A Stein Goodness-of-test for Exponential Random Graph Models Supplementary Material

A Proofs and Additional Lemmas

In Gorham et al. [2020] asymptotic consistency for stochastic Stein discrepancies is shown for Stein operators which yield continuous functions. Due to this continuity assumption the techniques applied in Gorham et al. [2020] differ considerably to then ones we shall employ here. Moreover, while the results in Gorham et al. [2020] are of asymptotic nature, the results in this section give computable bounds.

Proof of Theorem 1

For convenience we re-state the theorem here.

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 $\epsilon > 0$,*

$$\mathbb{P}(|\operatorname{gKSS}(q, X) - \operatorname{gKSS}(\tilde{q}, Y)| > \epsilon) \leq \left\{ ||\Delta(\operatorname{gKSS}(q,\cdot))^2||(1 + ||\Delta \operatorname{gKSS}(q,\cdot)||) + 4 \sup_x (||\Delta f_x^*||^2) \right\} \binom{n}{2} \frac{C}{\epsilon^2 \sqrt{n}}.$$

Under the null hypothesis, $X \sim q$ which is an ERGM satisfying Assumption 1. Let $Y \sim \tilde{q}$, where $\tilde{q}$ is the Bernoulli random graph with edge probability a^* and a^* is a solution to the equation in Assumption 1. We use the triangle inequality,

$$|\operatorname{gKSS}(q, x) - \operatorname{gKSS}(\tilde{q}, y)| \leq |\operatorname{gKSS}(q, x) - \operatorname{gKSS}(\tilde{q}, x)| + |\operatorname{gKSS}(\tilde{q}, y)|, x) - \operatorname{gKSS}(\tilde{q}, y)|. \tag{17}$$

This gives rise to two approximation terms. For the first summand in (17), we start with noting that

$$\operatorname{gKSS}(q, x) = \sup_{f \in \mathcal{H}, ||f|| \leq 1} |\mathcal{T}_q f(x)| = \sup_{f \in \mathcal{H}, ||f|| \leq 1} |(\mathcal{T}_q - \mathcal{T}_{\tilde{q}} + \mathcal{T}_{\tilde{q}}) f(x)| \leq \sup_{f \in \mathcal{H}, ||f|| \leq 1} |(\mathcal{T}_q - \mathcal{T}_{\tilde{q}}) f(x)| + \operatorname{gKSS}(\tilde{q}, x)$$

and this inequality also holds with the roles of q and $\tilde{q}$ reversed, so that

$$|\operatorname{gKSS}(q, x) - \operatorname{gKSS}(\tilde{q}, x)| \leq \sup_{f \in \mathcal{H}, ||f|| \leq 1} |(\mathcal{T}_q - \mathcal{T}_{\tilde{q}}) f(x)| \quad = \sup_{f \in \mathcal{H}, ||f|| \leq 1} |\langle f(\cdot), (\mathcal{T}_q - \mathcal{T}_{\tilde{q}}) k(x,\cdot) \rangle_{\mathcal{H}}|$$

where we used that due to the RKHS property, $f(x) = \langle f(\cdot), k(x,\cdot)\rangle_{\mathcal{H}}$. Thus we have an explicit form for the optimal f_x^* in this expression, namely $f_x^*(\cdot) = (\mathcal{T}_q - \mathcal{T}_{\tilde{q}})k(x,\cdot)/\|(\mathcal{T}_q - \mathcal{T}_{\tilde{q}})k(x,\cdot)\|_{\mathcal{H}}$, and

$$|\operatorname{gKSS}(q, x) - \operatorname{gKSS}(\tilde{q}, x)| \leq |(\mathcal{T}_q - \mathcal{T}_{\tilde{q}}) f_x^*(x)|.$$

Following the steps for the proof of Theorem 1.7 in Reinert and Ross [2019] but working directly with a function f without using that it is a solution of a Stein equation, it is straightforward to show that for all $f \in \mathcal{H}$, it holds that for $Y \sim \tilde{q}$,

$$|\mathbb{E}(\mathcal{T}_q f(Y) - \mathcal{T}_{\tilde{q}} f(Y))| \leq ||\Delta f|| \binom{n}{2} \frac{C(\beta, t)}{\sqrt{n}}$$

for an explicit constant C which depends only on the vectors β and t. Moreover inspecting the proof of Lemma 2.4 in Reinert and Ross [2019] the bound is indeed a stronger bound,

$$\frac{1}{N} \sum_{s \in N} \mathbb{E}|(\mathcal{T}_q^{(s)} f(Y) - \mathcal{T}_{\tilde{q}}^{(s)} f(Y))| \leq ||\Delta f|| \binom{n}{2} \frac{C(\beta, t)}{\sqrt{n}}.$$

In particular with the crude bound $|(\mathcal{T}_q^{(s)} - \mathcal{T}_{\tilde{q}}^{(s)}) f| \leq 2||\Delta f||$ it follows that

$$\mathbb{E}\left\{ \left(\frac{1}{N} \sum_{s \in N} (\mathcal{T}_q^{(s)} f(Y) - \mathcal{T}_{\tilde{q}}^{(s)} f(Y)) \right)^2 \right\} \leq 2||\Delta f||^2 \binom{n}{2} \frac{C(\beta, t)}{\sqrt{n}}.$$

Thus, using the Chebychev inequality, for all $\epsilon > 0$,

$$\mathbb{P}(|(\mathcal{T}_q - \mathcal{T}_{\tilde{q}})f_Y^*(Y)| > \epsilon) \leq \frac{1}{\epsilon^2}\mathbb{V}ar((\mathcal{T}_q - \mathcal{T}_{\tilde{q}})f_Y^*(Y)) \leq 4\sup_x(||\Delta f_x^*||^2)\binom{n}{2}\frac{C(\beta,t)}{\epsilon^2\sqrt{n}}.$$

Hence

$$\mathbb{P}(|\operatorname{gKSS}(q,Y) - \operatorname{gKSS}(\tilde{q},Y)| > \epsilon) \leq 4\sup_x(||\Delta f_x^*||^2)\binom{n}{2}\frac{C(\beta,t)}{\epsilon^2\sqrt{n}}.$$

For the second summand in Eq.(17), to bound $|\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y)|$ we consider the test function $h(x) = \operatorname{gKSS}(q,x)$ and apply Theorem 1.7 from Reinert and Ross [2019] to give that

$$|\mathbb{E}(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))| \leq ||\Delta \operatorname{gKSS}(q,\cdot)||\binom{n}{2}\frac{\tilde{C}}{\sqrt{n}}.$$

Here $\tilde{C}$ is a new constant which depends only on β and t. Similarly we can approximate the square of the expectation using that $(a-b)^2 = a^2 - b^2 + 2b(b-a)$ and write

$$\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))^2\} = \mathbb{E}\{(\operatorname{gKSS}(q,X)^2\} - \mathbb{E}\{\operatorname{gKSS}(q,Y)^2\} + 2\mathbb{E}\{\operatorname{gKSS}(q,X)(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y)\}.$$

The first summand can be bounded with Theorem 1.7 from Reinert and Ross [2019] using the test function $h(x) = \operatorname{gKSS}(q,x)^2$. For the second summand, we the Cauchy-Schwarz inequality gives

$$|\mathbb{E}\{\operatorname{gKSS}(q,X)(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y)\}| \leq [\mathbb{E}(\operatorname{gKSS}(q,X)^2)]^{\frac{1}{2}}[\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))^2\}]^{\frac{1}{2}}$$

As $|\operatorname{gKSS}(q,x)| \leq 1$ we obtain

$$\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))^2\} \leq \mathbb{E}\{\operatorname{gKSS}(q,X)^2\} - \mathbb{E}\{\operatorname{gKSS}(q,Y)^2\} + 2[\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))^2\}]^{\frac{1}{2}}.$$

Solving this quadratic inequality gives

$$\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))^2\} \leq \left(1 - \sqrt{1 - (\mathbb{E}\{\operatorname{gKSS}(q,X)^2\} - \mathbb{E}\{\operatorname{gKSS}(q,Y)^2\})}\right)^2$$

and $|\mathbb{E}\{\operatorname{gKSS}(q,X)^2\} - \mathbb{E}\{\operatorname{gKSS}(q,Y)^2| \leq 1$ we obtain that

$$\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))^2\} \leq |\mathbb{E}\{\operatorname{gKSS}(q,X)^2\} - \mathbb{E}\{\operatorname{gKSS}(q,Y)^2\}|.$$

With Theorem 1.7 from Reinert and Ross [2019] for the test function $h(x) = \operatorname{gKSS}(q,x)^2$ we obtain

$$\mathbb{E}\{(\operatorname{gKSS}(q,X) - \operatorname{gKSS}(q,Y))\}^2 \leq (||\Delta(\operatorname{gKSS}(q,\cdot))^2||\binom{n}{2}\frac{\hat{C}}{\sqrt{n}},$$

where $\hat{C}$ is another constant which depends only on β and t but not on n. With the Chebychev inequality and the triangle inequality we conclude that there is an explicitly computable constant C such that for all x

$$\mathbb{P}(|\operatorname{gKSS}(q,X) - \operatorname{gKSS}(\tilde{q},Y)| > \epsilon) \leq (\sup_x(||\Delta f_x^*||^2) + ||\Delta(\operatorname{gKSS}(q,\cdot))^2||(1 + ||\Delta \operatorname{gKSS}(q,\cdot)||)\binom{n}{2}\frac{C}{\epsilon^2\sqrt{n}}.$$

The assertion follows. □

For the approximate distribution of $\operatorname{gKSS}(\tilde{q},Y)$ it is more convenient to consider the square as given in Eq.(11); this is addressed by Theorem 2.

Proof of Theorem 2

For convenience we re-state the assumptions and the theorem here.

To approximate the distribution of gKSS^2 under the null hypothesis we make the following assumptions (Assumption 2 in the main text) on the kernel K for the RKHS $\mathcal{H}$, namely that for $x, y \in \{0,1\}^N$,

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 $s \in [N]$.

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*

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

For the Bernoulli random graph distribution $\tilde{q}$, and $s \in [N]$,

$$\mathcal{T}_{\tilde{q}}^{(s)} f(x) = a^* f(x^{(s,1)} - f(x)) + (1-a^*) f(x^{(s,0)} - f(x)).$$

Thus,

$$\begin{aligned}
\mathrm{gKSS}^2(\tilde{q}, x) &= \frac{1}{N^2} \sum_{s,s'\in[N]} \left\langle \mathcal{T}_{\tilde{q}}^{(s)} K(x,\cdot), \mathcal{T}_{\tilde{q}}^{(s')} K(x,\cdot) \right\rangle \\
&= \frac{1}{N^2} \sum_{s,s'\in[N]} \Big\langle a^* \left(K(x^{(s,1)},\cdot) - K(x,\cdot) \right) + (1-a^*) \left(K(x^{(s,0)},\cdot) - K(x,\cdot) \right), \\
&\qquad\qquad a^* \left(K(x^{(s',1)},\cdot) - K(x,\cdot) \right) + (1-a^*) \left(K(x^{(s',0)},\cdot) - K(x,\cdot) \right) \Big\rangle.
\end{aligned}$$

Under Assumptions (i), (ii) and (iii) we can write

$$\begin{aligned}
K(x^{(s,1)},\cdot) - K(x,\cdot) &= (l_s(1,\cdot) - l_s(x_s,\cdot)) \prod_{t\neq s} l_t(x_t,,\cdot) \\
&= (1-x_s)(l_s(1,\cdot) - l_s(0,\cdot)) l_{s'}(x_{s'},\cdot) \prod_{t\neq s,s'} l_t(x_t,\cdot).
\end{aligned}$$

Similarly,

$$K(x^{(s,0)},\cdot) - K(x,\cdot) = -x_s(l_s(1,\cdot) - l_s(0,\cdot)) l_{s'}(x_{s'},\cdot) \prod_{t\neq s,s'} l_t(x_t,\cdot).$$

Abbreviating $g(x^{-s,s'},\cdot) := \prod_{t\neq s,s'} l_t(x_t,\cdot)$ we obtain that

$$\begin{aligned}
\mathrm{gKSS}^2(\tilde{q}, x) &= \frac{1}{N^2} \sum_{s,s'\in[N]} (a^*(1-x_s) - (1-a^*)x_s)(a^*(1-x'_s) - (1-a^*)x'_s) \\
&\quad \langle (l_s(1,\cdot) - l_s(0,\cdot) l_{s'}(x_{s'},\cdot), \langle (l'_s(1,\cdot) - l'_s(0,\cdot)) l_s(x_s,\cdot)\rangle \langle g(x^{-s,s'},\cdot), g(x^{-s,s'},\cdot)\rangle \\
&= \frac{1}{N^2} \sum_{s,s'\in[N]} (a^* - x_s)(a^* - x'_s) \langle (l_s(1,\cdot) - l_s(0,\cdot)) l_{s'}(x_{s'},\cdot), \langle (l'_s(1,\cdot) - l'_s(0,\cdot)) l_s(x_s,\cdot)\rangle \\
&= \frac{1}{N^2} \sum_{s,s'\in[N]} (a^* - x_s)(a^* - x'_s) \langle l_s(x_s,\cdot), l_{s'}(x_{s'},\cdot)\rangle c(s,s')
\end{aligned}$$

with

$$c(s,s') = \langle l_s(1,\cdot) - l_s(0,\cdot), l_{s'}(1,\cdot) - l_{s'}(0,\cdot)\rangle$$

not depending on x. Here we used that by assumptions (ii) and (iii), $\langle g(x^{-s,s'},\cdot), g(x^{-s,s'},\cdot)\rangle = 1$. Thus, when replacing x by Y, a random vector in $\{0,1\}^N$ representing a Bernoulli random graph on n vertices with edge probability p, then $\mathrm{gKSS}^2(\tilde{q}, Y)$ is an average of locally dependent random variables. Hence, using Stein's method

we obtain a normal approximation with bound, as follows. Let $\mathcal{I} = \{(s, s') : s, s' \in [N]\}$ so that $|\mathcal{I}| = N^2$. For $\alpha = (s, s') \in \mathcal{I}$ set

$$X_\alpha = \frac{1}{N^2}(a^* - Y_s)(a^* - Y_{s'})\langle l_s(Y_s, \cdot), l_{s'}(Y_{s'}, \cdot)\rangle c(s, s');$$

then

$$\mathrm{gKSS}^2(\tilde{q}, Y) = \sum_{\alpha \in \mathcal{I}} X_\alpha$$

and unless α and β share at least one vertex, the random variables X_α and X_β are independent. Let $\mu_\alpha = \mathbb{E}X_\alpha$ and $\sigma^2 = \mathbb{V}ar(\mathrm{gKSS}^2(\tilde{q}, Y)))$; these quantities depend on the chosen kernels l_s. We use the standardised count

$$W = \sum_{\alpha \in \mathcal{I}} \frac{X_\alpha - \mu_\alpha}{\sigma} = \frac{1}{\sigma}\,\mathrm{gKSS}^2(\tilde{q}, Y) - \sum_{\alpha \in \mathcal{I}} \frac{\mu_\alpha}{\sigma};$$

then W has mean zero, variance 1, and results from Section 4.7 in Chen et al. [2010] apply. In their notation, with $A_{(s,s')} = \{\beta = (t, t') \in \mathcal{I} : |\{s, s'\} \cap \{t, t'\}| \neq \emptyset$, condition (LD1) is satisfied. Applying Theorem 4.13, p.134, from Chen et al. [2010] yields that, with $||\cdot||_1$ denoting L_1-distance, $\mathcal{L}$ denoting the law of a random variable, and Z denoting a standard normal variable,

$$||\mathcal{L}(W) - \mathcal{L}(Z)||_1 \leq \sqrt{\frac{2}{\pi}}\mathbb{E}\left|\sum_{\alpha \in \mathcal{I}}(\xi_\alpha \eta_\alpha - \mathbb{E}(\xi_\alpha \eta_\alpha))\right| + \sum_{\alpha \in \mathcal{I}} \mathbb{E}|\xi_\alpha \eta_\alpha^2| \leq \sqrt{\frac{2}{\pi}}\sqrt{\mathbb{V}ar(\sum_{\alpha \in \mathcal{I}} \xi_\alpha \eta_\alpha)} + \sum_{\alpha \in \mathcal{I}} \mathbb{E}|\xi_\alpha \eta_\alpha^2|. \quad (18)$$

with $\xi_\alpha = (X_\alpha - \mu_\alpha)/\sigma$ and $\eta_\alpha = \sum_{\beta \in A_\alpha} X_\beta$.

To obtain the dependence of the bound on N we assess its magnitude. First note that $|A_\alpha| \leq 2N$. Using that by the assumption (iii), $||l_s||^2 = 1$ for $s \in [N]$ and that $|a^* - Y_s| \leq 1$ we can use the crude bounds $|c(s, s')| \leq 4$, so that $|X_\alpha| \leq \frac{4}{N^2}$ and $\mu_\alpha \leq \frac{4}{N^2}$, In particular, $|\xi_\alpha| \leq \frac{8}{N^2\sigma}$ and $|\eta_\alpha| \leq \frac{16}{N\sigma}$. Thus,

$$\sum_{\alpha \in \mathcal{I}} \mathbb{E}|\xi_\alpha \eta_\alpha^2| \leq N^2 \times \frac{8}{N^2\sigma} \times \frac{256}{N^2\sigma^2} = \frac{2048}{N^2\sigma^3}.$$

To evaluate the variance σ^2,

$$\sigma^2 = \sum_{\alpha \in \mathcal{I}} \mathbb{V}ar X_\alpha + \sum_{\alpha \in \mathcal{I}} \sum_{\beta \in A_\alpha} Cov(X_\alpha, X_\beta).$$

We evaluate these terms in turn. First, if $\alpha = (s, s)$ then

$$\mathbb{V}ar X_\alpha \leq \frac{c(s, s)^2}{N^4} a^*(1 - a^*)$$

and if $\alpha = (s, s')$ with $s \neq s'$ then as $\langle l_s(x, \cdot), l_s(y, \cdot)\rangle \leq 1$ from the assumption (iii) and the Cauchy-Schwarz inequality,

$$\mathbb{V}ar X_\alpha \leq \mathbb{E}[X_\alpha^2] \leq \frac{c(s, s)^2}{N^4}[a^*(1 - a^*)]^2.$$

Thus,

$$\sum_{\alpha \in \mathcal{I}} \mathbb{V}ar X_\alpha \leq \frac{c(s, s)^2}{N^2} a^*(1 - a^*).$$

Moreover, if $\alpha = (s, s)$ and $\beta = (s, t) \in A_\alpha$ then

$$|Cov(X_\alpha, X_\beta)| = \left|\frac{c(s, s)c(s, t)}{N^4}\mathbb{E}\{(a^* - Y_s)^3(a^* - Y_t)\langle l_s(Y_s, \cdot), l_t(Y_t, \cdot)\rangle\} - \mu_\alpha\mu_\beta\right| \leq 2\frac{|c(s, s)c(s, t)|}{N^4}$$

and there are order N^2 such terms (α, β) in the variance. The main contributions to the variance stem from $Cov(X_\alpha, X_\beta)$ for $\beta \in \mathcal{I}_\alpha$ and for $\alpha = (s.s')$ with $s \neq s'$. Assumption (iv) guarantees that $c(s, s') \neq 0$. Then for $\beta = (s, t)$, with $t \neq s$,

$$\begin{aligned} Cov(X_\alpha, X_\beta) \quad = \quad & \frac{1}{N^4}c(\alpha)c(\beta)\mathbb{E}(a^* - Y_s)^2(a^* - Y_s')(a^* - Y_t)\langle l_s(Y_s, \cdot), l_{s'}(Y_{s'}, \cdot)\rangle\langle l_s(Y_s, \cdot), l_t(Y_t, \cdot)\rangle \\ & -\frac{1}{N^4}(a^*)^4(1 - a^*)^4 c(s, s')c(s, t) \end{aligned}$$

and expanding the expectation gives a contribution of the order N^{-4}. The overall contribution of such covariance terms, of which there are order N^3, to the variance is hence of order N^{-1}, and therefore σ^2 is of order N^{-1} and σ is of order $\sqrt{N}$.

Similarly,

$$\mathbb{V}ar\left(\sum_{\alpha\in\mathcal{I}}\xi_\alpha\eta_\alpha\right)=\mathbb{V}ar\left(\sum_{\alpha\in\mathcal{I}}\sum_{\beta\in A_\alpha}\xi_\alpha\xi_\beta\right)=\sum_{\alpha\in\mathcal{I}}\sum_{\beta\in A_\alpha}\sum_{\gamma\in\mathcal{I}}\sum_{\delta\in A_\gamma}Cov(\xi_\alpha\xi_\beta,\xi_\gamma\xi_\delta)$$

is dominated by the covariances between $\xi_\alpha\xi_\beta$ and $\xi_\gamma\xi_\delta$ such that α and β involve three distinct indices s, s', t, and γ and δ involve three distinct indices r, r', u, and these two sets of three indices have non-zero intersection. These summands give a contribution of order $N^5/(\sigma^4N^8)$, which is of order N^{-1}, to the variance $\mathbb{V}ar\left(\sum_{\alpha\in\mathcal{I}}\xi_\alpha\eta_\alpha\right)$. A crude bound is obtained as $\mathbb{V}ar(\sum_{\alpha\in\mathcal{I}}\xi_\alpha\eta_\alpha)\leq\frac{512}{\sigma^4N^3}$. These estimates give that the bound in Eq.(18) is of the order $N^{-\frac{1}{2}}$. All moment expressions can be bounded explicitly and thus the constant C can be computed explicitly. The conclusion follows.

Proof of Proposition 2

For convenience we re-state the result here again.

Proposition 2. *Let*

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

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,

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

For normal approximation in the presence of weak dependence, Charles Stein [Stein, 1986] introduced the method of exchangeable pairs: construct a sum W' such that (W,W') form an exchangeable pair, and such that $\mathbb{E}^W(W'-W)$ is (at least approximately) linear in W. This linearity condition arises naturally when thinking of correlated bivariate normals. As a multivariate generalisation, Reinert and Röllin [2009] considered the general setting that

$$\mathbb{E}^W(W'-W)=-\Lambda W+R \tag{19}$$

for a matrix Λ and a vector R with small $\mathbb{E}|R|$ is treated. In a followup paper [Meckes, 2009] the results by Chatterjee and Meckes [2008] and Reinert and Röllin [2009] are combined using slightly different smoothness conditions on test functions as compared to Reinert and Röllin [2009]. In Reinert and Röllin [2009] it was found that a statistic of interest can often be embedded into a larger vector of statistics such that (19) holds with $R=0$; this embedding does not directly correspond to Hoeffding projections, although it is related to the latter. In Reinert and Röllin [2010] this embedding is applied to complete non-degenerate U-statistics. among other examples. In this example the limiting covariance matrix is not of full rank; yet the bounds on the normal approximation are of the expected order.

The general setup is as follows. Denote by $W=(W_1,W_2,\ldots,W_d)^t$ random vectors in $\mathbb{R}^d$, where W_i are $\mathbb{R}$-values random variables for $i=1,\ldots,d$. We denote by Σ symmetric, non-negative definite matrices, and hence by $\Sigma^{1/2}$ the unique symmetric square root of Σ. Denote by Id the identity matrix, where we omit the dimension d. Let Z denote a random variable having standard d-dimensional multivariate normal distribution. We abbreviate the transpose of the inverse of a matrix Λ as $\Lambda^{-t}:=(\Lambda^{-1})^t$.

For derivatives of smooth functions $h:\mathbb{R}^d\to\mathbb{R}$, we use the notation ∇ for the gradient operator. Denote by $\|\cdot\|$ the supremum norm for both functions and matrices. If the corresponding derivatives exist for some function $g:\mathbb{R}^d\to\mathbb{R}$, we abbreviate $|g|_1:=\sup_i\left\|\frac{\partial}{\partial x_i}g\right\|$, $|g|_2:=\sup_{i,j}\left\|\frac{\partial^2}{\partial x_i\partial x_j}g\right\|$, and so on.

The following result is shown in Reinert and Röllin [2009].

Theorem A.1 (c.f. Theorem 2.1 Reinert and Röllin [2009]). *Assume that (W, W') is an exchangeable pair of $\mathbb{R}^d$-valued random variables such that*

$$\mathbb{E}W = 0, \qquad \mathbb{E}WW^t = \Sigma, \tag{20}$$

with $\Sigma \in \mathbb{R}^{d\times d}$ symmetric and positive definite. Suppose further that (19) is satisfied for an invertible matrix Λ and a $\sigma(W)$-measurable random variable R. Then, if Z has d-dimensional standard normal distribution, we have for every three times differentiable function g

$$\left|\mathbb{E}g(W) - \mathbb{E}g(\Sigma^{1/2}Z)\right| \leq \frac{|g|_2}{4}I + \frac{|g|_3}{12}II + \Big(|g|_1 + \tfrac{1}{2}d\|\Sigma\|^{1/2}|g|_2\Big)III,$$

where, with $\lambda^{(i)} = \sum_{m=1}^d |(\Lambda^{-1})_{m,i}|$,

$$\begin{aligned} I &= \sum_{i,j=1}^d \lambda^{(i)}\sqrt{\mathbb{V}ar\mathbb{E}^W(W_i' - W_i)(W_j' - W_j)}, \\ II &= \sum_{i,j,k=1}^d \lambda^{(i)}\mathbb{E}\left|(W_i' - W_i)(W_j' - W_j)(W_k' - W_k)\right|, \\ III &= \sum_i \lambda^{(i)}\sqrt{\mathbb{E}R_i^2}. \end{aligned}$$

Here we use the approach for statistics of the form $Y = \frac{1}{B^2}\sum_{s,t\in[N]}(k_sk_t - \mathbb{E}(k_sk_t))h(s,t)$. The subscript x is suppressed in h_x to simplify notation. To apply Theorem A.1 we employ two additional statistics; including Y as W_1,

$$\begin{aligned} W_1 &= \frac{1}{B^2}\sum_{s,t\in[N]}(k_sk_t - \mathbb{E}(k_sk_t))h(s,t) \\ W_2 &= \frac{1}{B^2}\sum_{s,t\in[N]}(k_s - \mathbb{E}(k_s))h(s,t) \\ W_3 &= \frac{1}{B^2}\sum_{s\in[N]}(k_s - \mathbb{E}(k_s))h(s,s). \end{aligned}$$

Given $\mathbf{k} = (k_1, \ldots, k_N)$ we construct an exchangeable pair $(\mathbf{k}, \mathbf{k}')$ by choosing an index $I \in [N]$ such that $\mathbb{P}(I = i) = \frac{k_i}{B}$ and if $I = i$ we set $k_i' = k_i - 1$ (we take a ball out of bin i in the multinomial construction). Then we pick $J \in [N]$ uniformly and if $J = j$ we set $k_j' = k_j + 1$ - we add the ball to bin j which we took away from bin i. All other $k_l's$ are left unchanged; $k_l' - k_l$ if $l \neq I, J$. Note that $I = J$ is possible in which case there is no change. Based on this exchangeable pair we set

$$\begin{aligned} W_1' &= \frac{1}{B^2}\sum_{s,t\in[N]}(k_s'k_t\text{`} - \mathbb{E}(k_s'k_t'))h(s,t) \\ W_2' &= \frac{1}{B^2}\sum_{s,t\in[N]}(k_s' - \mathbb{E}(k_s'))h(s,t) \\ W_3' &= \frac{1}{B^2}\sum_{s\in[N]}(k_s' - \mathbb{E}(k_s'))h(s,s). \end{aligned}$$

With $W = (W_1, W_2, W_3)$ and $W' = (W_1', W_2', W_3')$ we have obtained an exchangeable pair (W, W'). Moreover W has mean zero and finite covariance matrix. First we calculate $\mathbb{E}^W(W' - W)$ componentwise, starting with the easiest case to illustrate the argument. For this calculation we use that

$$\begin{aligned} k_I' - k_I &= -1 \\ k_J' - k_J &= 1 \\ k_s'k_t\text{`} - k_sk_t &= (k_s' - k_s)(k_t' - k_t) + k_s(k_t' - k_t) + k_t(k_s' - k_s). \end{aligned}$$

Then, conditioning on I and J,

$$\begin{aligned}
\mathbb{E}^W(W_3' - W_3) &= \frac{1}{B^2}\sum_{s\in[N]}\mathbb{E}^W(k_s' - k_s)h(s,s) \\
&= \frac{1}{B^2}\frac{1}{BN}\sum_{s\in[N]}\sum_{i\in[N]}k_i\sum_{j\in[N]}(-\mathbf{1}(s=i)h(i,i) + \mathbf{1}(s=j)h(j,j)) \\
&= -\frac{1}{B^2}\frac{1}{B}\sum_{i\in[N]}k_i h(i,i) + \frac{1}{N}\frac{1}{B^2}\sum_{j\in[N]}h(j,j) \\
&= -\frac{1}{B^2}\frac{1}{B}\sum_{i\in[N]}(k_i - \mathbb{E}(k_i))h(i,i) \\
&= -\frac{1}{B}W_3.
\end{aligned}$$

Similar arguments yield $\mathbb{E}^W(W_2' - W_2) = -\frac{1}{B}W_2$. Finally,

$$\begin{aligned}
&\mathbb{E}^W(W_1' - W_1) \\
&= \frac{1}{B^2}\sum_{s,t\in[N]}\mathbb{E}^W(k_s'k_t` - k_sk_t)h(s,t) \\
&= \frac{1}{B^2}\sum_{s,t\in[N]}\mathbb{E}^W[(k_s' - k_s)(k_t' - k_t) + k_s(k_t' - k_t) + k_t(k_s' - k_s)]h(s,t) \\
&= \frac{1}{B^2}\sum_{s,t\in[N]}\mathbb{E}^W[(k_s' - k_s)(k_t' - k_t)]h(s,t) + 2\sum_{s,t\in[N]}\mathbb{E}^W[k_t(k_s' - k_s)]h(s,t).
\end{aligned}$$

Here we used that $h(s,t) = h(t,s)$ in the last step. We tackle the conditional expectations separately. Again using $h(s,t) = h(t,s)$,

$$\begin{aligned}
&\sum_{s,t\in[N]}\mathbb{E}^W[(k_s' - k_s)(k_t' - k_t)]h(s,t) \\
&= \frac{1}{BN}\sum_{s,t\in[N]}\sum_{i\in[N]}k_i\sum_{j\in[N]}(\mathbf{1}(s=I,t=J) + \mathbf{1}(s=J,t=I)[(k_s' - k_s)(k_t' - k_t)]h(s,t) \\
&= -\frac{2}{BN}\sum_{i\in[N]}k_i\sum_{j\in[N]}\mathbf{1}(i\neq j)h(i,j) \\
&= -\frac{2}{BN}\sum_{i\in[N]}\sum_{j\in[N]}k_i h(i,j) + \frac{2}{BN}\sum_{i\in[N]}k_i h(i,i) \\
&= -\frac{2}{BN}W_2 + \frac{2}{BN}W_3.
\end{aligned}$$

Here the centering terms from W_2 and W_3 add up to 0 because the conditional expectation has mean zero, and are thus not included in the calculation.

Moreover,

$$\begin{aligned}
&\sum_{s,t\in[N]}\mathbb{E}^W[k_t(k_s' - k_s)]h(s,t) \\
&= \frac{1}{BN}\sum_{s,t\in[N]}\sum_{i\in[N]}k_i\sum_{j\in[N]}(-\mathbf{1}(s=i)k_t h(i,t) + \mathbf{1}(s=j)k_t h(j,t)) \\
&= -\frac{1}{B}\sum_{t\in[N]}\sum_{i\in[N]}k_i k_t h(i,t) + \frac{1}{N}\sum_{t\in[N]}\sum_{j\in[N]}k_t h(j,t) \\
&= -\frac{1}{B}W_1 + \frac{1}{N}W_2.
\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}^W(W_1' - W_1) &= -\frac{2}{BN}W_2 + \frac{2}{BN}W_3 - \frac{2}{B}W_1 + \frac{2}{N}W_2 \\ &= \frac{2}{BN}W_3 + \frac{2(B-1)}{BN}W_2 - \frac{2}{B}W_1.\end{aligned}$$

Hence (19) is satisfied with $R = 0$ and

$$\Lambda = \frac{1}{B}\begin{bmatrix} -2 & \frac{2(B-1)}{N} & \frac{2}{N} \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

giving

$$\lambda^{(1)} = \frac{B}{2}; \lambda^{(2)} = B\frac{|N - B + 1|}{N}; \lambda^{(3)} = \frac{B(N+1)}{N}.$$

With $B = FN$ we can bound

$$\lambda^{(i)} \leq \max(F, 1/2)B, \quad i = 1, 2, 3.$$

To complete the argument we need to bound I and II from Theorem A.1.

To bound the conditional variance term I from Theorem A.1,

$$I = \sum_{i,j=1}^{3} \lambda^{(i)} \sqrt{\mathbb{V}ar\mathbb{E}^W(W_i' - W_i)(W_j' - W_j)} \leq \max(F, 1/2)\, B \sum_{i,j=1}^{3} \sqrt{\mathbb{V}ar\mathbb{E}^W(W_i' - W_i)(W_j' - W_j)}.$$

Instead of conditioning on W we condition on $\mathbf{k}$ this conditioning would only increase the conditional variance. The largest variance contribution is from

$$\begin{aligned}&\mathbb{E}^{\mathbf{k}}(W_1' - W_1)^2 \\ &= \frac{1}{B^4}\sum_{s,t\in[N]}\sum_{u,v\in[N]} \mathbb{E}^{\mathbf{k}}[(k_s'k_t' - k_sk_t)(k_u'k_v' - k_uk_v)h(s,t)h(u,v)] \\ &= \frac{1}{B^4}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]} \mathbb{E}^{\mathbf{k}}[k_i\left((k_s' - k_s)(k_t' - k_t) + 2k_s(k_t' - k_t)\right) \times \\ &\qquad\qquad \left((k_u' - k_u)(k_v' - k_v) + 2k_u(k_v' - k_v)\right) h(s,t)h(u,v)] \\ &= \frac{1}{B^4}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]} \mathbb{E}^{\mathbf{k}}[k_i(k_s' - k_s)(k_t' - k_t)(k_u' - k_u)(k_v' - k_v)h(s,t)h(u,v)] \\ &\quad +2\frac{1}{B^4}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]} \mathbb{E}^{\mathbf{k}}[k_ik_u(k_s' - k_s)(k_t' - k_t)(k_v' - k_v)]h(s,t)h(u,v) \\ &\quad +2\frac{1}{B^4}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]} \mathbb{E}^{\mathbf{k}}[k_ik_s(k_t' - k_t)(k_u' - k_u)(k_v' - k_v)h(s,t)h(u,v)] \\ &\quad +4\frac{1}{B^4}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]} \mathbb{E}^{\mathbf{k}}[k_ik_sk_u(k_t' - k_t)(k_v' - k_v)h(s,t)h(u,v)].\end{aligned}$$

Due to the exchangeable pair construction many sums simplify and the largest contribution to the variance is

the last term;

$$
\begin{aligned}
&4\frac{1}{B^4}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]}\mathbb{E}^{\mathbf{k}}[k_ik_sk_u(k'_t-k_t)(k'_v-k_v)h(s,t)h(u,v)]\\
&= 4\frac{1}{B^5N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]}\mathbb{E}^{\mathbf{k}}[k_ik_sk_u(k'_t-k_t)(k'_v-k_v)h(s,t)h(u,v)](\mathbf{1}(t=i)+\mathbf{1}(t=j))\\
&= -4\frac{1}{B^5N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s\in[N]}\sum_{u,v\in[N]}\mathbb{E}^{\mathbf{k}}[k_ik_sk_u(k'_v-k_v)h(s,i)h(u,v)](\mathbf{1}(v=i)+\mathbf{1}(v=j))\\
&\quad +4\frac{1}{B^5N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s\in[N]}\sum_{u,v\in[N]}\mathbb{E}^{\mathbf{k}}[k_ik_sk_u(k'_v-k_v)h(s,j)h(u,v)](\mathbf{1}(v=i)+\mathbf{1}(v=j))\\
&= 4\frac{1}{B^5N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s\in[N]}k_ik_sk_uh(s,i)h(u,i)-4\frac{1}{B^5N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s\in[N]}k_ik_sk_uh(s,i)h(u,j)\\
&\quad -4\frac{1}{B^5N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s\in[N]}\sum_{u\in[N]}k_ik_sk_uh(s,j)h(u,i)+4\frac{1}{B^5N}B\sum_{j\in[N]}\sum_{s\in[N]}\sum_{u\in[N]}k_sk_uh(s,j)h(u,j).
\end{aligned}
$$

These terms have a variance contribution of order $\frac{1}{B^{10}N^2}\frac{B^6}{N^6}N^8=\frac{1}{B^4}$ as long as $h(i,j)$ is bounded. The mixed variances in I can be bounded using the Cauchy-Schwarz inequality. Overall the contribution to the term I of Theorem A.1 is thus of order $B\sqrt{\frac{1}{B^4}}=\frac{1}{B}$.

For the term II of Theorem A.1,

$$
\sum_{a,b,c=1}^{3}\lambda^{(a)}\mathbb{E}\big|(W'_a-W_a)(W'_b-W_b)(W'_c-W_c)\big| \;\leq\; \max(F,1/2)\,B\sum_{a,b,c=1}^{3}\mathbb{E}\big|(W'_a-W_a)(W'_b-W_b)(W'_c-W_c)\big|.
$$

The largest contribution to this term is

$$
\begin{aligned}
&\mathbb{E}\big|(W'_1-W_1)^3\big|\\
&\leq ||h||^3\frac{1}{B^6}\sum_{s,t\in[N]}\sum_{u,v\in[N]}\sum_{x,y\in[N]}\mathbb{E}|(k'_sk'_t-k_sk_t)(k'_uk'_v-k_uk_v)(k'_xk'_y-k_xk_y)|\\
&= ||h||^3\frac{1}{B^6}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]}\sum_{x,y\in[N]}\mathbb{E}\,|k_i\left((k'_s-k_s)(k'_t-k_t)+2k_s(k'_t-k_t)\right)\\
&\qquad ((k'_u-k_u)(k'_v-k_v)+2k_u(k'_v-k_v))\left((k'_x-k_x)(k'_y-k_y)+2k_x(k'_y-k_y)\right)\big|\\
&\leq ||h||^3\frac{1}{B^6}\frac{1}{BN}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]}\sum_{x,y\in[N]}\mathbb{E}[|k_i\left((k'_s-k_s)(k'_t-k_t)+2k_s(k'_t-k_t)\right)\\
&\qquad ((k'_u-k_u)(k'_v-k_v)+2k_u(k'_v-k_v))\left((k'_x-k_x)(k'_y-k_y)+2k_x(k'_y-k_y)\right)\big|.
\end{aligned}
$$

With $||h||=\max_{i,j}|h(i,j)|$ the leading term in this expression is

$$
\frac{8||h||^3}{B^7N}\sum_{i\in[N]}\sum_{j\in[N]}\sum_{s,t\in[N]}\sum_{u,v\in[N]}\sum_{x,y\in[N]}\mathbb{E}\left|k_ik_sk_x(k'_t-k_t)k_u(k'_v-k_v)(k'_y-k_y)\right|.
$$

Now, not all of t,v,y can be distinct for a non-zero contribution to this term; we can bound it by

$$
\frac{16||h||^3}{B^7N}\sum_{i,j,s,t,u,v,x\in[N]}\mathbb{E}k_ik_sk_uk_x(\mathbf{1}(t=i)+\mathbf{1}(t=j))(\mathbf{1}(v=i)+\mathbf{1}(v=j)) \;\leq\; \frac{64}{||h||^3B^3}.
$$

Here we used that $\sum_i k_i=B$. All other cross-expectations can be bounded using the Cauchy-Schwarz inequality. Hence we conclude that the term II in Theorem A.1 is of order B^{-2}. All higher moments can be bounded explicitly and hence C can be bounded explicitly. The conclusion follows.

B Graph Kernels

For a vertex-labeled graph $x = \{x_{ij}\}_{1\le i,j\le n} \in \mathcal{G}^{lab}$, with label range $\{1,\dots,c\} = [c]$, denote the vertex set by V, the edge set by E, and the label set by Σ. Consider an vertex-edge mapping $\psi : V \cup E \to [c]$. In this paper we use the following graph kernels.

Vertex-Edge Histogram Gaussian Kernels The *vertex-edge label histogram* $h = (h^{111}, h^{211}, \dots, h^{ccc}) = h(\psi, x)$ has as components $h^{l_1 l_2 l_3} = |\{v \in V, (v,u) \in E \,|\, \psi(v,u) = l_1, \psi(u) = l_2, \psi(v) = l_3\}|$, for $l_1, l_2, l_3 \in [c]$; it is a combination of vertex label counts and edge label counts. Let $\langle h(x), h(x')\rangle = \sum_{l_1,l_2,l_3} h(x)^{l_1,l_2,l_3} h(x')^{l_1,l_2,l_3}$. Following Sugiyama and Borgwardt [2015], we define the vertex-edge histogram Gaussian (VEG) kernel between two graphs x, x' as

$$K_{VEG}(x, x'; \sigma) = \exp\left\{-\frac{\|h(x) - h(x')\|^2}{2\sigma^2}\right\}.$$

The VEG kernel is a special case of histogram-based kernels for assessing graph similarity using feature maps, which are introduced in Kriege et al. [2016]. Adding a Gaussian RBF as in Sugiyama and Borgwardt [2015], yielding the VEG kernel, significantly improved problems such as classification accuracy, see [Kriege et al., 2020]. In our implementation, as in Sugiyama et al. [2018], ψ is induced by the vertex index. If the vertices are indexed by $i \in [n]$ then the label of vertex v_i is $\psi(v_i) = i$; for edges, $\psi(u,v) = 1$ if $(u,v) \in E$ is an edge and 0 otherwise.

Geometric Random Walk Graph Kernels A k-step random walk graph kernel [Sugiyama and Borgwardt, 2015] is built as follows. Take $A_\otimes$ as the adjacency matrix of the direct (tensor) product $G_\otimes = (V_\otimes, E_\otimes, \psi_\otimes)$ [Gärtner et al., 2003] between x and x' such that vertex labels match and edge labels match:

$$V_\otimes = \{(v, v') \in V \times V' | \psi(v) = \psi'(v')\},$$

$$E_\otimes = \{((v,u),(v',u'))) \in E \times E' \,|\, \psi(v,u) = \psi(v',u')\},$$

and use the corresponding label mapping $\psi_\otimes(v,v') = \psi(v) = \psi'(v')$; $\psi_\otimes((v,v'),(u,u')) = \psi(v,u) = \psi'(v',u')$. With input parameters $(\lambda_0, \dots, \lambda_k)$, the $k-$step random walk kernel between two graphs x, x' is defined as

$$K^k_\otimes(x, x') = \sum_{i,j=1}^{|V_\otimes|} \left[\sum_{t=0}^{k} \lambda_t A_\otimes^\top\right]_{i,j}.$$

A geometric random walk kernel between two graphs x, x' takes the λ-weighted infinite sum from the k step random walk kernels:

$$K_{GRW}(x, x') = \sum_{i,j=1}^{|V_\otimes|} \left[(I - \lambda A_\otimes)^{-1}\right]_{i,j}.$$

In our implementation we choose, $\lambda_l = \lambda, \forall l = 1, \dots, k$ and $\lambda = \frac{1}{3}$.

Shortest Path Graph Kernels Shortest Path Graph Kernels, introduced by Borgwardt and Kriegel [2005], are based on a transformation of the graph x, the Floyd transformation. The Floyd transformation F turns the original graph into the so-called shortest-path graph $y = F(x)$; the graph y is a complete graph with vertex set V with each edge labelled by the shortest distance in x between the vertices on either end of the edge. For two networks x and x' the 1-step random walk kernel $K^1_\otimes$ between the shortest-path graphs $y = F(x)$ and $y' = F(x')$ gives the shortest-path (SP) kernel between x and x';

$$K_{SP}(x, x') = K^1_\otimes(y, y').$$

Lemma 3 in Borgwardt and Kriegel [2005] showed that this kernel is positive definite.

Weisfeiler-Lehman Graph Kernels Weisfeiler-Lehman Graph Kernels have been proposed by Shervashidze et al. [2011]; these kernels are based on the Weisfeiler-Lehman test for graph isomorphisms and involve counting matching subtrees between two given graphs. Theorem 3 in Shervashidze et al. [2011] showed the positive definiteness of these kernels. In our implementation, we adapted an efficient implementation from the `graphkernel` package [Sugiyama et al., 2018].

C Vector-Valued RKHS

The general set-up for vector-valued RKHS for finite networks is as follows. Let $N = \binom{n}{2}$ denote the index set of vertex pairs in a graph $x \in \{0,1\}^N$. For $s \in [N]$ let $x^{-s} \in \{0,1\}^{N-1} =: \mathcal{X}^{-s}$ denote the collection of edge indicators except the one for s and let $x_s \in \{0,1\} =: \mathcal{X}^s$ denote the edge indicator for s. When the underlying graph is random, we use similar notation X^{-s}, X^s to denote the corresponding random variables . For $s \in [N]$, let $l_s : \mathcal{X}^s \times \mathcal{X}^s \to \mathbb{R}$ be reproducing kernels, with associated RKHS $\mathcal{H}_{l_s}$. Let $\varphi_s : x_s \in \mathcal{X}^s \mapsto l_s(\cdot, x^s) \in \mathcal{H}_{l_s}$ denote the corresponding feature maps of $(l_s)_{s\in[N]}$.

The RKHS kernels l_s, or those used in Chwialkowski et al. [2016] or Liu et al. [2016], have scalar outputs, while the RKHS kernel ℓ_{-s} has an output in $\mathcal{L}(\mathcal{H}_{l_s})$, the Banach space of bounded operators from $\mathcal{H}_{l_s}$ to $\mathcal{H}_{l_s}$; we refer to the space $\mathcal{H}_{l_s}$ for ℓ_{-s} as a vector-valued RKHS (vvRKHS). All the kernels used here are assumed to be positive definite and bounded. As composition preserves positive definiteness, we then consider the kernel: $K : (\mathcal{X}^s \otimes \mathcal{X}^{-s}) \times (\mathcal{X}^s \otimes \mathcal{X}^{-s}) \to \mathbb{R}$, with associate RKHS $\mathcal{H}_K$.

In our experiments we assume that the l_s corresponds to the same RKHS function: $l_s \equiv l, \forall s \in [N]$. We further assume the vvRKHS $\mathcal{H}_\ell$ has the form

$$\ell(x^{-s}, (x')^{-s'}) = k(x^{-s}, (x')^{-s'}) \mathbb{I}_{\mathcal{H}_l \times \mathcal{H}_l},$$

where $\mathbb{I}_{\mathcal{H}_l \times \mathcal{H}_l}$ is the identity map from $\mathcal{H}_l$ to $\mathcal{H}_l$ and k is the graph kernel of choice. The RKHS defined via composition reads

$$K((x^s, x^{-s}), ((x')^{s'}, (x')^{-s'})) = k(x^{-s}, (x')^{-s'}) l(x^s, (x')^{s'}).$$

For a single observed network x, as $\mathcal{H}_l$, $\mathcal{H}_\ell$ are the same for all s, it holds that for $s, s' \in [N]$:

$$K((x^s, x^{-s}), (x^{s'}, x^{-s'})) = l(x^s, x^{s'}).$$

In our implementation we use the kernels $k(x^{-s}, \cdot) = k(x^{(s,1)}, \cdot) + k(x^{(s,0)}, \cdot)$ from Section B, defined not on the whole graph x but on the set x^{-s}.

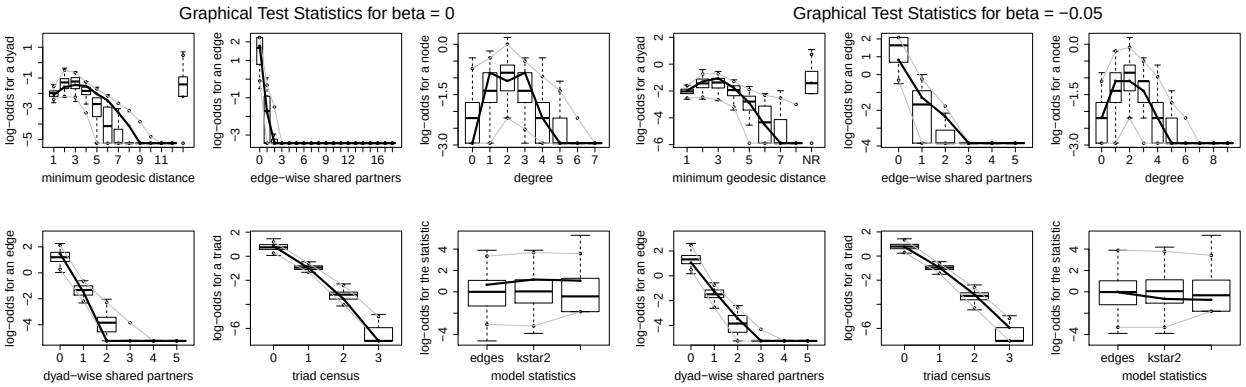
D Additional Details on Distance-based Test Statistics

D.1 Modified Graphical Tests with Total-Variation Distance

Here we give the details of the modified graphical tests based on Total-Variation (TV) distance, **mGra**, presented in Section 5. To assess the goodness-of-fit to a specific ERGM, Hunter et al. [2008] proposed to compare network statistics from the observed network to those of simulated networks from the null model via box plots and Monte-Carlo p-values. These network statistics are

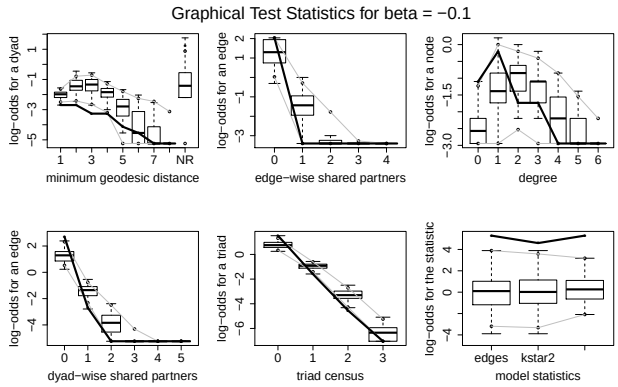
- the degree distribution, with d_k the number of vertices which have degree k;
- the number of edge-wise shared partners, which is the number of pairs of vertices which are neighbours and which have exactly k common neighbours;
- the number of dyad-wise shared partners, which is the number of pairs of vertices which have exactly k common neighbours (but are not necessarily themselves neighbours);
- the triad census, with 4 possible triads where triads are configurations on 3 vertices; the configurations are 0 edges, 1 edge, 2 edges and 3 edges;
- the statistics which are included in the ERGM as in Definition 1.

Fig. 2 shows an example of a graphical test based on the E2ST model Eq.(15) with the 2-star coefficient β_2 perturbed. By comparing whether the observed statistics (the bold line) deviates from the simulated null, one can visually assess whether the null hypothesis should be rejected. For instance, in Fig. 2(a) where the network is generated from the null distribution, the observed network statistics are all within the range in which 95 percent of the simulated observations fall.

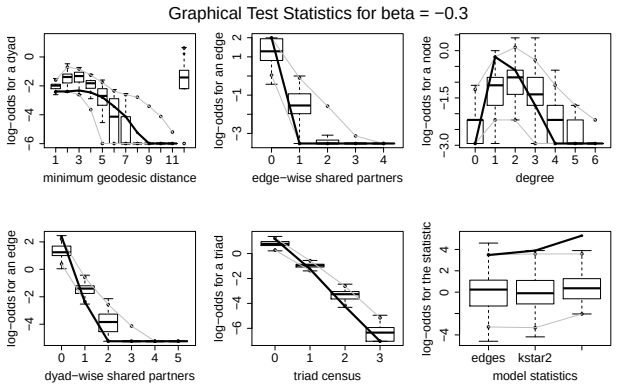


(a) The null model

(b) A small perturbation of the null model



(c) A moderate perturbation of the null model



(d) A larger perturbation of the null model

Figure 2: Graphical Tests with different beta parameters

When the difference between the null distribution and the distribution which generates the data is small, the graphical method may not easily distinguish the two models depending on the network statistics of choice. As shown in Fig. 2(b), with a network from a model with small perturbation from the null distribution, we see this effect. However, when the difference between data simulated under the null distribution and the data is substantial enough, we can see, e.g. from Fig. 2(c), that the minimum geodesic distance and the triad census from the observed network clearly differ from the simulated null, and the null hypothesis can be rejected. The box plots are also used to carry out Monte Carlo tests for each possible observation (for example a specific triad count) by giving a p-value for this specific test.

While every observed value can be used for a Monte Carlo test, Hunter et al. [2008] does not provide a systematic procedure to reach an overall conclusion about rejection. For instance, it is not clear whether the null is to be rejected when Fig. 2(d) is observed. To surpass such issue, we further develop the testing procedure by using the TV distance between distributions for the observed and simulated distributions of the summary statistics from Hunter et al. [2008]. Denote by S the random variable of a network statistic of choice and by $\mathcal{S}$ the space for S. Using the vertex degree of a simple undirected network on n vertices as an example, S is a discrete random variable taking values from 0 to $n-1$. Further denote by $S_{z'}$ the network statistic of an observation z' from the null model q and by s_x the network statistics from the observed x. Then with $\mathcal{R}$ denoting the set of possible values of S,

$$d_{TV}(S_{z'}; S_x) = \sup_{A \subset \mathcal{R}} |\mathbb{E}[h_A(S_{z'}) - h_A(S_x)]| = \frac{1}{2} \sum_{s \in \mathcal{S}} |P(S_{z'} = s) - P(S_x = s)|$$

where $h_A(s) = \mathbb{1}_{s \in A}$ is 1 if $s \in A$ and 0 otherwise. Our test statistic measures the distance between the

distribution of a network statistic S in the observed network x and under the null model q as follows:

$$D_{TV}(q, x; S) = \mathbb{E}_{z' \sim q}[d_{TV}(S_{z'}; S_x)].$$

To estimate $\mathbb{E}_q$, we simulate m' networks from the null model q, i.e. $z'_1, \ldots, z'_{m'} \sim q$ and use as empirical estimate for $D_{TV}(q, x; S)$

$$\widehat{D_{TV}}(q, x; S) = \frac{1}{m'} \sum_{j=1}^{m'} [d_{TV}(S_{z'_j}; S_x)].$$

To assess how the test statistics is distributed under the null hypothesis, i.e. $x \sim q$, we simulate $z \sim q$ from the null distribution. Similar to a Monte-Carlo test, we simulate independent network samples $z_1, \ldots, z_m \sim q$ and compute $\widehat{D_{TV}}(q, z_i; S)$, for $i \in [m]$. Then we reject the null if $\widehat{D_{TV}}(q, x; S)$ exceeds the $(1-\alpha)$-quantile level in the simulated observations $\{\widehat{D_{TV}}(q, z_1; S), \ldots, \widehat{D_{TV}}(q, z_m; S)\}$ under the null distribution.

D.2 Test Statistics with Mahalanobis Distance

Lospinoso and Snijders [2019] proposed using a Mahalanobis distance instead of the total variation distance. Suppose that a vector $S(x)$ of network summaries is observed and that the null distribution is parametrised by θ. Denote $\mu(\theta) = \mathbb{E}_\theta(X)$ as the expectation and $\Sigma(\theta) = Cov_\theta(X)$ as the covariance matrix under θ. The *Mahalanobis distance*

$$D_M(x, \theta; S) = (S(x) - \mu(\theta))^\top \Sigma(\theta)^{-1} (S(x) - \mu(\theta))$$

can then be used as test statistic. In practice, $\mu(\theta)$ and $\Sigma(\theta)$ are estimated using independent simulations $x_k, k = 1, \ldots, m$, from the distribution specified by θ;

$$\widehat{\mu} = \frac{1}{m} \sum_{k=1}^{m} S(x_k); \qquad \widehat{\Sigma} = \frac{1}{m} \sum_{k=1}^{m} (S(x_k) - \widehat{\mu})(S(x_k) - \widehat{\mu})^\top;$$

$$\widehat{D_M}(x) = (S(x) - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (S(x) - \widehat{\mu}).$$

The p-value of the test is estimated by the plug-in estimator

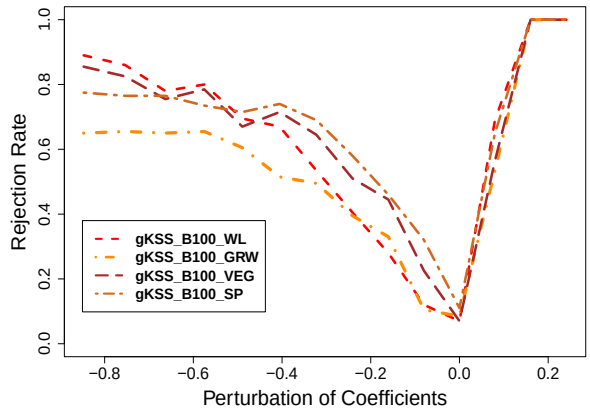
$$\widehat{p} = \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\widehat{D_M}(x_k) > \widehat{D_M}(x)\}.$$

In the main text this approach is abbreviated **MD** and applied to the degree distribution for ERGMs.

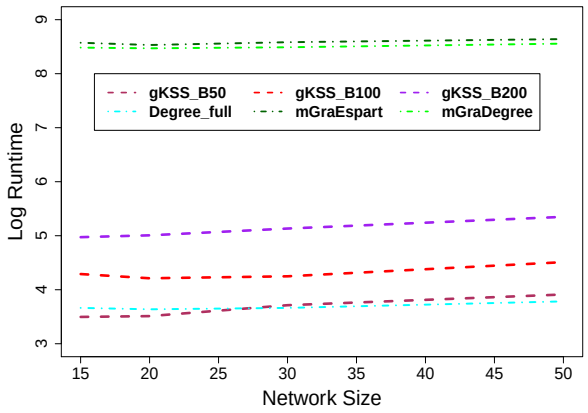
E Additional Experiment Results

Test performances with graph kernels Fig.3(a) shows the results for testing the E2ST model Eq.(15) with the 2-star coefficient β_2 perturbed using the different kernels described in Section B. Using the abbreviations from Section B, the relevant choices of kernel parameters are $\sigma = 1$ for the VEG kernel, level$= 5$ for the WL kernel, and $\lambda = \frac{1}{3}$ in the GRW kernel. Similar to the WL kernel used in the main text, the other choices of graph kernels achieve fairly good test power with the gKSS statistic. In our additional experimental results on the rejection rate, the re-sample size is $B = 100$ for all kernel choices. From Fig.3(a) we see that the test power is slightly higher with a small perturbed coefficient when the SP kernel and the VEG kernel are employed, while for larger perturbed coefficient (resulting in sparser graphs) the WL kernel better distinguishes the observation from the null. For large negative β_2 the GRW kernel has the poorest rejection rate. These differences in performance are no surprise as different choices of kernel emphasise different aspect of graph topology.

Computational time In Fig.3(b), we give more results for the computational time for one test, with 1,000 simulated networks. These results complement the reported results of Table 1 in the main text. As the number of vertices in the network increases, there is an increase in the computational complexity. However, as the main computation costs come from simulating the ERGM, we see from the plot that the slope is not substantial compared to the difference in testing procedures.



(a) $n = 20$, $\alpha = 0.05$, with β_2 in Eq.(15) perturbed



(b) log computational time for one test, in seconds, with $m = 1000$ simulated networks

Figure 3: Additional experiment results

F Comparison with the Kernel Discrete Stein Discrepancy on Testing Goodness-of-fit

F.1 Discrete Stein Operator

In this section, we compare our approach with the discrete Stein operator introduced in Yang et al. [2018]. First we need some definitions.

Definition 2. *[Definition 1 [Yang et al., 2018]](Cyclic permutation). For a set $\mathcal{N}$ of finite cardinality, a cyclic permutation $\neg : \mathcal{N} \to \mathcal{N}$ is a bijective function such that for some ordering $x^{[1]}, x^{[2]}, \ldots, x^{[|X|]}$ of the elements in $\mathcal{N}$, $\neg x^{[i]} = x^{[(i+1) mod |X|]}, \forall i = 1, 2, \ldots, |X|$.*

Definition 3. *[Definition 2 [Yang et al., 2018]] Given a cyclic permutation $\neg$ on $\mathcal{N}$, for any d-dimensional vector $x = (x_1, ..., x_d)^\top \in \mathcal{N}^d$, write $\neg_i x := (x_1, \ldots, x_{i-1}, \neg x_i, x_{i+1}, \ldots, x_d)^\top$. For any function $f : \mathcal{N}^d \to \mathbb{R}$, denote the (partial) difference operator as*

$$\Delta_{x_i} f(x) := f(x) - f(\neg_i x), \qquad i = 1, \ldots, d$$

and introduce the difference operator:

$$\Delta_\neg f(x) := (\Delta_{x_1} f(x), \ldots, \Delta_{x_d} f(x))^\top.$$

Here we use the notation $\Delta_\neg$ to distinguish it from the notation in the main text, where we used $\Delta_s h(x) = h(x^{(s,1)}) - h(x^{(s,0)})$ and $||\Delta h|| = \sup_{s \in [N]} |\Delta_s h(x)|$.

For discrete distributions q, Yang et al. [2018] proposed the following discrete Stein operator, which is based on the difference operator $\Delta_\neg$ constructed from a cyclic permutation:

$$\mathcal{T}_q^D f(x) = f(x) \frac{\Delta_\neg q(x)}{q(x)} - \Delta_\neg^* f(x), \tag{21}$$

where $\Delta_\neg^*$ denotes the adjoint operator of $\Delta_\neg$.

In particular, for q the distribution of an ERGM, with $\mathcal{N} = \{0, 1\}^N$, the discrete Stein operator proposed [Yang et al., 2018] can be written in the form of $\mathcal{T}_q^D f(x) = \sum_s \mathcal{T}_q^{D,s} f(x)$ where

$$\mathcal{T}_q^{D,s} f(x) = (-1)^{\mathbb{1}_{\{x = x^{(s,0)}\}}} \frac{f(x^{(s,1)}) q(x^{(s,0)}) - f(x^{(s,0)}) q(x^{(s,1)})}{q(x)}. \tag{22}$$

Recall the ERGM Stein operator of the form $\mathcal{T}_q f(x) = \frac{1}{N}\sum_{s\in[N]} \mathcal{T}_q^{(s)} f(x)$ and Eq.(5),

$$\begin{aligned}\mathcal{T}_q^{(s)} f(x) &= q(x^{(s,1)}|x)\Delta_s f(x) + \Big(f(x^{(s,0)}) - f(x)\Big)\\ &= \frac{q(x^{(s,1)})}{q(x^{(s,1)}) + q(x^{(s,0)})}\Big(f(x^{(s,1)}) - f(x^{(s,0)})\Big) + \Big(f(x^{(s,0)}) - f(x)\Big)\\ &= \frac{\mathbb{1}_{\{x=x^{(s,0)}\}} q(x^{(s,1)}) - \mathbb{1}_{\{x=x^{(s,1)}\}} q(x^{(s,0)})}{q(x^{(s,1)}) + q(x^{(s,0)})}\Big(f(x^{(s,1)}) - f(x^{(s,0)})\Big).\end{aligned}$$

We illustrate the difference between the ERGM Stein operator and the discrete Stein operator for a Bernoulli random graph with $\mathbb{P}(s=1) = q, \forall s$. Due to the independence, we have $q(x^{(s,1)}|x) = q$ and $q(x^{(s,0)}|x) = 1-q$. With Eq.(5), our Stein operator becomes

$$\mathcal{T}_q f(x) = \frac{1}{N}\sum_s (q - x_s)\left(f(x^{(s,1)}) - f(x^{(s,0)})\right). \tag{23}$$

The KDSD in this case can be written as:

$$\mathcal{T}_q^D f(x) = \frac{1}{q(x)}\sum_s (-1)^{1-x_s}\Big((1-q)f(x^{(s,1)}) - qf(x^{(s,0)})\Big)$$

with $q(x) = q^{\sum_s x_s}(1-q)^{N-\sum_s x_s}$. Thus, for different values, Eq.(23) is a weighted sum of the terms $\big(f(x^{s,1}) - f(x^{s,0})\big)$, while KDSD is a weighted sum of the terms $\big((1-q)f(x^{s,1}) - qf(x^{s,0})\big)$ and requires the calculation of the binomial probability $q(x)$.

The operators in Eq.(22) and Eq.(5) clearly differ in their scaling as well as in their repercussions for re-sampling. While the operator in Eq.(5) emerges from Glauber dynamics and hence has a natural re-sampling interpretation, no such interpretation is available for the operator in Eq.(22). Explicitly, the discrete Stein operator $\mathcal{T}_q^D$ has $q(x)$ in the denominator, indicating the fixed x realisation; however, the Stein ERGM operator $\mathcal{T}_q$ has $q(x^{-s})$ in the denominator which stems from the conditioning in Glauber dynamics. Consequently, the corresponding Stein discrepancy (called KSDS) differs from Eq.(2) in the main text, and, although usually only one network is available, the goodness-of-fit test in Yang et al. [2018] requires independent and identically distributed network observations.

A second key difference is that the test in Yang et al. [2018] requires the support of the unknown network distribution to be identical to the support of the ERGM which is described by q. In practice this condition is difficult if not impossible to verify. In contrast, $\widehat{\text{gKSS}}$ does not make any such assumption.

F.2 Comparison Between Graph Testing

Testing with multiple graph observations The relevant kernel discrete Stein discrepancy (KDSD) from the discrete Stein operator [Yang et al., 2018] is defined via taking the supreme over appropriate unit ball RKHS test functions, similar as in Eq.(2)

$$\text{KDSD}(q\|p;\mathcal{H}) = \sup_{\|f\|_{\mathcal{H}}\le 1} \mathbb{E}_p[\mathcal{T}_q^D f(x)]. \tag{24}$$

Yang et al. [2018] built a goodness-of-fit testing procedure based on the KDSD for ERGM for multiple graph observations. Let $x_1, \ldots, x_m \sim p$, be m independent identically distributed graph observations. The KDSD is empirically estimated from the observed samples; and as the number of observed samples $m \to \infty$, in probability,

$$\frac{1}{m}\sum_i [\mathcal{T}_q^D f(x_i)] \to \mathbb{E}_p[\mathcal{T}_q^D f(x)].$$

The rejection threshold is determined via a wild-bootstrap procedure [Chwialkowski et al., 2014].

While the $\widehat{\text{gKSS}}$ type of statistics based on the ERGM Stein operator in Eq.(4), $\mathcal{T}_q f(x) = \frac{1}{N}\sum_{s\in[N]} \mathcal{T}_q^{(s)} f(x)$, focuses on a single graph observation, this ERGM Stein operator could similarly be used to assess goodness-of-fit

when multiple graph observations are available. In particular, $\mathbb{E}_q[\mathcal{T}_q f(x)] = 0$. Hence, we introduce the graph kernel Stein discrepancy (gKSD) as

$$\text{gKSD}(q\|p;\mathcal{H}) = \sup_{\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_p[\mathcal{T}_q f(x)] = \sup_{\|f\|_{\mathcal{H}}\leq 1} \mathbb{E}_p\left[\frac{1}{N}\sum_s \mathcal{T}_q^{(s)} f(x)\right].$$

Here the sum is taken over all N pairs of vertices and the expectation is taken with respect to the ERGM q. When there are m independent observations $x_1, \ldots, x_m \sim p$ available then we can empirically estimate $\text{gKSD}(q\|p;\mathcal{H})$ by $\frac{1}{m}\sum_i[\frac{1}{N}\sum_s \mathcal{T}_q^{(s)} f(x_i)]$, which is weakly consistent by the law of large numbers. Then we use this statistic to build a goodness-of-fit test for multiple graph observations, determining the threshold via the same wild-bootstrap procedure as for the KDSD.

To compare the KDSD and the gKSD tests we consider the goodness-of-fit test setting as studied in Yang et al. [2018], using the E2ST model as presented in Eq.(15). We set the null parameters β to $(\beta_1, \beta_2, \beta_3) = (-2, 0.0, 0.01)$ and carry out a test at significance level $\alpha = 0.05$ using 100 repeats. For the alternative, we perturb the coefficient for 2-stars, β_2, and report the rejection rate in Table 3. Note that $\beta_2 = 0.00$ recovers the null distribution. In this experiment, with a small number of graph observations, $m = 30$, gKSD captures the difference between the null model and the alternative model better, resulting in a higher test power, compared to KDSD. Both gKSD and KDSD have higher power for $\beta_2 > 0$ than for $\beta_2 < 0$ of the same magnitude. This finding is plausible as increasing β_2 leads to denser networks.

β_2	-0.1	-0.08	-0.06	-0.04	-0.02	0.00	0.02	0.04	0.06	0.08	0.1
gKSD	0.32	0.30	0.24	0.14	0.10	0.04	0.08	0.22	0.18	0.28	0.54
KDSD	0.08	0.05	0.6	0.04	0.01	0.02	0.03	0.03	0.06	0.12	0.16

Table 3: Rejection rate for the E2ST model $(\beta_1, \beta_2, \beta_3) = (-2, 0, 0.01)$ with perturbation of the 2-star coefficient β_2: W.L. Kernel of level 3; sample size $m = 30$; graph size $n = 20$; test significance level $\alpha = 0.05$.

Testing with a single graph observation The ERGM Stein operator satisfies the mean zero property Eq.(6) when flipping each edge s given the rest of the graph x_{-s}. This is a key ingredient that KDSD does not satisfy; KDSD relies on a cyclic permutation as in Definition 2 to construct the partial difference operator in Definition 3, which depends on the order sequence of the cyclic permutation. As such, the mean zero property of their Stein operator is based on sign flips in each state of the discrete variable, instead of flipping the edge probability. Thus, the discrete Stein operator $\mathcal{T}_q^D$ could not easily be adapted to construct a subsampled Stein statistic such as $\widehat{\text{gKSS}}$ to perform goodness-of-fit testing with a single graph observation.

Testing with a few graph observations An interesting setting which is related to that of a single graph observation, is that a few graphs are observed, with the number of graphs assumed to be finite and not tending infinity with network size. With the proposed gKSD goodness-of-fit test for a single graph observation, a possible approach and a potential future research direction is applying multiple tests of goodness-of-fit, one for each observed network, with a Bonferroni correction [Bonferroni, 1936] to correct for multiple testing.