Chwialkowski et al., 2016; Liu et al., 2016], discrete distributions [Yang et al., 2018], point processes [Yang et al., 2019], directional distributions [Xu and Matsuda, 2020], and

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. 

censored data [Fernandez et al., 2020]. These scenarios are typically based on i.i.d. samples from the distributions. In addition, the properties of kernel mean embeddings [Berlinet and Thomas, 2004; Muandet et al., 2017] enable the extraction of distributional features to perform model comparison and model criticism [Jitkrittum et al., 2017b, 2018; Kanagawa et al., 2019; Jitkrittum et al., 2020].

Here we propose a novel goodness-of-fit testing procedure for ERGMs combining a Stein operator for ERGMs and functions in an RKHS. The class of ERGMs treated here are undirected networks which, when the number of vertices tends to infinity, can be approximated by a suitably chosen Bernoulli random graph, with edge probability parameter that generally does not equal the MLE. The test is based on only *one* observed network and estimates the Stein operator through re-sampling edge indicators. This test compares the test statistics from one observed network to the simulated distribution of the statistics under the null model. As the Stein operator characterises the target distribution for this class of ERGMs, under a member of this class serving as null hypothesis we derive theoretical results for the test statistic. We also provide a theoretical justification of the proposed re-sampling procedure.

To assess the performance of the test, we use simulated data as well as three real-world applications: Lazega's lawyer network [Lazega, 2001], a teenager friendship network [Steglich et al., 2006], and a larger network of legislation co-sponsorship [Fowler, 2006a,b]. We find that on synthetic data, our test is more reliable and has higher power than the standard tests even when only a small number of edge indicators is re-sampled. Moreover, the test can be applied to networks on more vertices than its competitor tests. For the lawyer network, we confirm the suggestion by Lazega [2001] of a Bernoulli random graph; for the friendship network we do not reject an ERGM with edges, two-stars and triangles as statistics. For the co-sponsorship network we do not reject a Bernoulli random graph fit whereas the model proposed in Schmid and Desmarais [2017] is rejected at level $\alpha = 0.05$.

The paper is structured as follows. We begin our presentation with a short review on ERGM, KSD and the ERGM Stein operator in Section 2. Section 3 introduces our re-sampling Stein operator for ERGM, Our goodness-of-fit testing procedure which is based on what we call the *graph kernel Stein statistic* (gKSS), and the relevant theoretical results, are given in Section 4. In Section 5, we illustrate the test performances on synthetic data as well as real network applications.

All the proofs are deferred to the Supplementary Material. The Supplementary Material also contains more discussions, details for experiment settings and additional experimental results, as well as a detailed comparison to the correesponding test in Yang et al. [2018]. The code and data sets for the experiments are available at `https://github.com/clemon13/gkss.git`.

## 2 BACKGROUND

### 2.1 Exponential Random Graph Models

Exponential random graph models (ERGM) are frequently used as parametric models for social network analysis [Wasserman and Faust, 1994; Holland and Leinhardt, 1981; Frank and Strauss, 1986]; they include Bernoulli random graphs as well as stochastic blockmodels as special cases. Here we restrict attention to undirected, unweighted simple graphs on $n$ vertices, without self-loops or multiple edges. To define such an ERGM, we introduce the following notations.

Let $\mathcal{G}_n^{lab}$ be a set of vertex-labeled graphs on $n$ vertices and, for $N = n(n-1)/2$, encode $x \in \mathcal{G}_n^{lab}$ by an ordered collection of $\{0,1\}$ valued variables $x = (x_{ij})_{1\le i<j\le n} \in \{0,1\}^N$ where $x_{ij} = 1$ if and only if there is an edge between $i$ and $j$. For a graph $H$ on at most $n$ vertices, let $V(H)$ denote the vertex set, and for $x \in \{0,1\}^N$, denote by $t(H,x)$ the number of *edge-preserving* injections from $V(H)$ to $V(x)$; an injection $\sigma$ preserves edges if for all edges $vw$ of $H$ with $\sigma(v) < \sigma(w)$, $x_{\sigma(v)\sigma(w)} = 1$. For $v_H = |V(H)| \ge 3$ set

$$t_H(x) = \frac{t(H,x)}{n(n-1)\cdots(n-v_H+3)}.$$

If $H = H_1$ is a single edge, then $t_H(x)$ is twice the number of edges of $x$. In the exponent this scaling of counts matches [Bhamidi et al., 2011, Definition 1] and [Chatterjee and Diaconis, 2013, Sections 3 and 4]. An ERGM for the collection $x \in \{0,1\}^N$ can be defined as follows, see Reinert and Ross [2019].

**Definition 1.** *Fix $n \in \mathbb{N}$ and $k \in \mathbb{N}$. Let $H_1$ be a single edge and for $l = 2,\dots,k$ let $H_l$ be a connected graph on at most $n$ vertices; set $t_l(x) = t_{H_l}(x)$. For $\beta = (\beta_1,\dots,\beta_k)^\top \in \mathbb{R}^k$ and $t(x) = (t_1(x),\dots,t_k(x))^\top \in \mathbb{R}^k$ $X \in \mathcal{G}_n^{lab}$ follows the exponential random graph model $X \sim \mathrm{ERGM}(\beta,t)$ if for $\forall x \in \mathcal{G}_n^{lab}$,*

$$\mathbb{P}(X = x) = \frac{1}{\kappa_n(\beta)} \exp\left(\sum_{l=1}^{k} \beta_l t_l(x)\right).$$

*Here $\kappa_n(\beta)$ is the normalisation constant.*

The vector $\beta \in \mathbb{R}^k$ is the parameter vector and the statistics $t(x) = (t_1(x),\dots,t_k(x))^\top \in \mathbb{R}^k$ are sufficient statistics. Moreover, exchangeability holds; letting

$[n] := \{1, \dots, n\}$, for any permutation $\sigma : [N] \to [N]$, $\mathbb{P}(x_1, \dots, x_N) = \mathbb{P}(x_{\sigma(1)}, \dots, x_{\sigma(N)})$.

Many random graph models can be set in this framework. The simplest example is the Bernoulli random graph (ER graph) with edge probability $0 < p < 1$; in this case, $l = 1$ and $H_1$ is a single edge. ERGMs can use other statistic in addition to subgraph counts, and many ERGMs model directed networks. Moreover, ERGMs can model network with covariates such as using dyadic statistics to model group interactions between vertices [Hunter et al., 2008]. Here we restrict attention to the case which is treated in Reinert and Ross [2019] because it is for this case that a Stein characterization is available.

As the network size increases, the number of possible network configurations increases exponentially in the number of possible edges, making the normalisation constant $\kappa_n(\beta)$ usually prohibitive to compute in closed form. Statistical inference on ERGM mainly relies on MCMC type methods that utilise the density ratio between proposed state and current state, where the normalisation constant cancels.

### 2.2 Kernel Stein Discrepancies

We briefly review the notion of kernel Stein discrepancy (KSD) for continuous distributions [Gorham and Mackey, 2015; Ley et al., 2017] and its associated statistical test [Chwialkowski et al., 2016; Liu et al., 2016].

Let $q$ be a smooth probability density on $\mathbb{R}^d$ that vanishes at the boundary. The operator $\mathcal{T}_q : (\mathbb{R}^d \to \mathbb{R}^d) \to (\mathbb{R}^d \to \mathbb{R})$ is called a *Stein operator* if the following *Stein's identity* holds: $\mathbb{E}_q[\mathcal{T}_q f] = 0$, where $f = (f_1, \dots, f_d) : \mathbb{R}^d \to \mathbb{R}^d$ is any bounded smooth function. The Stein operator $\mathcal{T}_q$ for continuous density [Chwialkowski et al., 2016; Liu et al., 2016] is defined as

$$\mathcal{T}_q f(x) = \sum_{i=1}^{d} \left( f_i(x) \frac{\partial}{\partial x^i} \log q(x) + \frac{\partial}{\partial x^i} f_i(x) \right). \quad (1)$$

This Stein operator is also called Lagenvin-diffusion Stein operator [Barp et al., 2019]. Since $q$ is assumed to vanish at the boundary and $f$ is bounded, the Stein identity holds due to integration by parts. As the Stein operator $\mathcal{T}_q$ only requires the derivatives of $\log q$ and thus does not involve computing the normalisation constant of $q$, it is useful for dealing with unnormalised models [Hyvärinen, 2005].

A suitable class of functions $\mathcal{F}$ is such that if $\mathbb{E}_p[\mathcal{T}_q f] = 0$ for all functions $f \in \mathcal{F}$, then $p = q$ follows. It is convenient to take $\mathcal{F} = B_1(\mathcal{H})$, the unit ball of a large enough RKHS $\mathcal{H}$. In particular, the kernel Stein discrepancies (KSD) between two densities $p$ and $q$ based on $\mathcal{T}_q$ is defined as

$$\mathrm{KSD}(p\|q, \mathcal{H}) = \sup_{f \in B_1(\mathcal{H})} \mathbb{E}_p[\mathcal{T}_q f]. \quad (2)$$

Under mild regularity conditions, $\mathrm{KSD}(p\|q, \mathcal{H}) \geq 0$ and $\mathrm{KSD}(p\|q, \mathcal{H}) = 0$ if and only if $p = q$ [Chwialkowski et al., 2016], making KSD a proper discrepancy measure between probability densities.

The KSD in Eq.(2) can be used for testing the model goodness-of-fit as follows. One can show that $\mathrm{KSD}^2(p\|q, \mathcal{H}) = \mathbb{E}_{x,\tilde{x} \sim p}[h_q(x, \tilde{x})]$, where $x$ and $\tilde{x}$ are independent random variables with density $p$ and $h_q(x, \tilde{x})$ is given in explicit form which does not involve $p$;

$$h_q(x, \tilde{x}) = \sum_{i=1}^{d} \left\langle \frac{\partial \log q(x)}{\partial x^i} k(x, \cdot) + \frac{\partial k(x, \cdot)}{\partial x^i}, \frac{\partial \log q(\tilde{x})}{\partial \tilde{x}^i} k(\tilde{x}, \cdot) + \frac{\partial k(\tilde{x}, \cdot)}{\partial \tilde{x}^i} \right\rangle_{\mathcal{H}}. \quad (3)$$

Suppose we have a set of samples $\{x_1, \dots, x_n\}$ from an unknown density $p$ on $\mathbb{R}^d$ and the goodness-of-fit test aims to check whether $p = q$. Then $\mathrm{KSD}^2(p\|q, \mathcal{H})$ can be empirically estimated by independent samples from $p$ using a U-statistics or V-statistics. The critical value is determined by bootstrap based on weighted chisquare approximations for U-statistics or V-statistics.

For goodness-of-fit test of discrete distributions, Yang et al. [2018] proposed a kernel discrete Stein discrepancy (KDSD). Essentially, the differential operator in Eq.(1) is replaced by an appropriately defined difference operator. KDSD is a useful method for assessing goodness-of-fit of ERGMs (as discrete random objects) when a large set of networks are observed [Yang et al., 2018, Figure 1(d)], but is not applicable when only one single network is observed. More details can be found in the Supplementary Material F.

### 2.3 The ERGM Stein Operator

Instead of using the Stein operator from Yang et al. [2018] we employ the Stein operator from Reinert and Ross [2019]. With $N = n(n-1)/2$ let $e_s \in \{0, 1\}^N$ be a vector with 1 in coordinate $s$ and 0 in all others; $x^{(s,1)} = x + (1 - x_s)e_s$ has the $s$-entry replaced of $x$ by the value 1, and $x^{(s,0)} = x - x_s e_s$ has the $s$-entry of $x$ replaced by the value 0; moreover, $x_{-s}$ is the set of edge indicators with entry $s$ removed. Then a (Glauber dynamics) Markov process on $\{0, 1\}^N$ is introduced with transition probabilities

$$\mathbb{P}(x \to x^{(s,1)}) = \frac{1}{N} - \mathbb{P}(x \to x^{(s,0)}) = \frac{1}{N} q_X(x^{(s,1)}|x)$$

where $q_X(x^{(s,1)}|x_{-s}) = \mathbb{P}(X_s = 1|X_{-s} = x_{-s})$. For the ERGM$(\beta, t)$ from Definition 1,

$$q(x^{(s,1)}|x_{-s}) := \exp\left\{\sum_{\ell=1}^{k} \beta_\ell t_\ell(x^{(s,1)})\right\} \times \left(\exp\left\{\sum_{\ell=1}^{k} \beta_\ell t_\ell(x^{(s,1)})\right\} + \exp\left\{\sum_{\ell=1}^{k} \beta_\ell t_\ell(x^{(s,0)})\right\}\right)^{-1}$$

and similarly the probability of the new edge being absent exchanges 1 and 0 in this formula to give $q(x^{(s,0)}|x_{-s})$. For $h : \{0,1\}^N \to \mathbb{R}$ let

$$\Delta_s h(x) = h(x^{(s,1)}) - h(x^{(s,0)}).$$

The generator $\mathcal{T}_{\beta,t}$ of this Markov process is the desired Stein operator and its expression simplifies to

$$\mathcal{T}_{\beta,t} f(x) = \frac{1}{N} \sum_{s\in[N]} \mathcal{T}_q^{(s)} f(x) \tag{4}$$

with

$$\mathcal{T}_q^{(s)} f(x) = q(x^{(s,1)}|x_{-s})\Delta_s f(x) + \Big(f(x^{(s,0)}) - f(x)\Big). \tag{5}$$

When the ERGM is such that the Markov process is irreducible, then its stationary distribution is unique, and if $\mathbb{E}_p[\mathcal{T}_{\beta,t} f] = 0$ for all smooth test functions $f$, then $p$ is the distribution of ERGM$(\beta, t)$. Thus, the Stein operator characterises ERGM$(\beta, t)$. Moreover, for each $s \in [N]$,

$$\mathbb{E}_q \mathcal{T}_q^{(s)} f = 0. \tag{6}$$

To see this, write

$$\mathbb{E}_q \mathcal{T}_q^{(s)} f = \sum_x q(x_{-s}) \Big(q(x^{(s,1)}|x_{-s})\mathcal{T}_q^{(s)} f(x^{(x,1)}) + q(x^{(s,0)}|x_{-s})\mathcal{T}_q^{(s)} f(x^{(x,0)})\Big).$$

Substituting $x^{(s,1)}$ and $x^{(s,0)}$ in (5) gives

$$\begin{aligned} &q(x^{(s,1)}|x_{-s})\mathcal{T}_q^{(s)} f(x^{(s,1)}) + q(x^{(s,0)}|x_{-s})\mathcal{T}_q^{(s)} f(x^{(s,0)}) \\ &= q(x^{(s,1)}|x_{-s})q(x^{(s,0)}|x_{-s})\Big(f(x^{(s,0)}) - f(x^{(s,1)}) \\ &\quad + f(x^{(s,1)}) - f(x^{(s,0)})\Big) = 0 \end{aligned}$$

and Eq.(6) follows.

Next, we introduce our kernel Stein statistic for testing the goodness-of-fit of an ERGM based on a single observed network as well as an estimator for it which is based on re-sampling of edge indicators.

# 3 KERNEL STEIN STATISTICS from RE-SAMPLING

**Kernel Stein Statistics** Based on the Stein operator representation Eq.(4), we develop the kernel Stein statistics (KSS)[1] for ERGMs. Similar to KSD in Eq.(2), we use the functions in a unit ball of an RKHS $\mathcal{H}$ as test functions.

The Stein operator in Eq.(4) can be written as expectation over edge variables $S$ with uniform probability $\mathbb{P}(S = s) = \frac{1}{N}, \forall s \in [N] := \{1, \ldots, N\}$, independently of $x$, namely

$$\mathcal{T}_q f(x) = \sum_{s\in[N]} \mathbb{P}(S = s)\mathcal{T}_q^{(s)} f(x) =: \mathbb{E}_S[\mathcal{T}_q^{(S)} f(x)]. \tag{7}$$

Note that the expectation is taken over $S$, with the network $x$ fixed except for the coordinate $S$.

After algebraic manipulations, Eq (5) has the form

$$\begin{aligned} &\mathcal{T}_q^{(s)} f(x) \\ &= q(x^{(s,1)}|x_{-s})f(x^{(s,1)}) + q(x^{(s,0)}|x_{-s})f(x^{(s,0)}) - f(x) \\ &= \mathbb{E}_{\{0,1\}}[f(X_s, x_{-s})|x_{-s})] - f(x). \end{aligned} \tag{8}$$

Here $\mathbb{E}_{\{0,1\}}$ refers to the expectation taken only over the value which $X_s$ takes on. Hence,

$$\mathcal{T}_q f(x) = \mathbb{E}_S\left[\mathbb{E}_{\{0,1\}}[f(X_s, x_{-s}|x)]\right] - f(x). \tag{9}$$

For a fixed network $x$, we seek a function $f \in \mathcal{H}$, s.t. $\|f\|_{\mathcal{H}} \leq 1$, that best distinguishes the difference in Eq.(9) when $X$ does not have distribution $q$; this rationale is similar as for Eq.(2). We define the *graph kernel Stein statistics* (gKSS) as

$$\text{gKSS}(q; x) = \sup_{\|f\|_{\mathcal{H}}\leq 1} \Big|\mathbb{E}_S[\mathcal{T}_q^{(S)} f(x)]\Big|. \tag{10}$$

It is often more convenient to consider $\text{gKSS}^2(q; x)$. Let the RKHS $\mathcal{H}$ have kernel $K$ and inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. By the reproducing property of RKHS functions, as for Eq.(3), algebraic manipulation allows the supremum to be computed in closed form:

$$\text{gKSS}^2(q; x) = \frac{1}{N^2} \sum_{s,s'\in[N]} h_x(s, s') \tag{11}$$

where $h_x(s, s') = \left\langle \mathcal{T}_q^{(s)} K(x, \cdot), \mathcal{T}_q^{(s')} K(\cdot, x)\right\rangle_{\mathcal{H}}$.

[1] We avoid calling it a *discrepancy* since our expectation is not taken over all ERGM samples as described in Yang et al. [2018], but instead based on a single network.

**Stein Operator from Edge Re-sampling** When the distribution of $X$ is known, the expectation in Eq.(9) can be computed for networks with a small number of vertices, but when the number of vertices is large, exhaustive evaluation is computationally intensive. For a fixed network $x$, we propose the following randomised Stein operator via edge re-sampling. This procedure mimics the Markov process which gives rise to the Stein operator. Let $B$ be the fixed number of edges to be re-sampled. Our re-sampled Stein operator is

$$\widehat{\mathcal{T}_q^B} f(x) = \frac{1}{B} \sum_{b\in[B]} \mathcal{T}_q^{(s_b)} f(x) \tag{12}$$

where $b \in B$ and $s_b$ are edge samples from $\{1,\dots,N\}$, chosen uniformly with replacement, independent of each other and of $x$. The expectation of $\widehat{\mathcal{T}_q^B} f(x)$ with respect to the re-sampling is

$$\mathbb{E}_B[\widehat{\mathcal{T}_q^B} f(x)] = \mathbb{E}_S[\mathcal{T}_q^{(S)} f(x)] = \mathcal{T}_q f(x).$$

We introduce the corresponding re-sampling gKSS:

$$\widehat{\mathrm{gKSS}}(q;x) = \sup_{\|f\|_{\mathcal{H}}\le 1} \Big| \frac{1}{B} \sum_{b\in[B]} \mathcal{T}_q^{(s_b)} f(x) \Big|. \tag{13}$$

This is a stochastic Stein discrepancy, see Gorham et al. [2020]. The supremum in Eq.(13) is achieved by

$$f^*(\cdot) = \frac{\frac{1}{B}\sum_b \mathcal{T}_q^{(s_b)} k(x,\cdot)}{\left\| \frac{1}{B}\sum_a \mathcal{T}_q^{(s_a)} k(x,\cdot) \right\|}.$$

Similar algebraic manipulations as for Eq.(11) yield

$$\widehat{\mathrm{gKSS}}^2(q;x) = \frac{1}{B^2} \sum_{b,b'\in[B]} h_x(s_b, s_{b'}). \tag{14}$$

# 4 GOODNESS-OF-FIT TEST with KERNEL STEIN STATISTICS

## 4.1 Goodness-of-fit Testing Procedures

We now describe the proposed procedure to assess the goodness-of-fit of an ERGM for a single network observation. The ERGM can be readily simulated from an unnormalised density via MCMC, see for example Morris et al. [2008]. Suppose that $q$ is the distribution of ERGM$(\beta, t)$ and $x$ is the observed network for which we want to assess the fit to $q$. We simulate independent networks $z_1,\dots,z_m \sim q$ and compare the observed $\widehat{\mathrm{gKSS}}^2(q;x)$ with the set of $\widehat{\mathrm{gKSS}}^2(q;z_i), i=1,\dots,m$ using a Monte Carlo test. As gKSS assesses the deviation from the null distribution, the test is one-sided; we reject the null model when the observed $\widehat{\mathrm{gKSS}}$ is large. The detailed test procedure is given in Algorithm 1.

**Algorithm 1** Kernel Stein Test for ERGM

**Input:**
Observed network $x$;
Null model $q$;
RKHS Kernel $K$;
Re-sample size $B$;
Confidence level $\alpha$;
Number of simulated networks $m$;

**Objective:**
Test $H_0 : x \sim q$ versus $H_1 : x \not\sim q$.

**Test procedure:**
1: Sample $\{s_1,\dots,s_B\}$ with replacement uniformly from $[N]$.
2: Compute $\tau = \widehat{\mathrm{gKSS}}^2(q;x)$ in Eq.(14).
3: Simulate $z_1,\dots,z_m \sim q$.
4: Compute $\tau_i = \widehat{\mathrm{gKSS}}^2(q;z_i)$ in Eq.(14). again with re-sampling, choosing new samples $\{s_{1,i},\dots,s_{B,i}\}$ uniformly from $[N]$ with replacement.
5: Estimate the empirical $(1-\alpha)$-quantile $\gamma_{1-\alpha}$ of $\tau_1,\dots,\tau_m$.

**Output:**
Reject $H_0$ if $\tau > \gamma_{1-\alpha}$; otherwise do not reject.

## 4.2 Kernel Choices

**Graph kernels** Apart from using simple kernels between adjacency vectors in $\{0,1\}^N$, we apply graph kernels that take into account graph topology via various measures. Various aspects of graph kernels have been studied [Borgwardt and Kriegel, 2005; Vishwanathan et al., 2010; Shervashidze et al., 2011; Sugiyama and Borgwardt, 2015]. We provide a brief review of some graph kernels in the Supplementary Material B. In our implementation in R, we utilise the `ergm` package related to Morris et al. [2008] for simulating ERGMs and the `graphkernels` package associated with Sugiyama et al. [2018] for computing relevant graph kernels.

**Vector-valued RKHS** As the operator in Eq.(9) has embedded a notion of conditional probability, we may tailor the RKHS accordingly. To incorporate the notion of $x_s$ conditioning on $x_{-s}$, we consider a separate treatment of $x_s$ and $x_{-s}$ and introduce a vector-valued reproducing kernel Hilbert Space (vvRKHS). Similar constructions are studied in Jitkrittum et al. [2020] when testing goodness-of-fit for conditional densities. In the Supplementary Material C, we provide a review on the vvRKHS we use as graph kernels; further details can be found in Caponnetto et al. [2008], Carmeli et al. [2010], or Sriperumbudur et al. [2011].

### 4.3 Theoretical Properties of gKSS

Let $X \sim q$ and $Y \sim \tilde{q}$, where $\tilde{q}$ is the distribution of an appropriately chosen ER graph. Our theoretical approximation argument has three steps: The first step, Theorem 1, is to approximate $\mathrm{gKSS}(q, X)$ by $\mathrm{gKSS}(\tilde{q}, Y)$, with an explicit bound on the approximation error, as the number of vertices $n \to \infty$. Secondly, Theorem 2 provides a normal approximation for $\mathrm{gKSS}(\tilde{q}, Y)^2$ of the approximating Bernoulli random graph as $n \to \infty$, again with an explicit bound, so that approximate confidence bounds for the test under the null hypothesis can be obtained. Finally, a normal approximation for $\widehat{\mathrm{gKSS}}(q, X)^2$ to a normal distribution with approximate mean $\mathrm{gKSS}(q, X)$, as $B \to \infty$ with $\lfloor B/N \rfloor$ fixed, is given in Proposition 2, again with an explicit error bound. These three results combined provide explicit control of the type 1 error.

In Chatterjee and Diaconis [2013] and Reinert and Ross [2019] it is shown that under some conditions on the parameters, an $\mathrm{ERGM}(\beta, t)$ is close to an appropriately chosen Bernoulli random graph, as follows. For $a \in [0, 1]$, define the following functions [Bhamidi et al., 2011; Eldan and Gross, 2018], with the notation in Definition 1 for $\mathrm{ERGM}(\beta, t)$:

$$\Phi(a) := \sum_{l=1}^{k} \beta_l e_l a^{e_l - 1}, \quad \varphi(a) := \frac{1 + \tanh(\Phi(a))}{2}$$

where $e_l$ is the number of edges in $H_l$. For a polynomial $f(x) = \sum_{\ell=0}^{k} c_\ell x^\ell$ set $|f(x)| := \sum_{\ell=1}^{k} |c_\ell| \, e_\ell \, x^\ell$. Moreover, $||f||$ denotes the supremum norm. The class of $\mathrm{ERGM}(\beta, t)$ in this section are assumed to satisfy the following standard technical assumption.

**Assumption 1.** (1) $\frac{1}{2}|\Phi|'(1) < 1$. (2) $\exists a^* \in [0, 1]$ *that solves the equation* $\varphi(a^*) = a^*$.

The value $a^*$ will be the edge probability in the approximating Bernoulli random graph, $\mathrm{ER}(a^*)$. Then the following result holds.

**Proposition 1.** [Theorem 1.7 and Corollary 1.10 [Reinert and Ross, 2019]] *Let* $\mathrm{ERGM}(\beta, t)$ *satisfy Assumption 1. Let* $X \sim \mathrm{ERGM}(\beta)$ *and* $Y \sim \mathrm{ER}(a^*)$. *Then for* $h : \{0, 1\}^N \to \mathbb{R}$,

$$|\mathbb{E}h(X) - \mathbb{E}h(Y)| \leq ||\Delta h|| N \frac{C_{a^*}(\beta, t, h)}{\sqrt{n}} \sum_{\ell=2}^{k} \beta_\ell.$$

*Here* $C_{a^*}(\beta, t, h)$ *is an explicit constant.*

Proposition 1 shows, that for large $n$, the ERGM can be approximated well by an appropriate ER graph for test functions $h$ which are properly scaled. In particular, if $H$ is a connected graph and $h(x) = t(H, x)n^{-|V(H)|}$, then there is an explicit constant $C = C(\beta, t, H)$ such that $|\mathbb{E}h(X) - \mathbb{E}h(Y)| \leq C/\sqrt{n}$. This result translates into an approximation for the gKSS, as follows.

**Theorem 1.** *Let* $q(x) = \mathrm{ERGM}(\beta, t)$ *satisfy Assumption 1 and let* $\tilde{q}$ *denote the distribution of* $ER(a^*)$. *For* $f \in \mathcal{H}$ *equipped with kernel* $K$, *let*

$$f_x^*(\cdot) = \frac{(\mathcal{T}_q - \mathcal{T}_{\tilde{q}})K(x, \cdot)}{\|(\mathcal{T}_q - \mathcal{T}_{\tilde{q}})K(x, \cdot)\|_{\mathcal{H}}}.$$

*Then there is an explicit constant* $C = C(\beta, t, K)$ *such that for all* $\epsilon > 0$,

$$\begin{aligned} &\mathbb{P}(|\,\mathrm{gKSS}(q, X) - \mathrm{gKSS}(\tilde{q}, Y)| \, > \, \epsilon) \\ &\leq \quad \Big\{||\Delta(\mathrm{gKSS}(q, \cdot))^2||(1 + ||\Delta\,\mathrm{gKSS}(q, \cdot)||) \\ &\qquad +4 \sup_x (||\Delta f_x^*||^2)\Big\} \binom{n}{2} \frac{C}{\epsilon^2 \sqrt{n}}. \end{aligned}$$

As our goodness-of-fit test statistic is based on the square of the gKSS, the asymptotic behaviour of $\mathrm{gKSS}^2(\tilde{q}, Y)$ is of interest. To approximate the distribution of $\mathrm{gKSS}^2$ under the null hypothesis we make some additional assumptions on kernel $K$ of RKHS.

**Assumption 2.** *Let* $\mathcal{H}$ *be the RKHS associated with the kernel* $K : \{0, 1\}^N \times \{0, 1\}^N \to \mathbb{R}$ *and for* $s \in [N]$ *let* $\mathcal{H}_s$ *be the RKHS associated with the kernel* $l_s : \{0, 1\} \times \{0, 1\} \to \mathbb{R}$. *Then*

*i)* $\mathcal{H}$ *is a tensor product RKHS,* $\mathcal{H} = \otimes_{s \in [n]} \mathcal{H}_s$;
*ii)* $k$ *is a product kernel,* $k(x, y) = \otimes_{s \in [N]} l_s(x_s, y_s)$;
*iii)* $\langle l_s(x_s, \cdot), l_s(x_s, \cdot)\rangle_{\mathcal{H}_s} = 1$;
*iv)* $l_s(1, \cdot) - l_s(0, \cdot) \neq 0$ *for all* $s \in [N]$.

These assumptions are satisfied for example for the suitably standardised Gaussian kernel $K(x, y) = \exp\{-\frac{1}{\sigma^2} \sum_{s \in [N]} (x_s - y_s)^2\}$.

Letting $|| \cdot ||_1$ denote $L_1$-distance, and $\mathcal{L}$ denote the law of a random variable, in Supplementary Material A, we show the following normal approximation.

**Theorem 2.** *Assume that the conditions i) - iv) in Assumption 2 hold. Let* $\mu = \mathbb{E}[\mathrm{gKSS}^2(\tilde{q}, Y)]$ *and* $\sigma^2 = \mathbb{V}ar[\mathrm{gKSS}^2(\tilde{q}, Y)]$. *Set* $W = \frac{1}{\sigma}(\mathrm{gKSS}^2(\tilde{q}, Y)] - \mu)$ *and let* $Z$ *denote a standard normal variable, Then there is an explicit constant* $C = C(a^*, l_s, s \in [N])$ *such that*

$$||\mathcal{L}(W) - \mathcal{L}(Z)||_1 \leq \frac{C}{\sqrt{N}}.$$

More details on $\mu$ and $\sigma^2$ are given in the Supplementary Material A. This normal approximation could also be used to assess the asymptotic distribution under an alternative $x \sim p$ where $p(x) = \mathrm{ERGM}(\beta', t')$ satisfies Assumption 1 with edge probability $b^*$ and $b^* \neq a^*$.

Then asymptotically we can compare the corresponding normal random variables with different means.

For the final step, the re-sampling version, let $k_s$ be the number of times that $s$ is included in the sample $\mathcal{B}$, where $|\mathcal{B}| = B$. Then, from Eq.(14),

$$\widehat{\text{gKSS}}^2(q;x) = \frac{1}{B^2}\sum_{s,s'\in[N]} k_s k_{s'} h_x(s,s').$$

In this expression, the randomness only lies in the counts $k_s$, where $s \in [N]$. These counts are exchangeable and $\mathbf{k} = (k_s, s \in [N])$ follows the multinomial $(B; N^{-1}, \ldots, N^{-1})$ distribution. Hence the statistic $\frac{1}{N^2}\sum_{s,t\in[N]} k_s k_t h_x(s,t)$ is a sum of weakly globally dependent random variables, although due to the network $x$ being fixed, this is not a classical $V$-statistic. Instead, Stein's method will be used to prove the following result in the Supplementary Material.

**Proposition 2.** *Let*

$$Y = \frac{1}{B^2}\sum_{s,t\in[N]} (k_s k_t - \mathbb{E}(k_s k_t)) h_x(s,t).$$

*Assume that $h_x$ is bounded such that $Var(Y)$ is non-zero. Then if $Z$ is mean zero normal with variance $Var(Y)$, there is an explicitly computable constant $C > 0$ such that for all three times continuously differentiable functions $g$ with bounded derivatives up to order 3,*

$$|\mathbb{E}[g(Y)] - \mathbb{E}[g(Z)]| \leq \frac{C}{B}.$$

When the sampling fraction $F = \frac{B}{N}$ is kept approximately constant as $N \to \infty$, noting that

$$\begin{aligned}\widehat{\text{gKSS}}^2(q;x) &= Y + \text{gKSS}^2 + \frac{N-1}{BN^2}\sum_{s\in[N]} h(s,s) \\ &\quad - \frac{1}{B^2N^2}\sum_{s,t\in[N],s\neq t} h(s,t)\end{aligned}$$

the normal approximation for $\widehat{\text{gKSS}}^2(q;x)$ with approximate mean $\text{gKSS}^2(q;x)$ follows for $N \to \infty$.

# 5 EXPERIMENTS

To assess the performance of the test, we replicate the synthetic benchmark-type settings from [Lusher et al., 2013; Rolls et al., 2015; Yang et al., 2018]. We then apply our tests to three real data networks: Lazega's lawyer network [Lazega, 2001] and a friendship network [Steglich et al., 2006] which are both studied in [Yin et al., 2019], as well as a co-sponsorship network from [Fowler, 2006a,b].

## 5.1 Synthetic Experiments

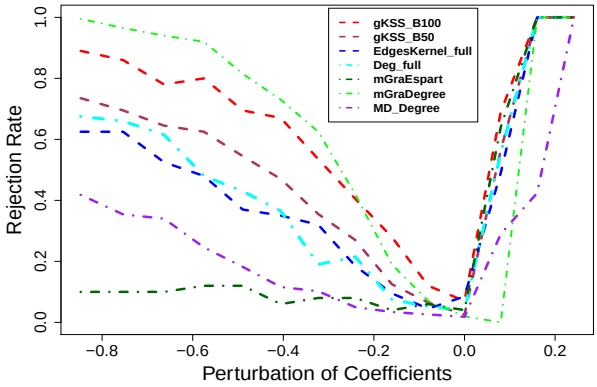
**Model** In the synthetic example, we assess the test performance on relatively simple but useful ERGMs, with three graphs $H_l$ in the statistic $t$, namely edge, 2-star, and triangle; we abbreviate this model as E2ST. Then the unnormalised density has the form

$$q(x) \propto \exp\Big(\beta_1 E_d(x) + \beta_2 S_2(x) + \beta_3 T_r(x)\Big), \quad (15)$$

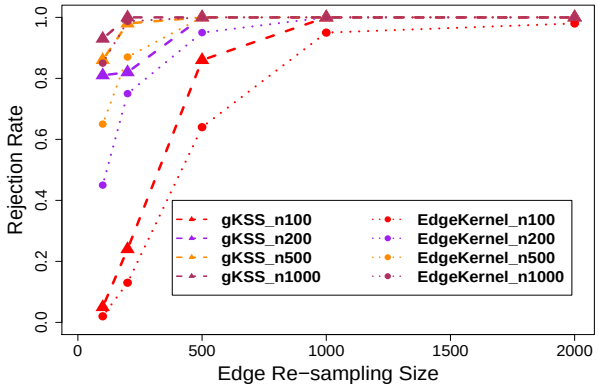
where $E_d(x)$ denotes the number of edges in $x$; $S_2(x)$ denotes the number of 2-stars in $x$ and $T_r(x)$ denotes the number of triangles in $x$. We choose the null parameter as $(\beta_1, \beta_2, \beta_3) = (-2, 0.0, 0.01)$, which satisfies Assumption 1 and gives $a^* = 0.1176$. For the alternative distributions, following similar settings in [Yang et al., 2018], we fix the coefficient $\beta_1 = -2$ and $\beta_3 = 0.01$ of the E2ST model in Eq.(15), and test the null model of $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$ with a perturbed $\beta_2$ so that the alternative model satisfies Assumption 1 also. For this model, Bresler and Nagaraj [2018] showed that even when using a (constant number of) i.i.d. samples from the network model are available, when $\beta_2\sqrt{n} \to 0$ as $n \to \infty$ then there is no test which can distinguish this model from the corresponding Bernoulli random graph. Hence poor power for $\beta_2$ close to 0 is to be expected.

**The Proposed Methods** We apply the proposed goodness-of-fit test procedures and compare with existing approaches. We use the following abbreviations: **gKSS** stands for the proposed test in Algorithm 1; **gKSS_B100** uses Eq.(14) as the test statistic where $B = 100$, and **gKSS_n20** denotes testing the problem with $n = 20$ vertices. Results shown in Fig. 1 are based on Weisfeiler-Lehman graph kernels [Shervashidze et al., 2011]; results using other kernels are shown in the Supplementary Material Section E. **EdgeKernel** denotes the gKSS with a kernel between binary edges which corresponds to a test based on edge counts. Re-sampling applies, e.g. **EdgeKernel_B100** indicates that 100 edges are re-sampled from the network.

**The Competing Approaches** We list the goodness-of-fit testing methods which serve as comparisons using the following abbreviations. **Degree_full** stands for degree-based tests [Ouadah et al., 2020], where the variance of degree counts on vertices are used as test statistics. The suffix "full" indicates that all vertices are used. The graphical tests for goodness-of-fit from Hunter et al. [2008] simulate the null distribution of a chosen network statistic from the null model as a visual guideline for goodness-of-fit. We quantify this idea by using total variation (TV) distance between distributions of network statistics of choice as test statistics;




(a) $n = 20$, $\alpha = 0.05$, $\beta_2$ perturbed




(b) $H_1 : \beta = (-2, -0.03, 0.01)$, $\alpha = 0.05$

Figure 1: Simulation Results for E2ST Model

**mGra** stands for the modified graphical test, where the TV distance is used to compare the distribution of the summary statistics of choice. Full details are provided in the Supplementary Material D. We append **mGra** by the summary statistics used, so that, for example, **mGraDegree** uses the TV distance between degree distributions as test statistics. **Espart** (or espartners) stands for edgewise shared partner [Hunter et al., 2008]; **MD_Degree** stands for the test based on Mahalanobis distance between chosen summary statistics [Lospinoso and Snijders, 2019]. The suffix after hyphen indicates that the vertex degree is used as network statistics.

**Test Results** The main results are shown in Fig. 1(a). We see that gKSS has higher power than the competitors, while, as expected, larger re-sampling size performs better. A denser networks can be easier to distinguish as higher subgraph counts are available compared to sparser networks. In our experimental set-up, the network size $n = 20$ is relatively small and the null model, with $\beta_1 = -2$, is fairly sparse. We observe that **mGraDegree** has slightly higher power than gKSS when $\beta_2 < -0.3$ so that the graph is sparser; it performs poorly when the alternative model is closer to the null, i.e. $|\beta_2|$ small [2]. This may relate to using the TV distance for comparing the degree distribution; the phenomenon does not occur for **MD_Degree**. Overall, **gKSS** is more reliable and has typically higher power compared to these competing methods.

**Increasing Edge Re-sampling Size B** Fig.1(b) shows the test power of large networks up to $n = 1000$ vertices. The results show that the tests achieve maximal power with a relatively small number of re-sampling edges indicators. With the choice of re-sampling size $B$ and good test power with a relatively small number of re-sampled edge indicators, gKSS is applicable to networks with a large number of vertices, beyond the reach of the graphical-based tests [Hunter et al., 2008]. In particular, the proposed tests can be useful in validating model assumptions in practical problems where the networks have a large number of vertices.

**Computational Time** The computational times for each test are shown in Table 1. The **gKSS** tests are faster than the **mGra** tests and of similar speed as the less accurate full degree method. The slow **mGra** tests are based on the computational demanding as well as hard-to-scale estimation associate with the graphical-based method in Hunter et al. [2008]. Its main computational cost stems from simulating the null graphs from `ergm` to compute the TV distances. Although the **Degree_full** test is supposed to be fast with computational complexity $O(n)$ , due to the estimation of the mean and variance of the degree statistics via simulating the null from `ergm`, its runtime is comparable with **gKSS_B50** with complexity $O(B^2)$ for $B = 100$.

| n | gKSS_B50 | gKSS_B100 | gKSS_B200 |
|---|---|---|---|
| 20 | 14.53 | 33.57 | 67.63 |
| 30 | 15.03 | 41.08 | 70.14 |
| 50 | 21.54 | 50.10 | 91.18 |
| n | Degree_full | mGraDegree | mGraEspart |
| 20 | 38.08 | 4596.67 | 4779.04 |
| 30 | 39.08 | 4840.66 | 4871.72 |
| 50 | 44.09 | 5127.74 | 5210.40 |

Table 1: The computational time for each test, in seconds, for 500 trials.

[2] In particular, it did not identify the slightly denser alternatives, which should be relatively easier problems.

### 5.2 Real Data Applications

Next we apply our test to two benchmark social network data sets which are analysed in Yin et al. [2019]; Lazega's lawyer network [Lazega, 2001] consists of a network between 36 laywers; the Teenager friendship network [Steglich et al., 2006] is a friendship data set of 50 secondary school students in Glasglow. Moreover, we apply our proposed test to large network, a co-sponsorship network for pieces of legislation in the U.S. Senate from Fowler [2006a,b]. The network data used here are from Schmid and Desmarais [2017] and consists of 2825 vertices and 28813 edges. For all three networks we fit an ER model with the maximum likelihood estimate as edge probability, an E2ST model, and an ER($a^*$) model using as edge probability $a^*$ calculated from the E2ST fit, or, for the co-sponsorship network, calculated from fitting an additional model detailed below. Table 2 summarises the results.

For the Lawyer network, Lazega [2001] suggests an ER model. Our test does not reject this null hypothesis when testing against the best fitted ER graph, with edge probability $p = 0.055$, which supports the assumed model. The fitted E2ST model with $\beta = (-2.8547, -0.0003, 0.6882)$ is rejected at $\alpha = 0.05$. This E2ST is close to an ER graph with $\beta_{E_d} = -2.774$ and the corresponding ER($a^*$) model is not rejected at significance level $\alpha = 0.05$.

For the Teenager network, the fitted ER model with $p = 0.046$ is rejected at $\alpha = 0.05$; for the fitted E2ST model in Eq.(15) with $\beta = (-2.3029, -0.3445, 2.8240)$ we do not have strong evidence to reject the null at $\alpha = 0.05$. The corresponding ER($a^*$) model is also not rejected at $\alpha = 0.05$. In particular the maximum likelihood estimator does not give the best fitting ER model.

The co-sponsorship network is well fitted by the ER graph with edge probability $p = 0.0072$. In contrast, the fitted E2ST with $\beta = (-6.4126, -0.0240, 2.4684)$, is rejected at $\alpha = 0.05$. Additionally we fit the ERGM proposed in Schmid and Desmarais [2017], which includes party homophily [Zhang et al., 2008] and the alternating $k$-star statistic [Snijders et al., 2006]:

$$q^*(x) \propto \exp\{\beta_1 E_d(x) + \beta_2 \Gamma(x; P) + \beta_3 \mathrm{S}_{\mathrm{alt}}(x; \lambda)\}, \tag{16}$$

where $P$ denotes the party assignment information between the pieces of legislations, and $\Gamma(x; P) = \sum_{ij} x_{ij} P_{ij}$; with the $k$-star count $S_k(x)$, the alternating $k$-star statistic is $\mathrm{S}_{\mathrm{alt}}(x; \lambda) = \sum_{k=2}^{n-1} (-\frac{1}{\lambda})^{k-2} S_k(x)$. We use the model $q^*$ with parameters fitted in Schmid and Desmarais [2017], $\beta_1 = -5.884$, $\beta_2 = 1.440$, $\beta_3 = 0.124$, and the parameter in alternating $k$-stars $\lambda = 0.4975$. This model (with p-value=0.022), as well as its corresponding ER($a^*$) model are rejected at significance level $\alpha = 0.05$.

| | $n$ | ER | E2ST | ER($a^*$) |
|---|---|---|---|---|
| Lawyer | 36 | 0.280 | 0.012 | 0.152 |
| Teenager | 50 | 0.016 | 0.060 | 0.336 |
| Co-sponsor | 2825 | 0.612 | 0.002 | 0.036 |

Table 2: Rejection rates for real networks. The results marked *blue* indicate not rejecting and *red* the null hypothesis at $\alpha = 0.05$, using $\widehat{\mathrm{gKSS}}^2$ with $B = 200$.

## 6 CONCLUSIONS AND DIRECTIONS FOR FURTHER WORK

In this paper we provide a novel goodness-of-fit test for exponential random graph models using Stein's method. A key feature is that the test relies on the observation of only one network. Probabilistic properties of the test statistic are analysed through comparison with Bernoulli random graphs.

Directions for future work include a thorough analysis of the interplay of the graph kernels used in the RKHS and the GKSD. Adaptive methods for tuning graph kernel hyper-parameters would be interesting; see for example Gretton et al. [2012] or Jitkrittum et al. [2017a].

Further, a large contribution to the computational cost of GKSD stems from sampling from the null model; an issue which affects all main methods for assessing goodness-of-fit for exponential random graph models. Developing a goodness-of-fit testing procedure based on a single network observation which does not require simulations from the null model is an exciting future challenge.

Finally, the approach is of independent interest and holds promise for adaptation to other random graph models.

**Acknowledgements** The authors would like to thank Arthur Gretton and Andrew Barbour for helpful discussions. Moreover, they thank the anonymous reviewers for many good comments which have improved the paper. G.R. acknowledges the support from EP/R018472/1 as well as from the COSTNET COST Action CA 15109. W.X. acknowledges the support from the Gatsby Charitable Foundation.

## References


A. Barbour and L. Chen. An introduction to Stein's method. *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*, 4, 2005.

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019.

A. Berlinet and C. Thomas. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. *The Annals of Applied Probability*, 21(6):2146–2170, 2011.

S. Bhattacharyya and P. J. Bickel. Subsampling bootstrap of count features of networks. *The Annals of Statistics*, 43(6):2384–2411, 2015.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

K. M. Borgwardt and H.-P. Kriegel. Shortest-path kernels on graphs. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE, 2005.

G. Bresler and D. Nagaraj. Optimal single sample tests for structured versus unstructured network data. In *Conference On Learning Theory*, pages 1657–1690. PMLR, 2018.

G. Bresler and D. Nagaraj. Stein's method for stationary distributions of Markov chains and application to Ising models. *The Annals of Applied Probability*, 29(5):3230–3265, 2019.

A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9(Jul):1615–1646, 2008.

C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461, 2013.

S. Chatterjee and E. Meckes. Multivariate normal approximation using exchangeable pairs. *Alea*, 4:257–283, 2008.

L. H. Y. Chen, L. Goldstein, and Q. M. Shao. *Normal approximation by Stein's method*. Springer, 2010.

S. Chen and J.-P. Onnela. A bootstrap method for goodness of fit and model selection with a single observed network. *Scientific Reports*, 9(1):1–12, 2019.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *JMLR: Workshop and Conference Proceedings*, 2016.

K. P. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*, pages 3608–3616, 2014.

R. Eldan and R. Gross. Exponential random graphs behave like mixtures of stochastic block models. *The Annals of Applied Probability*, 28(6):3698–3735, 2018.

T. Fernandez, N. Rivera, W. Xu, and A. Gretton. Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pages 3112–3122. PMLR, 2020.

J. H. Fowler. Connecting the Congress: A study of cosponsorship networks. *Political Analysis*, 14(4): 456–487, 2006a.

J. H. Fowler. Legislative cosponsorship networks in the US House and Senate. *Social Networks*, 28(4): 454–465, 2006b.

O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

T. Gärtner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer, 2003.

J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

J. Gorham, A. Raj, and L. Mackey. Stochastic Stein discrepancies. *arXiv preprint arXiv:2007.02857*, 2020.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, pages 1205–1213, 2012.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76 (373):33–50, 1981.

D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

W. Jitkrittum, Z. Szabó, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In *International Conference on Machine Learning*, pages 1742–1751. PMLR, 2017a.

W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271, 2017b.

W. Jitkrittum, H. Kanagawa, P. Sangkloy, J. Hays, B. Schölkopf, and A. Gretton. Informative features for model comparison. In *Advances in Neural Information Processing Systems*, pages 808–819, 2018.

W. Jitkrittum, H. Kanagawa, and B. Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR, 2020.

H. Kanagawa, W. Jitkrittum, L. Mackey, K. Fukumizu, and A. Gretton. A kernel Stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*, 2019.

N. M. Kriege, P.-L. Giscard, and R. Wilson. On valid optimal assignment kernels and applications to graph classification. In *Advances in Neural Information Processing Systems*, pages 1623–1631, 2016.

N. M. Kriege, F. D. Johansson, and C. Morris. A survey on graph kernels. *Applied Network Science*, 5(1):1–42, 2020.

E. Lazega. *The collegial phenomenon: The social mechanisms of cooperation among peers in a corporate law partnership*. Oxford University Press on Demand, 2001.

C. Ley, G. Reinert, and Y. Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284, 2016.

J. Lospinoso and T. A. Snijders. Goodness of fit for stochastic actor-oriented models. *Methodological Innovations*, 12(3):2059799119884282, 2019.

D. Lusher, J. Koskinen, and G. Robins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, 2013.

E. Meckes. On Stein's method for multivariate normal approximation. In *High dimensional probability V: the Luminy volume*, pages 153–178. Institute of Mathematical Statistics, 2009.

M. Morris, S. Goodreau, C. Butts, M. Handcock, and D. Hunter. ergm: A package to fit, simulate and diagnose exponential-family models for networks. 2008.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

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*, 7(2):226–253, 2019.

S. Ouadah, S. Robin, and P. Latouche. Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases. *Scandinavian Journal of Statistics*, 47(1):156–181, 2020.

G. Reinert and A. Röllin. Multivariate normal approximation with stein's method of exchangeable pairs under a general linearity condition. *The Annals of Probability*, 37(6):2150–2173, 2009.

G. Reinert and A. Röllin. Random subgraph counts and $u$-statistics: multivariate normal approximation via exchangeable pairs and embedding. *Journal of Applied Probability*, 47(2):378–393, 2010.

G. Reinert and N. Ross. Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *The Annals of Applied Probability*, 29(5):3201–3229, 2019.

D. A. Rolls, P. Wang, E. McBryde, P. Pattison, and G. Robins. A simulation study comparing epidemic dynamics on exponential random graph and edge-triangle configuration type contact network models. *PloS one*, 10(11):e0142181, 2015.

C. S. Schmid and B. A. Desmarais. Exponential random graph models with big networks: Maximum pseudolikelihood estimation and the parametric bootstrap. In *2017 IEEE International Conference on Big Data*, pages 116–121. IEEE, 2017.

M. Schweinberger. Statistical modelling of network panel data: Goodness of fit. *British Journal of Mathematical and Statistical Psychology*, 65(2):263–281, 2012.

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

J. Shore and B. Lubin. Spectral goodness of fit for network models. *Social Networks*, 43:16–27, 2015.

T. A. Snijders. Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.

T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1): 99–153, 2006.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

C. Steglich, T. A. Snijders, and P. West. Applying siena. *Methodology*, 2(1):48–56, 2006.

C. Stein. Approximate Cmputation of Expectations. IMS, 1986.

M. Sugiyama and K. Borgwardt. Halting in random walk kernels. In *Advances in Neural Information Processing Systems*, pages 1639–1647, 2015.

M. Sugiyama, M. E. Ghisu, F. Llinares-López, and K. Borgwardt. graphkernels: R and Python packages for graph comparison. *Bioinformatics*, 34(3): 530–532, 2018.

S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242, 2010.

S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, 1994.

W. Xu and T. Matsuda. A Stein goodness-of-fit test for directional distributions. *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning*, pages 5557–5566, 2018.

J. Yang, V. Rao, and J. Neville. A Stein–Papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 226–235, 2019.

F. Yin, N. E. Phillips, and C. T. Butts. Selection of exponential-family random graph models via Held-Out Predictive Evaluation (HOPE). *arXiv preprint arXiv:1908.05873*, 2019.

Y. Zhang, A. J. Friend, A. L. Traud, M. A. Porter, J. H. Fowler, and P. J. Mucha. Community structure in congressional cosponsorship networks. *Physica A: Statistical Mechanics and its Applications*, 387(7): 1705–1712, 2008.