Graph-based Two-Sample Tests for Data with Repeated Observations

Jingru Zhang and Hao Chen

University of Pennsylvania

and University of California, Davis

Supplementary Material

Section S1 provides proofs of the lemmas and theorems. Section S2 discusses the issues of the existing graph-based tests. Section S3 contains some additional results.

S1 Proofs for lemmas and theorems

S1.1 Proof of Theorem 2

Proof. In the following, we prove the decomposition of $S_{(a)}$ (Equation (3.5)) in details. The proof for the decomposition of $S_{(u)}$ (Equation (3.6)) can be obtained similarly and is omitted here.

Let

$$\mathbf{R}_{(a)} = \begin{pmatrix} R_{1,(a)} - \mathsf{E}(R_{1,(a)}) \\ R_{2,(a)} - \mathsf{E}(R_{2,(a)}) \end{pmatrix},$$

$$\mathbf{Z}_{(a)} = \begin{pmatrix} Z_{w,(a)} \\ \\ Z_{d,(a)} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}}\frac{n_2-1}{N-2} & \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}}\frac{n_1-1}{N-2} \\ \frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} & -\frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} \end{pmatrix} \mathbf{R}_{(a)} \triangleq \mathbf{B}_{(a)}\mathbf{R}_{(a)}.$$

It is easy to see that $B_{(a)}$ is invertible. From the definition of $S_{(a)}$ (Equation (3.3)), it can be written as

$$S_{(a)} = \mathbf{R}_{(a)}^T\mathbf{\Sigma}_{(a)}^{-1}\mathbf{R}_{(a)} = (\mathbf{B}_{(a)}^{-1}\mathbf{Z}_{(a)})^T\mathbf{\Sigma}_{(a)}^{-1}(\mathbf{B}_{(a)}^{-1}\mathbf{Z}_{(a)}) = \mathbf{Z}_{(a)}^T(\mathbf{B}_{(a)}\mathbf{\Sigma}_{(a)}\mathbf{B}_{(a)}^T)^{-1}\mathbf{Z}_{(a)}.$$

We calculate $\mathbf{B}_{(a)}\mathbf{\Sigma}_{(a)}\mathbf{B}_{(a)}^T$ as follows:

$$\mathbf{B}_{(a)}\mathbf{\Sigma}_{(a)}\mathbf{B}_{(a)}^T = \begin{pmatrix} \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}}\frac{n_2-1}{N-2} & \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}}\frac{n_1-1}{N-2} \\ \frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} & -\frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} \end{pmatrix} \cdot$$
$$\begin{pmatrix} \mathsf{Var}(R_{1,(a)}) & \mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) \\ \mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) & \mathsf{Var}(R_{2,(a)}) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}}\frac{n_2-1}{N-2} & \frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} \\ \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}}\frac{n_1-1}{N-2} & -\frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} \end{pmatrix}$$
$$\triangleq \begin{pmatrix} \textcircled{1} & \textcircled{2} \\ \textcircled{2} & \textcircled{3} \end{pmatrix},$$

where

$$\textcircled{1} = \frac{\mathsf{Var}(R_{1,(a)})(n_2-1)^2 + 2\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})(n_1-1)(n_2-1) + \mathsf{Var}(R_{2,(a)})(n_1-1)^2}{\mathsf{Var}(R_{w,(a)})(N-2)^2} = 1,$$
$$\textcircled{2} = \frac{(n_2-1)(\mathsf{Var}(R_{1,(a)}) - \mathsf{Cov}(R_{1,(a)}, R_{2,(a)})) + (n_1-1)(\mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) - \mathsf{Var}(R_{2,(a)}))}{(N-2)\sqrt{\mathsf{Var}(R_{d,(a)})\mathsf{Var}(R_{w,(a)})}}$$
$$=0, \text{ and}$$
$$\textcircled{3} = \frac{\mathsf{Var}(R_{1,(a)}) - 2\mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) + \mathsf{Var}(R_{2,(a)})}{\mathsf{Var}(R_{d,(a)})} = 1.$$

Thus, $\mathbf{B}_{(a)}\mathbf{\Sigma}_{(a)}\mathbf{B}_{(a)}^T = \mathbf{I}_{2\times 2}$ and we have $S_{(a)} = \mathbf{Z}_{(a)}^T\mathbf{Z}_{(a)} = Z_{w,(a)}^2 + Z_{d,(a)}^2$.

Note that $R_{d,(a)} = R_{1,(a)} - R_{2,(a)}$ and $R_{d,(u)} = R_{1,(u)} - R_{2,(u)}$. Plugging the analytic expressions of $\mathsf{E}(R_{1,(a)})$, $\mathsf{E}(R_{2,(a)})$, $\mathsf{Var}(R_{1,(a)})$, $\mathsf{Var}(R_{2,(a)})$, $\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})$, $\mathsf{E}(R_{1,(u)})$, $\mathsf{E}(R_{2,(u)})$, $\mathsf{Var}(R_{1,(u)})$, $\mathsf{Var}(R_{2,(u)})$ and $\mathsf{Cov}(R_{1,(u)}, R_{2,(u)})$ given in Lemmas 1 and 2 in the Supplementary Material S1.4, we can obtain the expectations and variances of $R_{d,(a)}$ and $R_{d,(u)}$. □

S1.2 Proof of Theorem 3

Proof. Applying Stein's method, we prove $(R_{1,(a)}, R_{2,(a)})'$ converges in distribution to a bivariate Gaussian distribution as $N \to \infty$ first. Consider sums of the form $W = \sum_{i\in\mathcal{J}} \xi_i$, where $\mathcal{J}$ is an index set and ξ_i are random variables with $\mathsf{E}(\xi_i) = 0$, and $\mathsf{E}(W^2) = 1$. The following assumption restricts the dependence between $\{\xi_i : i \in \mathcal{J}\}$.

Assumption S1.1. [Chen and Shao (2005), p. 17] For each $i \in \mathcal{J}$ there exists $S_i \subset T_i \subset \mathcal{J}$ such that ξ_i is independent of $\xi_{S_i^c}$ and ξ_{S_i} is independent of $\xi_{T_i^c}$.

We will use the following theorem.

Theorem 1. *[Chen and Shao (2005), Theorem 3.4] Under Assumption*

S1.1, we have

$$\sup_{h\in Lip(1)} |\mathsf{E}h(W) - \mathsf{E}h(Z)| \leq \delta$$

where $Lip(1) = \{h : \mathbb{R} \to \mathbb{R}, \ \ \| h' \| \leq 1\}$, *Z has* $\mathcal{N}(0, 1)$ *distribution and*

$$\delta = 2\sum_{i\in\mathcal{J}}(\mathsf{E}|\xi_i\eta_i\theta_i| + |\mathsf{E}(\xi_i\eta_i)|\mathsf{E}|\theta_i|) + \sum_{i\in\mathcal{J}}\mathsf{E}|\xi_i\eta_i^2|$$

with $\eta_i = \sum_{j\in S_i}\xi_j$ *and* $\theta_i = \sum_{j\in T_i}\xi_j$, *where* S_i *and* T_i *are defined in Assumption S1.1.*

To prove Theorem 3, we take one step back to study the statistic under the bootstrap null distribution, which is defined as follows: For each observation, we assign it to be from Sample A with probability n_1/N, and from Sample B with probability n_2/N, independently of other observations. Let n_1^B be the number of observations that are assigned to be from Sample A. Then, conditioning on $n_1^B = n_1$, the bootstrap null distribution becomes the permutation null distribution. We use $\mathsf{P}_\mathsf{B}, \mathsf{E}_\mathsf{B}, \mathsf{Var}_\mathsf{B}$ to denote the probability, expectation, and variance under the bootstrap null distribution, respectively.

Let

$$\mathsf{E}(R_{1,(a)}) \triangleq \mu_1, \quad \mathsf{E}(R_{2,(a)}) \triangleq \mu_2,$$

$$\mathsf{Var}(R_{1,(a)}) \triangleq (\sigma_1)^2, \quad \mathsf{Var}(R_{2,(a)}) \triangleq (\sigma_2)^2, \quad \mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) \triangleq \sigma_{12},$$

$$p_4 = \left(\frac{n_1}{N}\right)^2, \quad p_5 = \left(\frac{n_1}{N}\right)^3, \quad p_6 = \left(\frac{n_1}{N}\right)^4,$$

$$q_4 = \left(\frac{n_2}{N}\right)^2, \quad q_5 = \left(\frac{n_2}{N}\right)^3, \quad q_6 = \left(\frac{n_2}{N}\right)^4.$$

Using similar steps as those in Lemma 1 in the Supplementary Material S1.4, we have

$$\begin{aligned}
\mathsf{E}_{\mathsf{B}}(R_{1,(a)}) =&(N-K+|C_0|)p_4 \triangleq \mu_1^B,\\
\mathsf{E}_{\mathsf{B}}(R_{2,(a)}) =&(N-K+|C_0|)q_4 \triangleq \mu_2^B,\\
\mathsf{Var}_{\mathsf{B}}(R_{1,(a)}) =&4(p_5-p_6)(N-K+2|C_0|+\sum_u \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u} - \sum_u \frac{|\mathcal{E}_u^{C_0}|}{m_u}) + 2(p_4-4p_5\\
&+3p_6)(K-\sum_u \frac{1}{m_u}) + (p_4-2p_5+p_6)\sum_{(u,v)\in C_0}\frac{1}{m_u m_v} \triangleq (\sigma_1^B)^2,\\
\mathsf{Var}_{\mathsf{B}}(R_{2,(a)}) =&4(q_5-q_6)(N-K+2|C_0|+\sum_u \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u} - \sum_u \frac{|\mathcal{E}_u^{C_0}|}{m_u}) + 2(q_4-4q_5\\
&+3q_6)(K-\sum_u \frac{1}{m_u}) + (q_4-2q_5+q_6)\sum_{(u,v)\in C_0}\frac{1}{m_u m_v} \triangleq (\sigma_2^B)^2.
\end{aligned}$$

Let

$$\begin{aligned}
W_1^B &= \frac{R_{1,(a)}-\mu_1^B}{\sigma_1^B}, \quad W_1 = \frac{R_{1,(a)}-\mu_1}{\sigma_1},\\
W_2^B &= \frac{R_{2,(a)}-\mu_2^B}{\sigma_2^B}, \quad W_2 = \frac{R_{2,(a)}-\mu_2}{\sigma_2},\\
W_3^B &= \frac{n_1^B - Np_a}{\sigma_0}, \text{ where } p_a = \frac{n_1}{N}, \sigma_0^2 = Np_a(1-p_a).
\end{aligned}$$

Under the conditions of Theorem 3, as $N \to \infty$, we can prove the following results:

(1) (W_1^B, W_2^B, W_3^B) becomes multivariate Gaussian distributed under the bootstrap null.

(2)

$$\frac{\sigma_1^B}{\sigma_1} \to c_1, \quad \frac{\mu_1^B - \mu_1}{\sigma_1^B} \to 0; \quad \frac{\sigma_2^B}{\sigma_2} \to c_2, \quad \frac{\mu_2^B - \mu_2}{\sigma_2^B} \to 0,$$

where c_1 and c_2 are constants.

(3) $|\lim_{N\to\infty} \mathsf{Cor}(W_1, W_2)| < 1$.

From (1) and given that $\mathsf{Var_B}(W_3^B) = 1$, the conditional distribution of $(W_1^B, W_2^B)'$ given W_3^B is a bivariate Gaussian distribution under the bootstrap null distribution as $N \to \infty$. Since the permutation null distribution is equivalent to the bootstrap null distribution given $W_3^B = 0$, (W_1^B, W_2^B) follows a bivariate Gaussian distribution under the permutation null distribution as $N \to \infty$. Since

$$W_1 = \frac{\sigma_1^B}{\sigma_1}\left(W_1^B + \frac{\mu_1^B - \mu_1}{\sigma_1^B}\right), \quad W_2 = \frac{\sigma_2^B}{\sigma_2}\left(W_2^B + \frac{\mu_2^B - \mu_2}{\sigma_2^B}\right),$$

given (2), we have (W_1, W_2) follows a bivariate Gaussian distribution under the permutation null distribution as $N \to \infty$. Together with (3), we have the conclusion that $(R_{1,(a)}, R_{2,(a)})'$ converges in distribution to a bivariate Gaussian distribution as $N \to \infty$. In the following, we prove the results (1)—(3).

To prove (1), by Cramér-Wold device, we only need to show that $W = a_1W_1^B + a_2W_2^B + a_3W_3^B$ is asymptotically Gaussian distributed for any combination of a_1, a_2, a_3 such that $\mathsf{Var_B}(W) > 0$.

We first define more notations.

For any node u of C_0, *i.e.* $u \in \mathcal{J}_1 = \{1, \cdots, K\}$, let

$$R_u^{(1)} = \frac{n_{1u}(n_{1u}-1)}{m_u}, \quad d_u^{(1)} = \mathsf{E}_{\mathsf{B}}(R_u^{(1)}) = (m_u - 1)p_4, \quad \xi_u^{(1)} = \frac{R_u^{(1)} - d_u^{(1)}}{\sigma_1^B},$$

$$R_u^{(2)} = \frac{n_{2u}(n_{2u}-1)}{m_u}, \quad d_u^{(2)} = \mathsf{E}_{\mathsf{B}}(R_u^{(2)}) = (m_u - 1)q_4, \quad \xi_u^{(2)} = \frac{R_u^{(2)} - d_u^{(2)}}{\sigma_2^B}.$$

For any edge (u, v) of C_0, *i.e.* $uv \in \mathcal{J}_2 = \{uv : u < v, (u, v) \in C_0\}$, let

$$R_{uv}^{(1)} = \frac{n_{1u}n_{1v}}{m_u m_v}, \quad d_{uv}^{(1)} = \mathsf{E}_{\mathsf{B}}(R_{uv}^{(1)}) = p_4, \quad \xi_{uv}^{(1)} = \frac{R_{uv}^{(1)} - d_{uv}^{(1)}}{\sigma_1^B},$$

$$R_{uv}^{(2)} = \frac{n_{2u}n_{2v}}{m_u m_v}, \quad d_{uv}^{(2)} = \mathsf{E}_{\mathsf{B}}(R_{uv}^{(2)}) = q_4, \quad \xi_{uv}^{(2)} = \frac{R_{uv}^{(2)} - d_{uv}^{(2)}}{\sigma_2^B}.$$

And for any $i \in \mathcal{J}_3 = \{|\mathcal{J}_0| + 1, \cdots, |\mathcal{J}_0| + K\}$, $\mathcal{J}_0 = \mathcal{J}_1 \cup \mathcal{J}_2$, let

$$\xi_i^{(3)} = \frac{n_{1i'} - p_a m_{i'}}{\sigma_0}, \quad i' = i - |\mathcal{J}_0|.$$

Thus,

$$W_1^B = \frac{R_{1,(a)} - \mu_1^B}{\sigma_1^B} = \sum_{i \in \mathcal{J}_0} \xi_i^{(1)},$$

$$W_2^B = \frac{R_{2,(a)} - \mu_2^B}{\sigma_2^B} = \sum_{i \in \mathcal{J}_0} \xi_i^{(2)},$$

$$W_3^B = \frac{n_1^B - Np_a}{\sigma_0} = \sum_{i \in \mathcal{J}_3} \xi_i^{(3)},$$

$$W = a_1 W_1^B + a_2 W_2^B + a_3 W_3^B = \sum_{i \in \mathcal{J}_0} (a_1 \xi_i^{(1)} + a_2 \xi_i^{(2)}) + \sum_{i \in \mathcal{J}_3} a_3 \xi_i^{(3)} \triangleq \sum_{i \in \mathcal{J}} \xi_i,$$

where $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_3$,

$$\xi_i = \begin{cases} a_1\xi_u^{(1)} + a_2\xi_u^{(2)}, & \text{if } i = u \in \mathcal{J}_1, \\ a_1\xi_{uv}^{(1)} + a_2\xi_{uv}^{(2)}, & \text{if } i = uv \in \mathcal{J}_2, \\ a_3\xi_u^{(3)}, & \text{if } i = u \in \mathcal{J}_3. \end{cases} \tag{S1.1}$$

We introduce following index sets to satisfy Assumption S1.1.

For $u \in \mathcal{J}_1$, let

$$\begin{aligned} S_u =& \{u, u + |\mathcal{J}_0|\} \cup \{uv, vu : (u, v) \in C_0\}, \\ T_u =& S_u \cup \{v, v + |\mathcal{J}_0|, vw, wv : (u, v), (v, w) \in C_0\}. \end{aligned}$$

For $uv \in \mathcal{J}_2$, let

$$\begin{aligned} S_{uv} =& \{uv, u, v, u + |\mathcal{J}_0|, v + |\mathcal{J}_0|\} \cup \{uw, wu : (u, w) \in C_0\} \\ & \cup \{vw, wv : (v, w) \in C_0\}, \\ T_{uv} =& S_{uv} \cup \{w, w + |\mathcal{J}_0|, wy, yw : (u, w), (w, y) \in C_0\} \\ & \cup \{w, w + |\mathcal{J}_0|, wy, yw : (v, w), (w, y) \in C_0\}. \end{aligned}$$

And for $u \in \mathcal{J}_3$, let

$$\begin{aligned} S_u =& \{u, u'\} \cup \{u'v, vu' : (u', v) \in C_0\}, \quad u' = u - |\mathcal{J}_0|, \\ T_u =& S_u \cup \{v, v + |\mathcal{J}_0|, vw, wv : (u', v), (v, w) \in C_0\}. \end{aligned}$$

Let $a = \max\{|a_1|, |a_2|, |a_3|\}, \sigma = \min(\sigma_1^B, \sigma_2^B, \sigma_0)$. Since $R_u^{(1)} \in [0, m_u - 1], p_4 \in [0, 1]$, and $R_{uv}^{(1)} \in [0, 1]$, we have $d_u^{(1)} \in [0, m_u - 1], d_{uv}^{(1)} \in [0, 1]$, and therefore

$$|\xi_u^{(1)}| \le \frac{m_u}{\sigma_1^B} \le \frac{m_u}{\sigma} \quad u \in \mathcal{J}_1; \quad |\xi_{uv}^{(1)}| \le \frac{1}{\sigma_1^B} \le \frac{1}{\sigma} \quad uv \in \mathcal{J}_2.$$

Similarly,

$$|\xi_u^{(2)}| \le \frac{m_u}{\sigma_2^B} \le \frac{m_u}{\sigma} \quad u \in \mathcal{J}_1; \quad |\xi_{uv}^{(2)}| \le \frac{1}{\sigma_2^B} \le \frac{1}{\sigma} \quad uv \in \mathcal{J}_2.$$

Since $n_{1u'} \in [0, m_{u'}], p_a \in [0,1]$, we have $|\xi_u^{(3)}| \le \frac{m_{u'}}{\sigma_0} \le \frac{m_{u'}}{\sigma}$ with $u \in \mathcal{J}_3, u' = u - |\mathcal{J}_0|$. Plugging these inequations into (S1.1),

$$|\xi_i| \le \begin{cases} \frac{2a}{\sigma} m_u, & \text{if } i = u \in \mathcal{J}_1, \\ \frac{2a}{\sigma}, & \text{if } i = uv \in \mathcal{J}_2, \\ \frac{a}{\sigma} m_{u'}, & \text{if } i = u \in \mathcal{J}_3. \end{cases}$$

Hence,

$$\sum_{j \in S_u} |\xi_j| \le \frac{2a}{\sigma} 2m_u + \frac{2a}{\sigma} |\mathcal{E}_u^{C_0}| \le \frac{4a}{\sigma}(m_u + |\mathcal{E}_u^{C_0}|), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,$$

$$\sum_{j \in T_u} |\xi_j| \le \frac{2a}{\sigma}(2m_u + 2 \sum_{v \in \mathcal{V}_u^{C_0}} m_v) + \frac{2a}{\sigma} |\mathcal{E}_{u,2}^{C_0}|$$
$$\le \frac{4a}{\sigma}(m_u + \sum_{v \in \mathcal{V}_u^{C_0}} m_v + |\mathcal{E}_{u,2}^{C_0}|), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,$$

$$\sum_{j \in S_{uv}} |\xi_j| \le \frac{2a}{\sigma}(2m_u + 2m_v) + \frac{2a}{\sigma}(|\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|)$$
$$\le \frac{4a}{\sigma}(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|), \quad uv \in \mathcal{J}_2,$$

$$\begin{aligned}\sum_{j\in T_{uv}}|\xi_j| \leq \frac{2a}{\sigma}(2m_u+2m_v+2\sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}}m_w)+\frac{2a}{\sigma}(|\mathcal{E}_{u,2}^{C_0}|+|\mathcal{E}_{v,2}^{C_0}|)\\ \leq \frac{4a}{\sigma}(m_u+m_v+\sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}}m_w+|\mathcal{E}_{u,2}^{C_0}|+|\mathcal{E}_{v,2}^{C_0}|), \quad uv\in\mathcal{J}_2.\end{aligned}$$

For $i=u\in\mathcal{J}_1\cup\mathcal{J}_3$, the terms $\mathsf{E}_\mathsf{B}|\xi_i\eta_i\theta_i|, |\mathsf{E}_\mathsf{B}(\xi_i\eta_i)|\mathsf{E}_\mathsf{B}|\theta_i|$, and $\mathsf{E}_\mathsf{B}|\xi_u\eta_i^2|$ are all bounded by

$$\frac{32a^3}{\sigma^3}m_u(m_u+|\mathcal{E}_u^{C_0}|)(m_u+\sum_{v\in\mathcal{V}_u^{C_0}}m_v+|\mathcal{E}_{u,2}^{C_0}|),$$

and for $i=uv\in\mathcal{J}_2$, the terms $\mathsf{E}_\mathsf{B}|\xi_i\eta_i\theta_i|, |\mathsf{E}_\mathsf{B}(\xi_i\eta_i)|\mathsf{E}_\mathsf{B}|\theta_i|$, and $\mathsf{E}_\mathsf{B}|\xi_u\eta_i^2|$ are all bounded by

$$\frac{32a^3}{\sigma^3}(m_u+m_v+|\mathcal{E}_u^{C_0}|+|\mathcal{E}_v^{C_0}|)(m_u+m_v+\sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}}m_w+|\mathcal{E}_{u,2}^{C_0}|+|\mathcal{E}_{v,2}^{C_0}|).$$

So we have $\sup_{h\in Lip(1)}|\mathsf{E}_\mathsf{B}h(\bar{W})-\mathsf{E}_\mathsf{B}h(Z_0)|\leq\delta$ with $\bar{W}=\frac{W}{\sqrt{\mathsf{Var}_\mathsf{B}(W)}}$, $Z_0\sim\mathcal{N}(0,1)$, and

$$\begin{aligned}\delta =&\frac{1}{\sqrt{\mathsf{Var}_\mathsf{B}^3(W)}}\left\{2\sum_{i\in\mathcal{J}}(\mathsf{E}_\mathsf{B}|\xi_i\eta_i\theta_i|+|\mathsf{E}_\mathsf{B}(\xi_i\eta_i)|\mathsf{E}_\mathsf{B}|\theta_i|)+\sum_{i\in\mathcal{J}}\mathsf{E}_\mathsf{B}|\xi_i\eta_i^2|\right\}\\ =&\frac{1}{\sqrt{\mathsf{Var}_\mathsf{B}^3(W)}}\left\{2\sum_{i\in\mathcal{J}_1\cup\mathcal{J}_3}(\mathsf{E}_\mathsf{B}|\xi_i\eta_i\theta_i|+|\mathsf{E}_\mathsf{B}(\xi_i\eta_i)|\mathsf{E}_\mathsf{B}|\theta_i|)+\sum_{i\in\mathcal{J}_1\cup\mathcal{J}_3}\mathsf{E}_\mathsf{B}|\xi_i\eta_i^2|\right.\\ &\left.+2\sum_{i\in\mathcal{J}_2}(\mathsf{E}_\mathsf{B}|\xi_i\eta_i\theta_i|+|\mathsf{E}_\mathsf{B}(\xi_i\eta_i)|\mathsf{E}_\mathsf{B}|\theta_i|)+\sum_{i\in\mathcal{J}_2}\mathsf{E}_\mathsf{B}|\xi_i\eta_i^2|\right\}\\ \leq&\frac{320a^3}{\sigma^3\sqrt{\mathsf{Var}_\mathsf{B}^3(W)}}\left\{\sum_{u=1}^K m_u(m_u+|\mathcal{E}_u^{C_0}|)(m_u+\sum_{v\in\mathcal{V}_u^{C_0}}m_v+|\mathcal{E}_{u,2}^{C_0}|)+\sum_{(u,v)\in C_0}(m_u\right.\\ &\left.+m_v+|\mathcal{E}_u^{C_0}|+|\mathcal{E}_v^{C_0}|)(m_u+m_v+\sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}}m_w+|\mathcal{E}_{u,2}^{C_0}|+|\mathcal{E}_{v,2}^{C_0}|)\right\}.\end{aligned}$$

Since σ_1^B, σ_2^B are of order $\sqrt{N}$ or higher and σ_0 is of order $\sqrt{N}$ or higher, σ is at least of order $\sqrt{N}$ by Condition 1. Thus, under Condition 2, $\delta \to 0$ as $N \to \infty$.

Next we prove result (2). The equations for $(\sigma_1)^2$ and $(\sigma_1^B)^2$ can be reorganized as

$$\begin{aligned}
(\sigma_1)^2 =& \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{ 4\frac{n_1-2}{n_2-1}\left[\sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \frac{(|C_0|-K)^2}{N}\right]\right. \\
&\left. + 2(K - \sum_u \frac{1}{m_u}) + \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} - \frac{2}{N(N-1)}(|C_0|+N-K)^2 \right\}, \\
(\sigma_1^B)^2 =& \frac{n_1^2 n_2^2}{N^4}\left\{ 4\frac{n_1}{n_2}\left[N - 2K + 2|C_0| + \sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u}\right]\right. \\
&\left. + 2(K - \sum_u \frac{1}{m_u}) + \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} \right\}.
\end{aligned}$$

Assume $n_1/N \to p > 0$ and $n_2/N \to q > 0$ as $N \to \infty$. According to Conditions 1 and 3, we assume

$$\frac{1}{N}\left[\sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \frac{(|C_0|-K)^2}{N}\right] \longrightarrow b_1,$$

$$\frac{1}{N}\sum_u \frac{1}{m_u} \longrightarrow b_2, \quad \frac{1}{N}\sum_{(u,v)\in C_0} \frac{1}{m_u m_v} \longrightarrow b_3, \quad \frac{1}{N}|C_0| \longrightarrow b_4, \quad \frac{K}{N} \longrightarrow b_5,$$

as $N \to \infty$, where $b_1, b_3, b_4 \in (0,\infty); b_2, b_5 \in [0,1], b_2 \leq b_5$. The value ranges of b_2, b_3, b_4, b_5 are obvious. $b_1 > 0$ can be proved by solving the following optimization problem.

$$\begin{cases} \text{minimize } h(\mathbf{m}) = \sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \frac{(|C_0|-K)^2}{N}, \quad \mathbf{m} = (m_1, \cdots, m_K)', \\ \text{s.t. } \sum_{u=1}^K m_u = N, \quad m_u > 0, \end{cases}$$

$$\Longrightarrow \begin{cases} \min\ h(\mathbf{m}) = \frac{1}{4N}\left\{\left[\sum_{u=1}^K \left||\mathcal{E}_u^{C_0}| - 2\right|\right]^2 - 4(|C_0|-K)^2\right\} \geq \\ \frac{1}{4N}\left\{\left[\sum_{u=1}^K (|\mathcal{E}_u^{C_0}| - 2)\right]^2 - 4(|C_0|-K)^2\right\} = 0, \\ \hat{m}_u = \arg\min\ h(\mathbf{m}) = \frac{N\left||\mathcal{E}_u^{C_0}|-2\right|}{\sum_{u=1}^K \left||\mathcal{E}_u^{C_0}|-2\right|}, \quad u = 1, \cdots, K. \end{cases}$$

Then

$$\frac{1}{N}\left[N - 2K + 2|C_0| + \sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u}\right] \longrightarrow b_1 + (1 + b_4 - b_5)^2.$$

So as $N \to \infty$,

$$\begin{aligned} \frac{(\sigma_1)^2}{N} \longrightarrow & p^2q^2\left\{4\frac{p}{q}b_1 + 2(b_5 - b_2) + b_3\right\}, \\ \frac{(\sigma_1^B)^2}{N} \longrightarrow & p^2q^2\left\{4\frac{p}{q}\left[b_1 + (1 + b_4 - b_5)^2\right] + 2(b_5 - b_2) + b_3\right\}, \\ \frac{\sigma_1^B}{\sigma_1} \longrightarrow & \sqrt{\frac{4p\left[b_1 + (1 + b_4 - b_5)^2\right] + 2(b_5 - b_2)q + b_3q}{4pb_1 + 2(b_5 - b_2)q + b_3q}} \\ = & \sqrt{1 + \frac{4p(1 + b_4 - b_5)^2}{4pb_1 + 2(b_5 - b_2)q + b_3q}}. \end{aligned}$$

Similarly, we have

$$\frac{\sigma_2^B}{\sigma_2} \longrightarrow \sqrt{1 + \frac{4q(1 + b_4 - b_5)^2}{4qb_1 + 2(b_5 - b_2)p + b_3p}}.$$

Also,

$$\mu_1^B - \mu_1 = (N - K + |C_0|)(p_4 - p_1) = (N - K + |C_0|)\frac{n_1n_2}{N^2(N-1)},$$

so

$$\lim_{N\to\infty}\frac{\mu_1^B-\mu_1}{\sigma_1^B}=\lim_{N\to\infty}\frac{(1+b_4-b_5)pq}{\sigma_1^B}=0,$$

since $\sigma_1^B=O(N^{0.5})$. Similarly, we have

$$\lim_{N\to\infty}\frac{\mu_2^B-\mu_2}{\sigma_2^B}=0.$$

Last, we prove result (3). Rewrite $\mathsf{Cov}(R_{1,(a)},R_{2,(a)})$ as

$$\begin{aligned}\sigma_{12}=&\frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{-4\left[\sum_u\frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u}-\frac{(|C_0|-K)^2}{N}\right]\right.\\&\left.+2(K-\sum_u\frac{1}{m_u})+\sum_{(u,v)\in C_0}\frac{1}{m_um_v}-\frac{2}{N(N-1)}(|C_0|+N-K)^2\right\}.\end{aligned}$$

As $N\to\infty$,

$$\begin{aligned}\frac{\sigma_{12}}{N}\longrightarrow&p^2q^2\{-4b_1+2(b_5-b_2)+b_3\},\\\sqrt{\frac{(\sigma_1)^2}{N}\frac{(\sigma_2)^2}{N}}\longrightarrow&p^2q^2\sqrt{\left[4\frac{p}{q}b_1+2(b_5-b_2)+b_3\right]\left[4\frac{q}{p}b_1+2(b_5-b_2)+b_3\right]}\\=&p^2q^2\sqrt{[-4b_1+2(b_5-b_2)+b_3]^2+\frac{4b_1}{pq}[2(b_5-b_2)+b_3]},\end{aligned}$$

$$\begin{aligned}\lim_{N\to\infty}\mathsf{Cor}(W_1,W_2)=&\lim_{N\to\infty}\frac{\sigma_{12}}{\sqrt{(\sigma_1)^2(\sigma_2)^2}}\\=&\frac{-4b_1+2(b_5-b_2)+b_3}{\sqrt{[-4b_1+2(b_5-b_2)+b_3]^2+\frac{4b_1}{pq}[2(b_5-b_2)+b_3]}}.\end{aligned}$$

Strictly positive $\frac{4b_1}{pq}[2(b_5-b_2)+b_3]$ implies (3).

Since $Z_{w,(a)}$ and $Z_{d,(a)}$ are the standardizations of $R_{w,(a)}$ and $R_{d,(a)}$, the proof above implies $(Z_{w,(a)}, Z_{d,(a)})'$ converges in distribution to a bivariate Gaussian distribution $\mathbf{N}_2(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right))$. If we can prove $\rho = \lim_{N\to\infty} \mathsf{Cov}(Z_{w,(a)}, Z_{d,(a)}) = 0$, then the proof of Theorem 3 would be completed.

Since

$$\begin{aligned}
\mathsf{Cov}(R_{w,(a)}, R_{d,(a)}) &= \mathsf{Cov}(\hat{q}R_{1,(a)} + \hat{p}R_{2,(a)}, R_{1,(a)} - R_{2,(a)}) \\
&= \hat{q}[(\sigma_1)^2 - \sigma_{12}] - \hat{p}[(\sigma_2)^2 - \sigma_{12}] \\
&= 0,
\end{aligned}$$

where the last line can be shown by plugging

$$\hat{p} = \frac{\mathsf{Var}(R_{1,(a)}) - \mathsf{Cov}(R_{1,(a)}, R_{2,(a)})}{\mathsf{Var}(R_{1,(a)}) + \mathsf{Var}(R_{2,(a)}) - 2\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})}$$

and $\hat{q} = 1 - \hat{p}$, we have

$$\rho = \lim_{N\to\infty} \mathsf{Cov}(Z_{w,(a)}, Z_{d,(a)}) = \lim_{N\to\infty} \frac{\mathsf{Cov}(R_{w,(a)}, R_{d,(a)})}{\sqrt{\mathsf{Var}(R_{w,(a)})\mathsf{Var}(R_{d,(a)})}} = 0.$$

□

S1.3 Proof of Theorem 4

Proof. We will use Assumption S1.1 and Theorem 1 in Proof S1.2. The proof is similar to Proof S1.2, so we omit some arguments and notations here. Applying Stein's method, we prove $(R_{1,(u)}, R_{2,(u)})'$ converges in distribution to a bivariate Gaussian distribution as $N \to \infty$ first. Consider sums of the form $\tilde{W} = \sum_{i \in \mathcal{J}} \tilde{\xi}_i$, where $\mathcal{J}$ is an index set and $\tilde{\xi}_i$ are random variables with $\mathsf{E}(\tilde{\xi}_i) = 0$, and $\mathsf{E}(\tilde{W}^2) = 1$.

Let

$$\mathsf{E}(R_{1,(u)}) \triangleq \nu_1, \quad \mathsf{E}(R_{2,(u)}) \triangleq \nu_2,$$

$$\mathsf{Var}(R_{1,(u)}) \triangleq (\delta_1)^2, \quad \mathsf{Var}(R_{2,(u)}) \triangleq (\delta_2)^2, \quad \mathsf{Cov}(R_{1,(u)}, R_{2,(u)}) \triangleq \delta_{12},$$

$$p_4 = \left(\frac{n_1}{N}\right)^2, \quad p_5 = \left(\frac{n_1}{N}\right)^3, \quad p_6 = \left(\frac{n_1}{N}\right)^4,$$
$$q_4 = \left(\frac{n_2}{N}\right)^2, \quad q_5 = \left(\frac{n_2}{N}\right)^3, \quad q_6 = \left(\frac{n_2}{N}\right)^4.$$

Using similar steps as those in Lemma 2 in the Supplementary Material S1.4, we have

$$\begin{aligned}
\mathsf{E}_{\mathsf{B}}(R_{1,(u)}) =& |\bar{G}| p_4 \triangleq \nu_1^B, \\
\mathsf{E}_{\mathsf{B}}(R_{2,(u)}) =& |\bar{G}| q_4 \triangleq \nu_2^B, \\
\mathsf{Var}_{\mathsf{B}}(R_{1,(u)}) =& (p_4 - p_6)|\bar{G}| + (p_5 - p_6) \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1) \triangleq (\delta_1^B)^2, \\
\mathsf{Var}_{\mathsf{B}}(R_{2,(u)}) =& (q_4 - q_6)|\bar{G}| + (q_5 - q_6) \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1) \triangleq (\delta_2^B)^2.
\end{aligned}$$

Let

$$\begin{aligned}
&\tilde{W}_1^B = \frac{R_{1,(u)} - \nu_1^B}{\delta_1^B}, \quad \tilde{W}_1 = \frac{R_{1,(u)} - \nu_1}{\delta_1}, \\
&\tilde{W}_2^B = \frac{R_{2,(u)} - \nu_2^B}{\delta_2^B}, \quad \tilde{W}_2 = \frac{R_{2,(u)} - \nu_2}{\delta_2}, \\
&\tilde{W}_3^B = \frac{n_1^B - Np_a}{\delta_0}, \text{ where } p_a = \frac{n_1}{N}, \delta_0^2 = Np_a(1 - p_a).
\end{aligned}$$

Under the conditions of Theorem 4, as $N \to \infty$, we can prove the following results:

(1) $(\tilde{W}_1^B, \tilde{W}_2^B, \tilde{W}_3^B)$ becomes multivariate Gaussian distributed under the bootstrap null.

(2)
$$\frac{\delta_1^B}{\delta_1} \to c_1, \quad \frac{\nu_1^B - \nu_1}{\delta_1^B} \to 0; \quad \frac{\delta_2^B}{\delta_2} \to c_2, \quad \frac{\nu_2^B - \nu_2}{\delta_2^B} \to 0,$$
where c_1 and c_2 are constants.

(3) $|\lim_{N\to\infty} \mathsf{Cor}(\tilde{W}_1, \tilde{W}_2)| < 1$.

With the same argument as that in Proof S1.2. Theorem 4 could be proved if (1)–(3) are satisfied.

To prove (1), by Cramér-Wold device, we only need to show that $\tilde{W} = a_1\tilde{W}_1^B + a_2\tilde{W}_2^B + a_3\tilde{W}_3^B$ is asymptotically Gaussian distributed for any combination of a_1, a_2, a_3 such that $\mathsf{Var}_\mathsf{B}(\tilde{W}) > 0$.

We first define more notations.

For any node u of C_0, *i.e.* $u \in \mathcal{J}_1 = \{1, \cdots, K\}$, let

$$\tilde{R}_u^{(1)} = \frac{n_{1u}(n_{1u}-1)}{2}, \quad \tilde{d}_u^{(1)} = \mathsf{E}_\mathsf{B}(\tilde{R}_u^{(1)}) = \frac{m_u(m_u-1)}{2}p_4, \quad \tilde{\xi}_u^{(1)} = \frac{\tilde{R}_u^{(1)} - \tilde{d}_u^{(1)}}{\delta_1^B},$$

$$\tilde{R}_u^{(2)} = \frac{n_{2u}(n_{2u}-1)}{2}, \quad \tilde{d}_u^{(2)} = \mathsf{E}_\mathsf{B}(\tilde{R}_u^{(2)}) = \frac{m_u(m_u-1)}{2}q_4, \quad \tilde{\xi}_u^{(2)} = \frac{\tilde{R}_u^{(2)} - \tilde{d}_u^{(2)}}{\delta_2^B}.$$

For any edge (u,v) of C_0, *i.e.* $uv \in \mathcal{J}_2 = \{uv : u < v, (u,v) \in C_0\}$, let

$$\tilde{R}_{uv}^{(1)} = n_{1u}n_{1v}, \quad \tilde{d}_{uv}^{(1)} = \mathsf{E}_{\mathsf{B}}(\tilde{R}_{uv}^{(1)}) = m_u m_v p_4, \quad \tilde{\xi}_{uv}^{(1)} = \frac{\tilde{R}_{uv}^{(1)} - \tilde{d}_{uv}^{(1)}}{\delta_1^B},$$

$$\tilde{R}_{uv}^{(2)} = n_{2u}n_{2v}, \quad \tilde{d}_{uv}^{(2)} = \mathsf{E}_{\mathsf{B}}(\tilde{R}_{uv}^{(2)}) = m_u m_v q_4, \quad \tilde{\xi}_{uv}^{(2)} = \frac{\tilde{R}_{uv}^{(2)} - \tilde{d}_{uv}^{(2)}}{\delta_2^B},$$

And for any $i \in \mathcal{J}_3 = \{|\mathcal{J}_0| + 1, \cdots, |\mathcal{J}_0| + K\}$, $\mathcal{J}_0 = \mathcal{J}_1 \cup \mathcal{J}_2$, let

$$\tilde{\xi}_i^{(3)} = \frac{n_{1i'} - p_a m_{i'}}{\delta_0}, \quad i' = i - |\mathcal{J}_0|.$$

Thus,

$$\tilde{W}_1^B = \frac{R_{1,(u)} - \nu_1^B}{\delta_1^B} = \sum_{i \in \mathcal{J}_0} \tilde{\xi}_i^{(1)},$$

$$\tilde{W}_2^B = \frac{R_{2,(u)} - \nu_2^B}{\delta_2^B} = \sum_{i \in \mathcal{J}_0} \tilde{\xi}_i^{(2)},$$

$$\tilde{W}_3^B = \frac{n_1^B - Np_a}{\delta_0} = \sum_{i \in \mathcal{J}_3} \tilde{\xi}_i^{(3)},$$

$$\tilde{W} = a_1\tilde{W}_1^B + a_2\tilde{W}_2^B + a_3\tilde{W}_3^B = \sum_{i \in \mathcal{J}_0}(a_1\tilde{\xi}_i^{(1)} + a_2\tilde{\xi}_i^{(2)}) + \sum_{i \in \mathcal{J}_3} a_3\tilde{\xi}_i^{(3)} \triangleq \sum_{i \in \mathcal{J}} \tilde{\xi}_i,$$

where $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_3$,

$$\tilde{\xi}_i = \begin{cases} a_1\tilde{\xi}_u^{(1)} + a_2\tilde{\xi}_u^{(2)}, & \text{if } i = u \in \mathcal{J}_1, \\ a_1\tilde{\xi}_{uv}^{(1)} + a_2\tilde{\xi}_{uv}^{(2)}, & \text{if } i = uv \in \mathcal{J}_2, \\ a_3\tilde{\xi}_u^{(3)}, & \text{if } i = u \in \mathcal{J}_3. \end{cases} \tag{S1.2}$$

The definition of index sets $(S_u, T_u, S_{uv}, T_{uv})$ are same as those in Proof S1.2.

Let $a = \max\{|a_1|, |a_2|, |a_3|\}, \sigma = \min(\delta_1^B, \delta_2^B, \delta_0)$. Since $\tilde{R}_u^{(1)} \in [0, \frac{m_u(m_u-1)}{2}], p_4 \in [0,1]$, and $\tilde{R}_{uv}^{(1)} \in [0, m_u m_v]$, we have $\tilde{d}_u^{(1)} \in [0, \frac{m_u(m_u-1)}{2}], \tilde{d}_{uv}^{(1)} \in [0, m_u m_v]$, and therefore

$$|\tilde{\xi}_u^{(1)}| \leq \frac{m_u^2}{2\delta_1^B} \leq \frac{m_u^2}{2\sigma} \quad u \in \mathcal{J}_1; \quad |\tilde{\xi}_{uv}^{(1)}| \leq \frac{m_u m_v}{\delta_1^B} \leq \frac{m_u m_v}{\sigma} \quad uv \in \mathcal{J}_2.$$

Similarly,

$$|\tilde{\xi}_u^{(2)}| \leq \frac{m_u^2}{2\delta_2^B} \leq \frac{m_u^2}{2\sigma} \quad u \in \mathcal{J}_1; \quad |\tilde{\xi}_{uv}^{(2)}| \leq \frac{m_u m_v}{\delta_2^B} \leq \frac{m_u m_v}{\sigma} \quad uv \in \mathcal{J}_2.$$

Since $n_{1u'} \in [0, m_{u'}], p_a \in [0,1]$, we have $|\tilde{\xi}_u^{(3)}| \leq \frac{m_{u'}}{\delta_0} \leq \frac{m_{u'}}{\sigma}$ with $u \in \mathcal{J}_3, u' = u - |\mathcal{J}_0|$. Plugging these inequations into (S1.2),

$$|\tilde{\xi}_i| \leq \begin{cases} \frac{a}{\sigma} m_u^2, & \text{if } i = u \in \mathcal{J}_1, \\ \frac{2a m_u m_v}{\sigma}, & \text{if } i = uv \in \mathcal{J}_2, \\ \frac{a}{\sigma} m_{u'}, & \text{if } i = u \in \mathcal{J}_3. \end{cases}$$

Hence,

$$\sum_{j \in S_u} |\tilde{\xi}_j| \leq \frac{2a}{\sigma}(m_u^2 + m_u \sum_{v \in \mathcal{V}_u^{C_0}} m_v), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,$$

$$\sum_{j \in T_u} |\tilde{\xi}_j| \leq \frac{2a}{\sigma}(m_u^2 + \sum_{v \in \mathcal{V}_u^{C_0}} m_v^2 + m_u \sum_{v \in \mathcal{V}_u^{C_0}} m_v + \sum_{v \in \mathcal{V}_u^{C_0}, w \in \mathcal{V}_v^{C_0}} m_v m_w), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,$$

$$\sum_{j\in S_{uv}} |\tilde{\xi}_j| \le \frac{2a}{\sigma}(m_u^2 + m_v^2 + m_u \sum_{w\in\mathcal{V}_u^{C_0}} m_w + m_v \sum_{w\in\mathcal{V}_v^{C_0}} m_w), \quad uv \in \mathcal{J}_2,$$

$$\sum_{j\in T_{uv}} |\tilde{\xi}_j| \le \frac{2a}{\sigma}(m_u^2 + m_v^2 + \sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}} m_w^2 + m_u \sum_{w\in\mathcal{V}_u^{C_0}} m_w + m_v \sum_{w\in\mathcal{V}_v^{C_0}} m_w + \sum_{\substack{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}\\ y\in\mathcal{V}_w^{C_0}}} m_w m_y),$$

$$uv \in \mathcal{J}_2.$$

For $i = u \in \mathcal{J}_1 \cup \mathcal{J}_3$, the terms $\mathsf{E}_\mathsf{B}|\tilde{\xi}_i\tilde{\eta}_i\tilde{\theta}_i|, |\mathsf{E}_\mathsf{B}(\tilde{\xi}_i\tilde{\eta}_i)|\mathsf{E}_\mathsf{B}|\tilde{\theta}_i|$, and $\mathsf{E}_\mathsf{B}|\tilde{\xi}_u\tilde{\eta}_i^2|$ are all bounded by

$$\frac{8a^3}{\sigma^3} m_u^3 (m_u + \sum_{v\in\mathcal{V}_u^{C_0}} m_v) \sum_{v\in\{u\}\cup\mathcal{V}_u^{C_0}} m_v (m_v + \sum_{w\in\mathcal{V}_v^{C_0}} m_w),$$

and for $i = uv \in \mathcal{J}_2$, the terms $\mathsf{E}_\mathsf{B}|\tilde{\xi}_i\tilde{\eta}_i\tilde{\theta}_i|, |\mathsf{E}_\mathsf{B}(\tilde{\xi}_i\tilde{\eta}_i)|\mathsf{E}_\mathsf{B}|\tilde{\theta}_i|$, and $\mathsf{E}_\mathsf{B}|\tilde{\xi}_u\tilde{\eta}_i^2|$ are all bounded by

$$\frac{8a^3}{\sigma^3} m_u m_v \left[m_u (m_u + \sum_{w\in\mathcal{V}_u^{C_0}} m_w) + m_v (m_v + \sum_{w\in\mathcal{V}_v^{C_0}} m_w) \right] \left[\sum_{\substack{w\in\{u\}\cup\{v\}\cup\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}\\ y\in\mathcal{V}_w^{C_0}}} m_w (m_w + m_y) \right].$$

So we have $\sup_{h\in Lip(1)} |\mathsf{E}_\mathsf{B} h(\bar{W}) - \mathsf{E}_\mathsf{B} h(Z_0)| \le \delta$ with $\bar{W} = \frac{\tilde{W}}{\sqrt{\mathsf{Var}_\mathsf{B}(\tilde{W})}}$,

$Z_0 \sim \mathcal{N}(0,1)$, and

$$
\begin{aligned}
\delta =& \frac{1}{\sqrt{\mathsf{Var}_{\mathsf{B}}^3(\tilde{W})}}\left\{2\sum_{i\in\mathcal{J}}(\mathsf{E}_{\mathsf{B}}|\tilde{\xi}_i\tilde{\eta}_i\tilde{\theta}_i| + |\mathsf{E}_{\mathsf{B}}(\tilde{\xi}_i\tilde{\eta}_i)|\mathsf{E}_{\mathsf{B}}|\tilde{\theta}_i|) + \sum_{i\in\mathcal{J}}\mathsf{E}_{\mathsf{B}}|\tilde{\xi}_i\tilde{\eta}_i^2|\right\}\\
=& \frac{1}{\sqrt{\mathsf{Var}_{\mathsf{B}}^3(\tilde{W})}}\left\{2\sum_{i\in\mathcal{J}_1\cup\mathcal{J}_3}(\mathsf{E}_{\mathsf{B}}|\tilde{\xi}_i\tilde{\eta}_i\tilde{\theta}_i| + |\mathsf{E}_{\mathsf{B}}(\tilde{\xi}_i\tilde{\eta}_i)|\mathsf{E}_{\mathsf{B}}|\tilde{\theta}_i|) + \sum_{i\in\mathcal{J}_1\cup\mathcal{J}_3}\mathsf{E}_{\mathsf{B}}|\tilde{\xi}_i\tilde{\eta}_i^2|\right.\\
&\left. + 2\sum_{i\in\mathcal{J}_2}(\mathsf{E}_{\mathsf{B}}|\tilde{\xi}_i\tilde{\eta}_i\tilde{\theta}_i| + |\mathsf{E}_{\mathsf{B}}(\tilde{\xi}_i\tilde{\eta}_i)|\mathsf{E}_{\mathsf{B}}|\tilde{\theta}_i|) + \sum_{i\in\mathcal{J}_2}\mathsf{E}_{\mathsf{B}}|\tilde{\xi}_i\tilde{\eta}_i^2|\right\}\\
\leq& \frac{80a^3}{\sigma^3\sqrt{\mathsf{Var}_{\mathsf{B}}^3(\tilde{W})}}\left\{\sum_{u=1}^{K} m_u^3(m_u + \sum_{v\in\mathcal{V}_u^{C_0}} m_v)\sum_{v\in\{u\}\cup\mathcal{V}_u^{C_0}} m_v(m_v + \sum_{w\in\mathcal{V}_v^{C_0}} m_w)\right.\\
&+ \sum_{(u,v)\in C_0} m_u m_v\left[m_u(m_u + \sum_{w\in\mathcal{V}_u^{C_0}} m_w) + m_v(m_v + \sum_{w\in\mathcal{V}_v^{C_0}} m_w)\right]\\
&\left.\cdot\left[\sum_{\substack{w\in\{u\}\cup\{v\}\cup\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}\\ y\in\mathcal{V}_w^{C_0}}} m_w(m_w + m_y)\right]\right\}.
\end{aligned}
$$

Since δ_1^B, δ_2^B are of order $\sqrt{N}$ or higher and δ_0 is of order $\sqrt{N}$ or higher, σ is at least of order $\sqrt{N}$ by Condition 4. Thus, under Condition 6, $\delta \to 0$ as $N \to \infty$.

Next we prove result (2). The equations for $(\delta_1)^2$ and $(\delta_1^B)^2$ can be reorganized as

$$\begin{aligned}(\delta_1)^2 =&\frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{|\bar{G}| + \frac{n_1-2}{n_2-1}\left[\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2 - \frac{4n_1}{(n_1-2)(N-1)}|\bar{G}|^2\right]\right.\\ &\left.+6\frac{N+n_1-1}{(n_2-1)N(N-1)}|\bar{G}|^2\right\},\\ (\delta_1^B)^2 =&\frac{n_1^2n_2^2}{N^4}\left\{|\bar{G}| + \frac{n_1}{n_2}\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2\right\}.\end{aligned}$$

Assume $n_1/N \to p > 0$ and $n_2/N \to q > 0$ as $N \to \infty$. According to Conditions 4 and 5, we assume

$$\frac{|\bar{G}|}{N} \longrightarrow d_1, \quad \frac{\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2 - \frac{4}{N}|\bar{G}|^2}{N} \longrightarrow d_2,$$

as $N \to \infty$, where $d_1, d_2 \in (0, \infty)$. Then

$$\frac{\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2}{N} \longrightarrow 4d_1^2 + d_2.$$

So as $N \to \infty$,

$$\begin{aligned}\frac{(\delta_1)^2}{N} \longrightarrow& p^2q^2\left\{d_1 + \frac{p}{q}d_2\right\},\\ \frac{(\delta_1^B)^2}{N} \longrightarrow& p^2q^2\left\{d_1 + \frac{p}{q}(4d_1^2+d_2)\right\},\\ \frac{\delta_1^B}{\delta_1} \longrightarrow& \sqrt{\frac{d_1q + (4d_1^2+d_2)p}{d_1q+d_2p}} = \sqrt{1 + \frac{4d_1^2p}{d_1q+d_2p}}.\end{aligned}$$

Similarly, we have

$$\frac{\delta_2^B}{\delta_2} \longrightarrow \sqrt{1 + \frac{4d_1^2q}{d_1p+d_2q}}.$$

Also,

$$\nu_1^B - \nu_1 = |\bar{G}|(p_4 - p_1) = |\bar{G}|\frac{n_1n_2}{N^2(N-1)},$$

so

$$\lim_{N\to\infty} \frac{\nu_1^B - \nu_1}{\delta_1^B} = \lim_{N\to\infty} \frac{d_1 pq}{\delta_1^B} = 0,$$

since $\delta_1^B = O(N^{0.5})$. Similarly, we have

$$\lim_{N\to\infty} \frac{\nu_2^B - \nu_2}{\delta_2^B} = 0.$$

Last, we prove result (3). Rewrite $\mathsf{Cov}(R_{1,(u)}, R_{2,(u)})$ as

$$\delta_{12} = \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{|\bar{G}| - \left[\sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|^2 - \frac{4N-6}{N(N-1)}|\bar{G}|^2\right]\right\}.$$

So as $N \to \infty$,

$$\begin{aligned} \frac{(\delta_1)^2}{N} &\longrightarrow p^2q^2\{d_1 + \frac{p}{q}d_2\}, \\ \frac{(\delta_2)^2}{N} &\longrightarrow p^2q^2\{d_1 + \frac{q}{p}d_2\}, \\ \frac{\delta_{12}}{N} &\longrightarrow p^2q^2\{d_1 - d_2\}, \end{aligned}$$

$$\begin{aligned} \lim_{N\to\infty} \mathsf{Cor}(\tilde{W}_1, \tilde{W}_2) = \lim_{N\to\infty} \frac{\delta_{12}}{\sqrt{(\delta_1)^2(\delta_2)^2}} &= \frac{d_1 - d_2}{\sqrt{(d_1 + \frac{p}{q}d_2)(d_1 + \frac{q}{p}d_2)}} \\ &= \frac{d_1 - d_2}{\sqrt{(d_1 - d_2)^2 + \frac{d_1 d_2}{pq}}}. \end{aligned}$$

Strictly positive $\frac{d_1 d_2}{pq}$ implies (3).

Last, with some computation as in the last part of Proof S1.2, we obtain

$$\rho = \lim_{N\to\infty} \mathsf{Cov}(Z_{w,(u)}, Z_{d,(u)}) = 0.$$

□

S1.4 Analytic expressions of the expectation and variance for the extended basic quantities

Lemma 1. *The means, variances and covariance of $R_{1,(a)}$ and $R_{2,(a)}$ under the permutation null are*

$$
\begin{aligned}
\mathsf{E}(R_{1,(a)}) &= (N-K+|C_0|)p_1,\\
\mathsf{E}(R_{2,(a)}) &= (N-K+|C_0|)q_1,\\
\mathsf{Var}(R_{1,(a)}) =&4(p_2-p_3)(N-K+2|C_0|+\sum_u \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u}-\sum_u \frac{|\mathcal{E}_u^{C_0}|}{m_u})\\
&+(p_3-p_1^2)(N-K+|C_0|)^2+(p_1-2p_2+p_3)\sum_{(u,v)\in C_0}\frac{1}{m_u m_v}\\
&+2(p_1-4p_2+3p_3)(K-\sum_u \frac{1}{m_u}),\\
\mathsf{Var}(R_{2,(a)}) =&4(q_2-q_3)(N-K+2|C_0|+\sum_u \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u}-\sum_u \frac{|\mathcal{E}_u^{C_0}|}{m_u})\\
&+(q_3-q_1^2)(N-K+|C_0|)^2+(q_1-2q_2+q_3)\sum_{(u,v)\in C_0}\frac{1}{m_u m_v}\\
&+2(q_1-4q_2+3q_3)(K-\sum_u \frac{1}{m_u}),\\
\mathsf{Cov}(R_{1,(a)},R_{2,(a)}) =&(f_1-p_1q_1)(N-K+|C_0|)^2\\
&+f_1\Bigg[-4(N-K+2|C_0|+\sum_u \frac{|\mathcal{E}_u^{C_0}|^2}{4m_u}-\sum_u \frac{|\mathcal{E}_u^{C_0}|}{m_u})\\
&+6(K-\sum_u \frac{1}{m_u})+\sum_{(u,v)\in C_0}\frac{1}{m_u m_v}\Bigg],
\end{aligned}
$$

where

$$
p_1=\frac{n_1(n_1-1)}{N(N-1)},\quad p_2=\frac{n_1(n_1-1)(n_1-2)}{N(N-1)(N-2)},\quad p_3=\frac{n_1(n_1-1)(n_1-2)(n_1-3)}{N(N-1)(N-2)(N-3)},
$$

$$q_1 = \frac{n_2(n_2-1)}{N(N-1)}, \quad q_2 = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}, \quad q_3 = \frac{n_2(n_2-1)(n_2-2)(n_2-3)}{N(N-1)(N-2)(N-3)},$$

$$f_1 = \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}.$$

Proof. (I) Compute $\mathsf{E}(R_{1,(a)}), \mathsf{Var}(R_{1,(a)}), \mathsf{E}(R_{2,(a)})$ and $\mathsf{Var}(R_{2,(a)})$.

Define

$$\begin{aligned}
X_{1A} &= \sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u, i \neq j} \mathbf{I}_{g_i = g_j = 1}, \\
X_{1B} &= \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} \mathbf{I}_{g_i = g_j = 1}, \\
X_{2A} &= \sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u, i \neq j} \mathbf{I}_{g_i = g_j = 2}, \\
X_{2B} &= \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} \mathbf{I}_{g_i = g_j = 2}.
\end{aligned}$$

Since

$$\mathsf{P}(g_i = g_j = 1) = \frac{n_1(n_1-1)}{N(N-1)} = p_1 \quad \text{for } i \neq j,$$

$$\mathsf{P}(g_i = g_j = g_k = g_l = 1)$$

$$= \begin{cases} \frac{n_1(n_1-1)}{N(N-1)} = p_1, & \text{if } \begin{cases} i = k, j = l \\ i = l, j = k \end{cases} \\ \frac{n_1(n_1-1)(n_1-2)}{N(N-1)(N-2)} = p_2, & \text{if } \begin{cases} i = k, j \neq l \\ i = l, j \neq k \\ j = k, i \neq l \\ j = l, i \neq k \end{cases} \\ \frac{n_1(n_1-1)(n_1-2)(n_1-3)}{N(N-1)(N-2)(N-3)} = p_3, & \text{if } i, j, k, l \text{ are all different} \end{cases} \quad \text{for } i \neq j, k \neq l,$$

we have

$$
\begin{aligned}
&\mathsf{E}(R_{1,(a)})\\
=&\mathsf{E}(X_{1A})+\mathsf{E}(X_{1B})\\
=&\sum_{u=1}^{K}\frac{1}{m_u}\sum_{i,j\in\mathcal{C}_u,i\neq j}\mathsf{P}(g_i=g_j=1)\\
&+\sum_{(u,v)\in C_0}\frac{1}{m_um_v}\sum_{i\in\mathcal{C}_u,j\in\mathcal{C}_v}\mathsf{P}(g_i=g_j=1)\\
=&\sum_{u=1}^{K}\frac{1}{m_u}m_u(m_u-1)p_1+\sum_{(u,v)\in C_0}\frac{1}{m_um_v}m_um_vp_1\\
=&(N-K+|C_0|)p_1.
\end{aligned}
$$

Now, to compute the second moment, first note that

$$\mathsf{E}(R_{1,(a)}^2)=\mathsf{E}(X_{1A}^2)+\mathsf{E}(X_{1B}^2)+2\mathsf{E}(X_{1A}X_{1B}).$$

We calculate every summand on the right side of the above equation as follows.

$$
\begin{aligned}
&\mathsf{E}(X_{1A}^2)\\
=&\sum_{u,v=1}^{K}\frac{1}{m_u m_v}\sum_{i,j\in\mathcal{C}_u,k,l\in\mathcal{C}_v}\mathsf{P}(g_i=g_j=g_k=g_l=1)\\
=&\sum_{u=1}^{K}\frac{1}{m_u^2}\sum_{i,j,k,l\in\mathcal{C}_u}\mathsf{P}(g_i=g_j=g_k=g_l=1)\\
&+\sum_{u=1}^{K}\sum_{v\neq u}\frac{1}{m_u m_v}\sum_{i,j\in\mathcal{C}_u,k,l\in\mathcal{C}_v}\mathsf{P}(g_i=g_j=g_k=g_l=1)\\
=&\sum_{u=1}^{K}\frac{1}{m_u^2}[2m_u(m_u-1)p_1+4m_u(m_u-1)(m_u-2)p_2\\
&+m_u(m_u-1)(m_u-2)(m_u-3)p_3]\\
&+\sum_{u=1}^{K}\sum_{v\neq u}\frac{1}{m_u m_v}m_u(m_u-1)m_v(m_v-1)p_3\\
=&2Kp_1+4(N-3K)p_2+(8p_2-2p_1)\sum_{u=1}^{K}\frac{1}{m_u}\\
&+(N-K)(N-K-4)p_3+6(K-\sum_{u=1}^{K}\frac{1}{m_u})p_3,
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{E}(X_{1B}^2)\\
={}&\sum_{(u,v)\in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k\in\mathcal{C}_u, j,l\in\mathcal{C}_v} \mathsf{P}(g_i=g_j=g_k=g_l=1)\\
&+ \sum_{\substack{(u,v),(u,w)\in C_0\\ v\neq w}} \frac{1}{m_u^2 m_v m_w} \sum_{\substack{i,k\in\mathcal{C}_u\\ j\in\mathcal{C}_v, l\in\mathcal{C}_w}} \mathsf{P}(g_i=g_j=g_k=g_l=1)\\
&+ \sum_{\substack{(u,v),(w,h)\in C_0,\\ u,v,w,h \text{ all different}}} \frac{1}{m_u m_v m_w m_h} \sum_{\substack{i\in\mathcal{C}_u, j\in\mathcal{C}_v,\\ k\in\mathcal{C}_w, l\in\mathcal{C}_h}} \mathsf{P}(g_i=g_j=g_k=g_l=1)\\
={}&\sum_{(u,v)\in C_0}\frac{1}{m_u^2 m_v^2}[m_u m_v p_1 + (m_u m_v(m_v-1)\\
&+ m_v m_u(m_u-1))p_2 + m_u(m_u-1)m_v(m_v-1)p_3]\\
&+ \sum_{(u,v),(u,w)\in C_0} \frac{1}{m_u^2 m_v m_w}[m_u m_v m_w p_2 + m_u(m_u-1)m_v m_w p_3]\\
&+ \sum_{(u,v),(w,h)\in C_0} \frac{1}{m_u m_v m_w m_h} m_u m_v m_w m_h p_3\\
={}&\sum_{(u,v)\in C_0}\frac{1}{m_u m_v}[p_1+(m_u+m_v-2)p_2+(m_u-1)(m_v-1)p_3]\\
&+ \sum_{(u,v),(u,w)\in C_0}\frac{1}{m_u}[p_2+(m_u-1)p_3] + \sum_{\substack{(u,v),(w,h)\in C_0\\ u,v,w,h \text{ all different}}} p_3\\
={}&\sum_{u=1}^{K} \frac{|\mathcal{E}_u^{C_0}|^2}{m_u}(p_2-p_3) + |C_0|^2 p_3 + \sum_{(u,v)\in C_0}\frac{1}{m_u m_v}(p_1-2p_2+p_3),
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{E}(X_{1A}X_{1B})\\
=&\sum_{u=1}^{K}\sum_{(v,w)\in C_0}\frac{1}{m_u m_v m_w}\sum_{i,j\in\mathcal{C}_u,k\in\mathcal{C}_v,l\in\mathcal{C}_w}\mathsf{P}(g_i=g_j=g_k=g_l=1)\\
=&\sum_{u=1}^{K}\sum_{(u,v)\in\mathcal{E}_u^{C_0}}\frac{1}{m_u^2 m_v}\sum_{i,j,k\in\mathcal{C}_u,l\in\mathcal{C}_v}\mathsf{P}(g_i=g_j=g_k=g_l=1)\\
&+\sum_{u=1}^{K}\sum_{(v,w)\in C_0\setminus\mathcal{E}_u^{C_0}}\frac{1}{m_u m_v m_w}\sum_{\substack{i,j\in\mathcal{C}_u\\k\in\mathcal{C}_v,l\in\mathcal{C}_w}}\mathsf{P}(g_i=g_j=g_k=g_l=1)\\
=&\sum_{u=1}^{K}\sum_{(u,v)\in\mathcal{E}_u^{C_0}}\frac{1}{m_u^2 m_v}[2m_u(m_u-1)m_v p_2+m_u(m_u-1)(m_u-2)m_v p_3]\\
&+\sum_{u=1}^{K}\sum_{(v,w)\in C_0\setminus\mathcal{E}_u^{C_0}}\frac{1}{m_u m_v m_w}m_u(m_u-1)m_v m_w p_3\\
=&\sum_{u=1}^{K}\sum_{(u,v)\in\mathcal{E}_u^{C_0}}\left[\frac{2(m_u-1)p_2}{m_u}+\frac{(m_u-1)(m_u-2)}{m_u}p_3\right]\\
&+\sum_{u=1}^{K}\sum_{(v,w)\in C_0\setminus\mathcal{E}_u^{C_0}}(m_u-1)p_3\\
=&\sum_{u=1}^{K}\left[\left(2p_2|\mathcal{E}_u^{C_0}|-2p_2\frac{|\mathcal{E}_u^{C_0}|}{m_u}\right)+m_u|\mathcal{E}_u^{C_0}|p_3-3|\mathcal{E}_u^{C_0}|p_3+2p_3\frac{|\mathcal{E}_u^{C_0}|}{m_u}\right]\\
&+\sum_{u=1}^{K}(|C_0|-|\mathcal{E}_u^{C_0}|)(m_u-1)p_3\\
=&2(p_2-p_3)\left(2|C_0|-\sum_{u=1}^{K}\frac{|\mathcal{E}_u^{C_0}|}{m_u}\right)+|C_0|(N-K)p_3.
\end{aligned}
$$

$\mathsf{Var}(R_{1,(a)})$ follows by combining the equations above in computing $\mathsf{E}(R_{1,(a)}^2)$, and then subtracting $\mathsf{E}^2(R_{1,(a)})$.

Similarly, we can get $\mathsf{E}(R_{2,(a)})$ and $\mathsf{Var}(R_{2,(a)})$ with

$$\mathsf{P}(g_i = g_j = 2) = \frac{n_2(n_2-1)}{N(N-1)} = q_1 \quad \text{for } i \neq j,$$

$$\mathsf{P}(g_i = g_j = g_k = g_l = 2)$$

$$= \begin{cases} \frac{n_2(n_2-1)}{N(N-1)} = q_1, & \text{if } \begin{cases} i = k, j = l \\ i = l, j = k \end{cases} \\ \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} = q_2, & \text{if } \begin{cases} i = k, j \neq l \\ i = l, j \neq k \\ j = k, i \neq l \\ j = l, i \neq k \end{cases} \\ \frac{n_2(n_2-1)(n_2-2)(n_2-3)}{N(N-1)(N-2)(N-3)} = q_3, & \text{if } i, j, k, l \text{ are all different} \end{cases} \quad \text{for } i \neq j, k \neq l.$$

(II) Compute $\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})$.

Note that

$$\begin{aligned} & \mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) \\ =& \mathsf{Cov}(X_{1A} + X_{1B}, X_{2A} + X_{2B}) \\ =& \mathsf{Cov}(X_{1A}, X_{2A}) + \mathsf{Cov}(X_{1B}, X_{2B}) \\ & + \mathsf{Cov}(X_{1B}, X_{2A}) + \mathsf{Cov}(X_{1B}, X_{2B}). \end{aligned} \qquad \text{(S1.3)}$$

First $\mathsf{E}(X_{1A}), \mathsf{E}(X_{1B}), \mathsf{E}(X_{2A}), \mathsf{E}(X_{2B})$ can be calculated easily.

$$\begin{aligned}
\mathsf{E}(X_{1A}) &= \frac{n_1(n_1-1)}{N(N-1)}(N-K) = p_1(N-K),\\
\mathsf{E}(X_{1B}) &= \frac{n_1(n_1-1)}{N(N-1)}|C_0| = p_1|C_0|,\\
\mathsf{E}(X_{2A}) &= \frac{n_2(n_2-1)}{N(N-1)}(N-K) = q_1(N-K),\\
\mathsf{E}(X_{2B}) &= \frac{n_2(n_2-1)}{N(N-1)}|C_0| = q_1|C_0|.
\end{aligned}$$

Then we calculate each term on the right side of (S1.3). Since

$$\begin{aligned}
&\mathsf{P}(g_i = g_j = 1, g_k = g_l = 2) = \mathsf{P}(g_i = g_j = 2, g_k = g_l = 1)\\
=&\frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)} = f_1, \text{ if } i,j,k,l \text{ are all different},
\end{aligned}$$

we have

$$\begin{aligned}
\mathsf{E}(X_{1A}X_{2A}) =&\sum_{u=1}^{K}\frac{1}{m_u^2}\sum_{i,j,k,l\in\mathcal{C}_u}\mathsf{P}(g_i = g_j = 1, g_k = g_l = 2)\\
&+\sum_{u=1}^{K}\sum_{v\neq u}\frac{1}{m_u m_v}\sum_{\substack{i,j\in\mathcal{C}_u\\ k,l\in\mathcal{C}_v}}\mathsf{P}(g_i = g_j = 1, g_k = g_l = 2)\\
=&\sum_{u=1}^{K}\frac{1}{m_u^2}m_u(m_u-1)(m_u-2)(m_u-3)f_1\\
&+\sum_{u=1}^{K}\sum_{v\neq u}\frac{1}{m_u m_v}m_u(m_u-1)m_v(m_v-1)f_1\\
=&(N-K)(N-K-4)f_1 + 6(K - \sum_{u=1}^{K}\frac{1}{m_u})f_1,
\end{aligned}$$

$$\begin{aligned}
&\mathsf{Cov}(X_{1A}, X_{2A})\\
=&\mathsf{E}(X_{1A}X_{2A}) - \mathsf{E}(X_{1A})\mathsf{E}(X_{2A})\\
=&(N-K)(N-K-4))f_1 + 6(K - \sum_{u=1}^{K}\frac{1}{m_u})f_1 - p_1q_1(N-K)^2,
\end{aligned}$$

$$\begin{aligned}
&\mathsf{E}(X_{1A}X_{2B})\\
=&\sum_{u=1}^{K}\sum_{(u,v)\in\mathcal{E}_u^{C_0}}\frac{1}{m_u^2m_v}\sum_{\substack{i,j,k\in\mathcal{C}_u\\ l\in\mathcal{C}_v}}\mathsf{P}(g_i=g_j=1,g_k=g_l=2)\\
&+\sum_{u=1}^{K}\sum_{(v,w)\in C_0\backslash\mathcal{E}_u^{C_0}}\frac{1}{m_um_vm_w}\sum_{\substack{i,j\in\mathcal{C}_u\\ k\in\mathcal{C}_v,l\in\mathcal{C}_w}}\mathsf{P}(g_i=g_j=1,g_k=g_l=2)\\
=&\sum_{u=1}^{K}\sum_{(u,v)\in\mathcal{E}_u^{C_0}}\frac{1}{m_u^2m_v}m_u(m_u-1)(m_u-2)m_vf_1\\
&+\sum_{u=1}^{K}\sum_{(v,w)\in C_0\backslash\mathcal{E}_u^{C_0}}\frac{1}{m_um_vm_w}m_u(m_u-1)m_vm_wf_1\\
=&|C_0|(N-K)f_1 - 2(2|C_0| - \sum_{u=1}^{K}\frac{|\mathcal{E}_u^{C_0}|}{m_u})f_1,
\end{aligned}$$

$$\begin{aligned}
&\mathsf{Cov}(X_{1A}, X_{2B})\\
=&\mathsf{E}(X_{1A}X_{2B}) - \mathsf{E}(X_{1A})\mathsf{E}(X_{2B})\\
=&|C_0|(N-K)f_1 - 2(2|C_0| - \sum_{u=1}^{K}\frac{|\mathcal{E}_u^{C_0}|}{m_u})f_1 - p_1q_1|C_0|(N-K),
\end{aligned}$$

$$
\begin{aligned}
&\mathsf{Cov}(X_{1B}, X_{2A})\\
=&\mathsf{E}(X_{1B}X_{2A}) - \mathsf{E}(X_{1B})\mathsf{E}(X_{2A})\\
=&\mathsf{E}(X_{1A}X_{2B}) - \mathsf{E}(X_{1A})\mathsf{E}(X_{2B}) = \mathsf{Cov}(X_{1A}, X_{2B}),
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{E}(X_{1B}X_{2B})\\
=&\sum_{(u,v)\in C_0} \frac{1}{m_u^2 m_v^2} \sum_{\substack{i,k\in\mathcal{C}_u\\ j,l\in\mathcal{C}_v}} \mathsf{P}(g_i = g_j = 1, g_k = g_l = 2)\\
&+ \sum_{(u,v),(u,w)\in C_0} \frac{1}{m_u^2 m_v m_w} \sum_{\substack{i,k\in\mathcal{C}_u\\ j\in\mathcal{C}_v, l\in\mathcal{C}_w}} \mathsf{P}(g_i = g_j = 1, g_k = g_l = 2)\\
&+ \sum_{(u,v),(w,h)\in C_0} \frac{1}{m_u m_v m_w m_h} \sum_{\substack{i\in\mathcal{C}_u, j\in\mathcal{C}_v\\ k\in\mathcal{C}_w, l\in\mathcal{C}_h}} \mathsf{P}(g_i = g_j = 1, g_k = g_l = 2)\\
=&\sum_{(u,v)\in C_0} \frac{1}{m_u^2 m_v^2} m_u(m_u-1)m_v(m_v-1)f_1\\
&+ \sum_{(u,v),(u,w)\in C_0} \frac{1}{m_u^2 m_v m_w} m_u(m_u-1)m_v m_w f_1\\
&+ \sum_{(u,v),(w,h)\in C_0} \frac{1}{m_u m_v m_w m_h} m_u m_v m_w m_h f_1\\
=&-\sum_{u=1}^{K} \frac{|\mathcal{E}_u^{C_0}|^2}{m_u} f_1 + |C_0|^2 f_1 + \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} f_1,
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{Cov}(X_{1B}, X_{2B})\\
=&\mathsf{E}(X_{1B}X_{2B}) - \mathsf{E}(X_{1B})\mathsf{E}(X_{2B})\\
=&\left(\sum_{(u,v)\in C_0} \frac{1}{m_u m_v} - \sum_{u=1}^{K} \frac{|\mathcal{E}_u^{C_0}|^2}{m_u}\right) f_1 + (f_1 - p_1 q_1)|C_0|^2.
\end{aligned}
$$

Plugging these equations into (S1.3), we immediately get the result.

□

Lemma 2. *The means, variances and covariance of $R_{1,(u)}$ and $R_{2,(u)}$ under the permutation null are*

$$\mathsf{E}(R_{1,(u)}) = |\bar{G}|p_1,$$

$$\mathsf{E}(R_{2,(u)}) = |\bar{G}|q_1,$$

$$\mathsf{Var}(R_{1,(u)}) = (p_1 - p_3)|\bar{G}| + (p_2 - p_3)\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1) + (p_3 - p_1^2)|\bar{G}|^2,$$

$$\mathsf{Var}(R_{2,(u)}) = (q_1 - q_3)|\bar{G}| + (q_2 - q_3)\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1) + (q_3 - q_1^2)|\bar{G}|^2,$$

$$\mathsf{Cov}(R_{1,(u)}, R_{2,(u)}) = f_1\left[|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1)\right] - p_1 q_1|\bar{G}|^2.$$

where $p_1, p_2, p_3, q_1, q_2, q_3, f_1$ are defined as those in Lemma 1.

Proof. With $p_1, p_2, p_3, q_1, q_2, q_3, f_1, |\bar{G}|$ and $\mathcal{E}_i^{\bar{G}}$ defined previously, we have

$$\mathsf{E}(R_{1,(u)}) = \sum_{(i,j)\in\bar{G}} \mathsf{P}(g_i = g_j = 1) = |\bar{G}|p_1,$$

$$\begin{aligned}
\mathsf{E}(R_{1,(u)}^2) &= \sum_{(i,j),(k,l)\in\bar{G}} \mathsf{P}(g_i = g_j = g_k = g_l = 1) \\
&= \sum_{(i,j)\in\bar{G}} \mathsf{P}(g_i = g_j = 1) + \sum_{\substack{(i,j),(i,k)\in\bar{G} \\ j\neq k}} \mathsf{P}(g_i = g_j = g_k = 1) \\
&\quad + \sum_{\substack{(i,j),(k,l)\in\bar{G} \\ i,j,k,l \text{ all different}}} \mathsf{P}(g_i = g_j = g_k = g_l = 1) \\
&= |\bar{G}|p_1 + \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1)p_2 + \left[|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}| \left(|\mathcal{E}_i^{\bar{G}}| - 1\right)\right] p_3 \\
&= (p_1 - p_3)|\bar{G}| + (p_2 - p_3)\sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}| \left(|\mathcal{E}_i^{\bar{G}}| - 1\right) + p_3|\bar{G}|^2, \\
\mathsf{Var}(R_{1,(u)}) &= (p_1 - p_3)|\bar{G}| + (p_2 - p_3)\sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}| \left(|\mathcal{E}_i^{\bar{G}}| - 1\right) + (p_3 - p_1^2)|\bar{G}|^2.
\end{aligned}$$

Similarly, we can get $\mathsf{E}(R_{2,(u)})$ and $\mathsf{Var}(R_{2,(u)})$.

Since

$$\begin{aligned}
\mathsf{E}(R_{1,(u)}R_{2,(u)}) &= \sum_{\substack{(i,j),(k,l)\in\bar{G} \\ i,j,k,l \text{ all different}}} \mathsf{P}(g_i = g_j = 1, g_k = g_l = 2) \\
&= \left[|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}| \left(|\mathcal{E}_i^{\bar{G}}| - 1\right)\right] f_1,
\end{aligned}$$

we have

$$\begin{aligned}
\mathsf{Cov}(R_{1,(u)}, R_{2,(u)}) &= \mathsf{E}(R_{1,(u)}R_{2,(u)}) - \mathsf{E}(R_{1,(u)})\mathsf{E}(R_{2,(u)}) \\
&= f_1 \left[|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1)\right] - p_1 q_1 |\bar{G}|^2.
\end{aligned}$$

□

S2 Issues of existing graph-based tests

S2.1 Problem of the graph-based tests for data with repeated observations

To illustrate this problem, we use a phone-call network dataset analyzed in both Chen and Friedman (2017) and Chen et al. (2018). This dataset has 330 networks, corresponding to 330 consecutive days, respectively. Each network represents the phone-call activity among the same group of people on a particular day (a more detailed description of this dataset see in Section 6). In both papers, the authors tested whether the distribution of phone-call networks on weekdays is the same as that on weekends. The distance between two networks is defined as the number of different edges between them. In this dataset, phone-call networks on some days are the same and the distance matrix on the distinct networks has ties. According to their results, the 9-MST was a good choice for the similarity graph. However, the 9-MST is not uniquely defined due to the repeated observations (networks) and the ties in the distance matrix. We randomly selected four such 9-MSTs and the results of the generalized edge-count test (S_G) and the weighted edge-count test (Z_w) under each of the 9-MSTs are listed in Table 1. We see that the test statistics based on different 9-MSTs vary a lot and the

p-values could be very small under some choices of 9-MSTs but very large under some other choices, leading to completely different conclusions (see Table 1 in the main context).

S2.2 Variance boosting problem for the extended edge-count test

To illustrate the problem, we use a preference ranking set up, where two groups of people are asked to rank six objects, and we test whether the two samples have the same preference over these six objects or not. Let Ξ be the set of all permutations of the set $\{1, 2, 3, 4, 5, 6\}$. We use the following probability model introduced by Mallows (1957) to generate data:

$$\mathsf{P}_{\theta,\eta}(\zeta) = \frac{1}{\psi(\theta)} \exp\{-\theta d(\zeta, \eta)\}, \quad \zeta, \eta \in \Xi, \quad \theta \in \mathbf{R},$$

where $d(\cdot, \cdot)$ is a distance function such as Kendall's or Spearman's distance and ψ is a normalizing constant. There are two parameters, θ and η, where η can be viewed as the "center" of the distribution and θ controls the "spread" of the distribution — the larger θ is, the less the distribution spreads. In the following, we let $d(\zeta, \eta)$ be the Spearman's distance between ζ and η and let C_0 be the 3-NNL on distinct values.

Let $\theta_1 = \theta_2 = 5$, $\eta_1 = \{1, 2, 3, 4, 5, 6\}$ and $\eta_2 = \{1, 2, 5, 4, 3, 6\}$ in the example. We check the performance under unbalanced sample sizes. The power of $R_{0,(a)}$ and $R_{0,(u)}$ are 0.804 and 0.832 respectively when $n_1 = n_2 = 80$. However, if we increase the sample size of Sample 2 to $n_2 = 400$ and keep all other parameters unchanged, the power of $R_{0,(a)}$ and $R_{0,(u)}$ decreases to 0.49 and 0.815, respectively (Table 1).

Table 1: The fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level in the preference ranking example. Here, $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = \theta_2 = 5$.

Power	$n_1 = n_2 = 80$	$n_1 = 80,\ n_2 = 400$
$R_{0,(a)}$	0.804	0.49
$R_{0,(u)}$	0.832	0.815

S3 Additional results

S3.1 Additional results in examining the extended test statistics

- S1 (Both η and θ differ with $\theta_1 > \theta_2$) :

  $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = 5.5$, $\theta_2 = 4$ with balanced ($n_1 = n_2 = 100$) and unbalance ($n_1 = 100, n_2 = 300$) sample sizes.

- S2 (Both η and θ differ with $\theta_1 < \theta_2$) :

  $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = 4$, $\theta_2 = 5.5$ with balanced ($n_1 = n_2 = 100$) and unbalance ($n_1 = 100, n_2 = 300$) sample sizes.

Table 2: S1: $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = 5.5$, $\theta_2 = 4$.

$n_1 = n_2 = 100$						
Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.865	0.775	0.865	0.846	0.824	0.792
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.915**	0.863	**0.915**	**0.895**	**0.886**	**0.872**
Statistic	LR	Pearson	Ker			
Estimated Power	0.222	0.221	0.227			

$n_1 = 100, n_2 = 300$						
Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.805	0.898	**0.952**	**0.936**	**0.927**	**0.916**
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.489	**0.961**	**0.980**	**0.975**	**0.970**	**0.964**
Statistic	LR	Pearson	Ker			
Estimated Power	0.187	0.181	0.363			

Table 3: S2: $\eta_1 = \{1, 2, 3, 4, 5, 6\}$, $\eta_2 = \{1, 2, 5, 4, 3, 6\}$, $\theta_1 = 4$, $\theta_2 = 5.5$.

$n_1 = n_2 = 100$						
Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.873**	0.783	**0.873**	**0.851**	0.837	0.812
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.891**	**0.858**	**0.891**	**0.888**	**0.872**	**0.864**
Statistic	LR	Pearson	Ker			
Estimated Power	0.217	0.218	0.233			

$n_1 = 100, n_2 = 300$						
Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.796	0.903	**0.956**	0.943	0.932	0.920
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.993**	**0.984**	**0.980**	**0.985**	**0.983**	**0.980**
Statistic	LR	Pearson	Ker			
Estimated Power	0.689	0.700	0.301			

S3.2 Choice of κ for max-type edge-count test statistics

We discuss the choice of κ by examining the test on 100-dimensional multivariate normal distributions $\mathcal{N}_d(\mu_1, \Sigma_1)$ and $\mathcal{N}_d(\mu_2, \Sigma_2)$ with mean and/or variance difference:

- Scenario 1, Only mean differs, $\|\mu_1 - \mu_2\|_2 = 1.5,\ \Sigma_1 = \Sigma_2 = \mathbf{I}$;
- Scenario 2, Only variance differs, $\mu_1 = \mu_2,\ \Sigma_1 = \mathbf{I},\ \Sigma_2 = 0.9\mathbf{I}$;
- Scenario 3, Only variance differs, $\mu_1 = \mu_2,\ \Sigma_1 = \mathbf{I},\ \Sigma_2 = 1.1\mathbf{I}$.

For each scenario, we examine both balanced setting $n_1 = n_2 = 80$ and unbalanced setting $n_1 = 80,\ n_2 = 150$.

Since the data is continuous, the optimal graph is uniquely determined (with probability 1). We compare the power of $M(\kappa)$ with the edge-count test (R_0), the generalized edge-count test (S_G) and the weighted edge-count test ($R_{w,G}$) to have a better understanding of the max-type statistic. Figures 1–3 plot the estimate power of the tests based on 1000 trials under each scenario. We see that $M(\kappa)(\kappa = \{1.31, 1.14, 1\})$ always perform well under various scenarios.

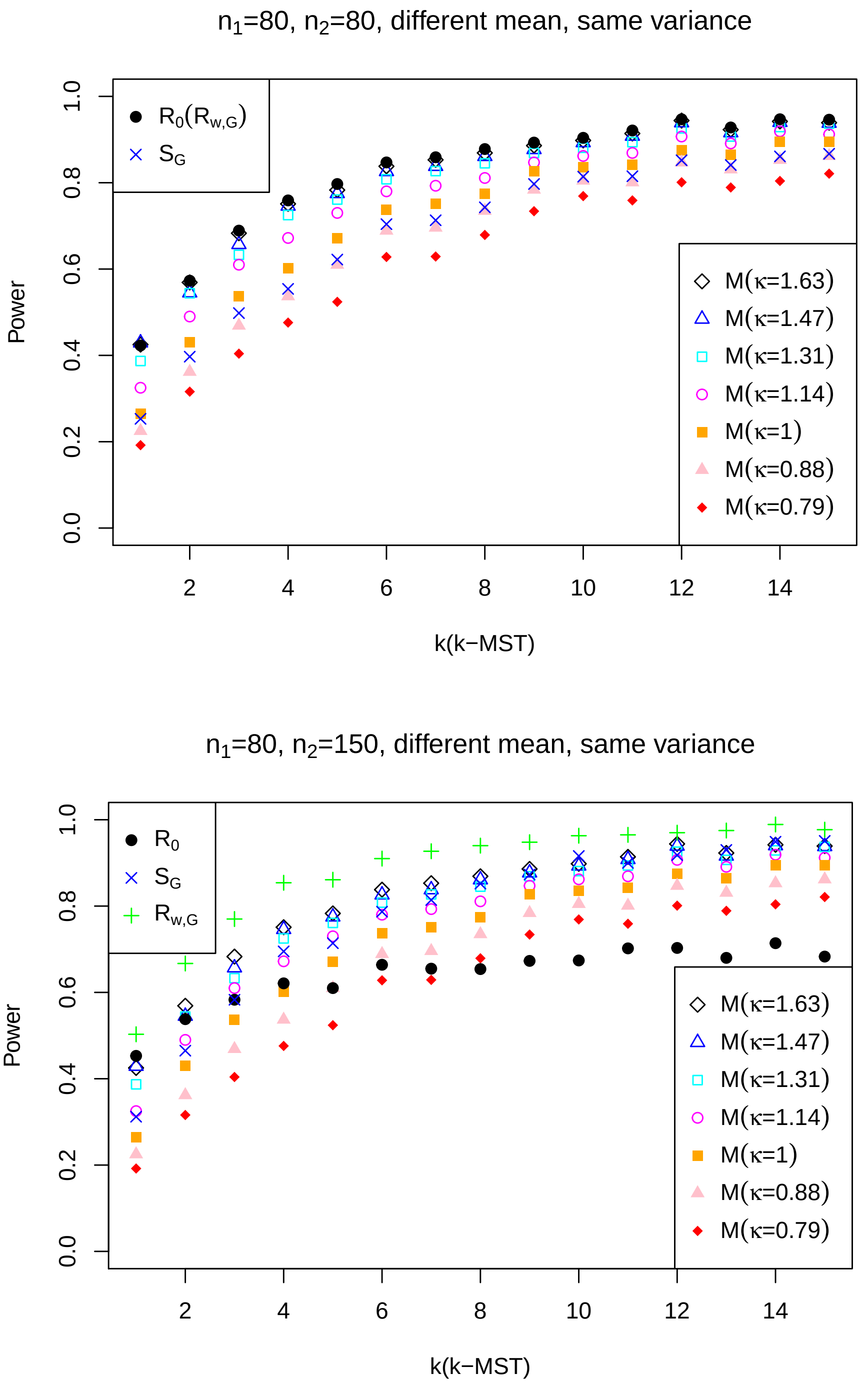


Figure 1: Under Scenario 1, the fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level.

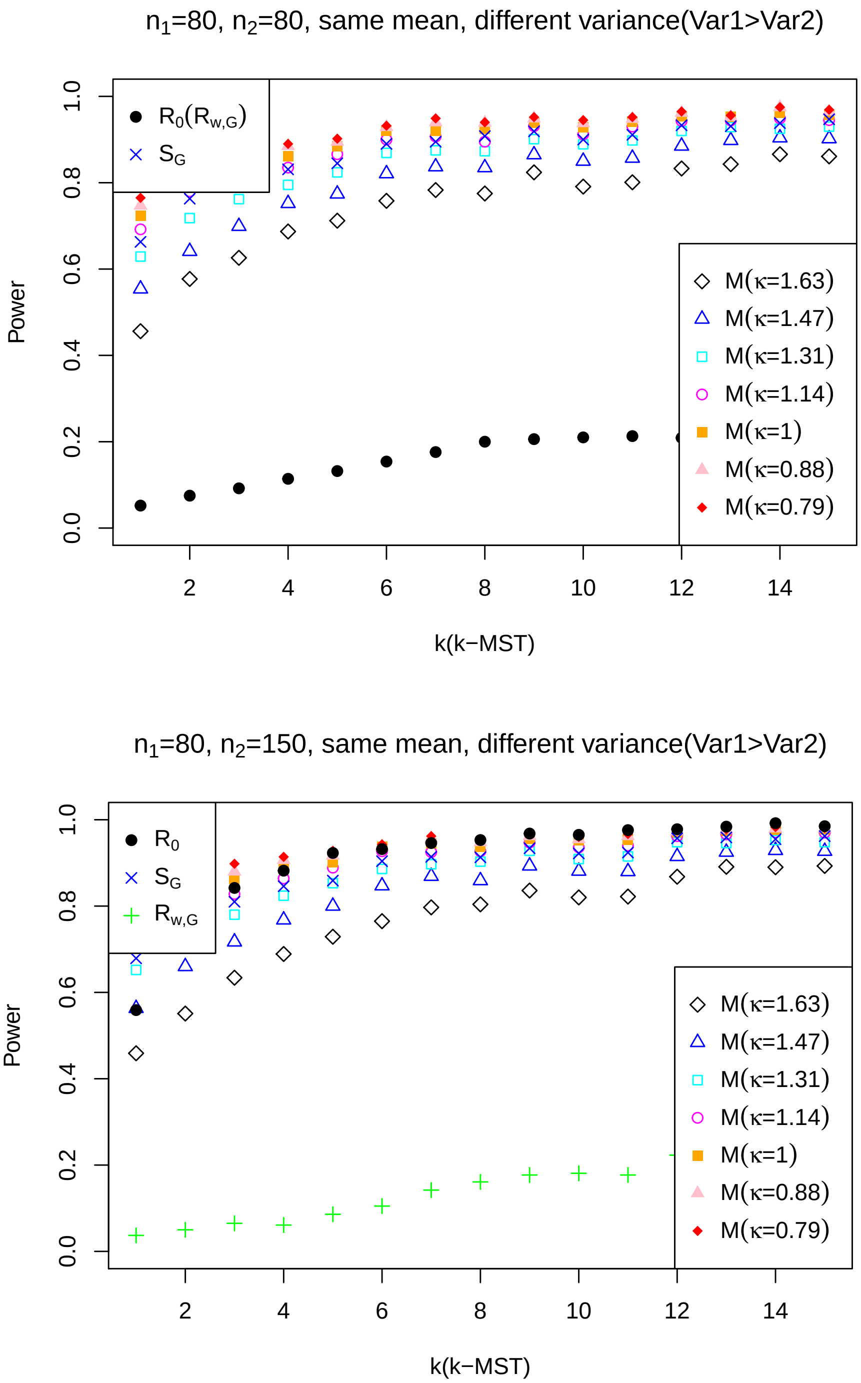
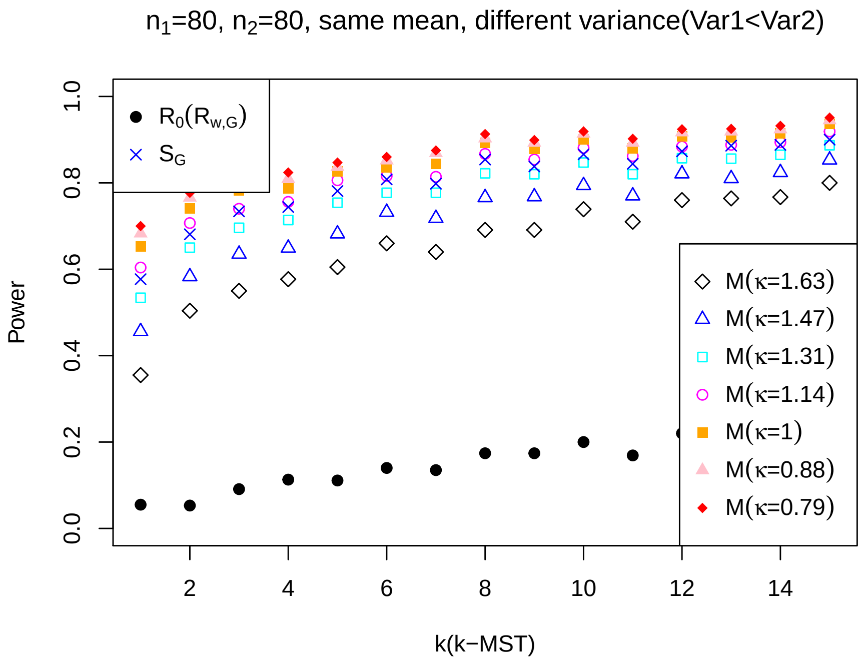


Figure 2: Under Scenario 2, the fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level.



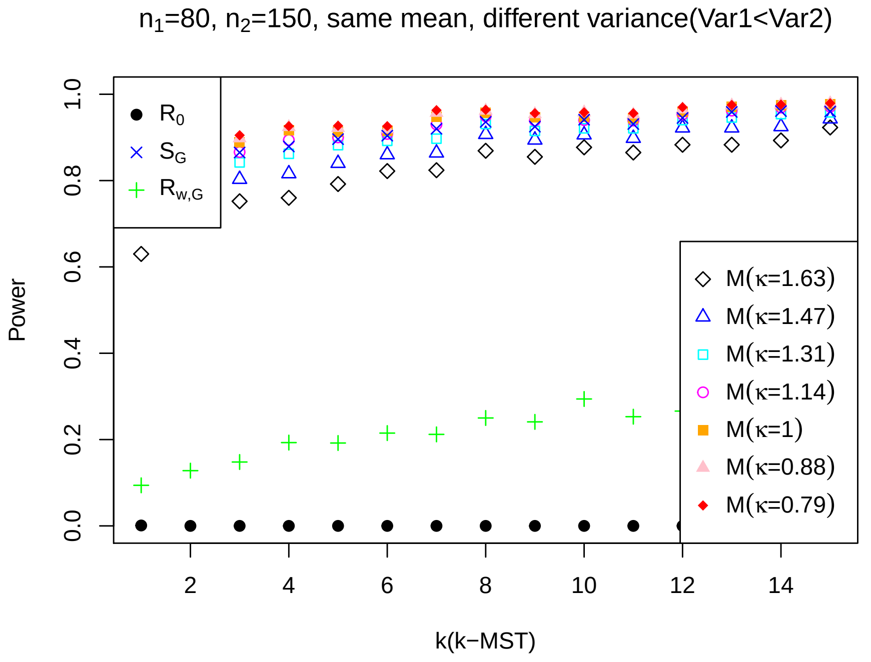


Figure 3: Under Scenario 3, the fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level.

S3.3 Analytic p-value approximations

The asymptotic results in Sections 5.1 and 5.2 provide theoretical bases for analytic p-value approximations. Here we check how well the analytic p-value approximations based on asymptotic results work under finite samples by comparing them with permutation p-values calculated from 10,000 random permutations.

Preference ranking

In the following, we generate data from mechanism (i) in Section 4 with $\theta_1 = \theta_2 = 5$, $\eta_1 = \{1, 2, 3, 4, 5\}$ and $\eta_2 = \{1, 4, 3, 2, 5\}$. We set C_0 be the NNL and examine the difference of the asymptotic p-value and permutation p-value under various settings.

Figures 4–6 show boxplots for the differences of the two p-values (asymptotic p-value minus permutation p-value) with different choices of n_1 and n_2 for $S_{(a)}, S_{(u)}, R_{w,(a)}, R_{w,(u)}, M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$. We see that when both n_1 and n_2 are over 100, the asymptotic p-value is very close to the permutation p-value for all new test statistics.

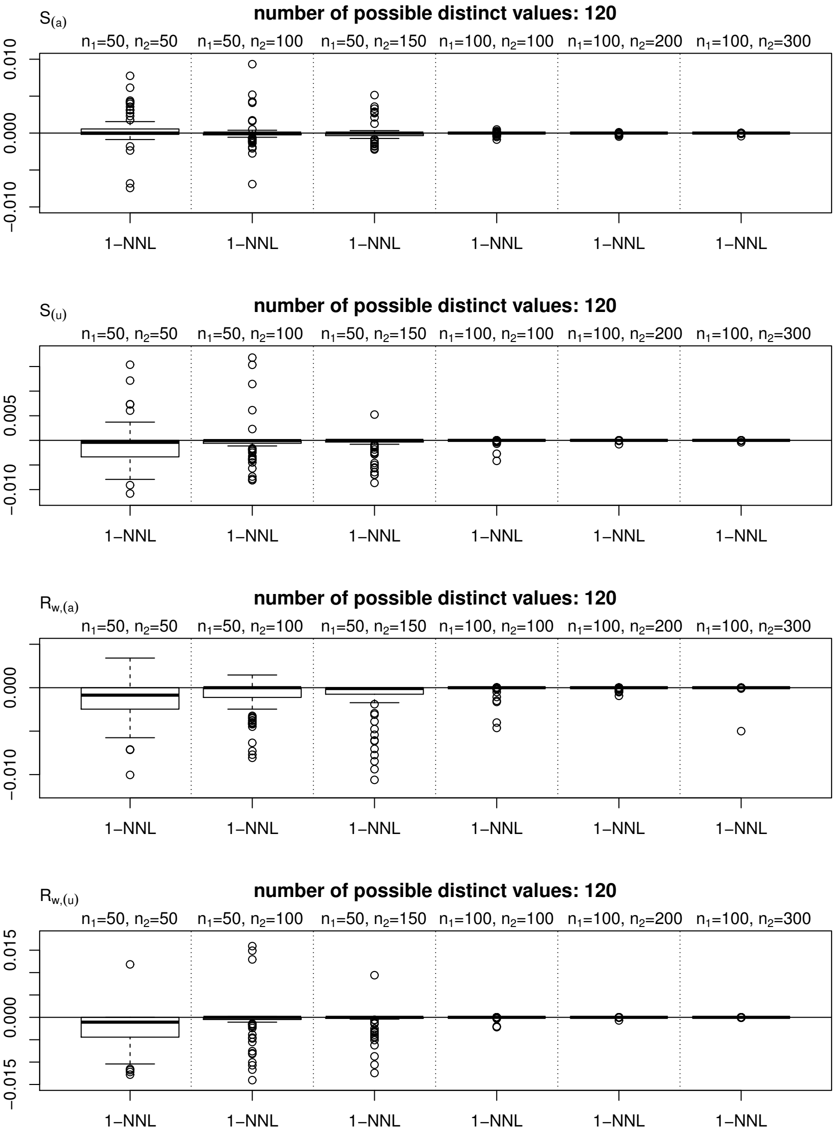


Figure 4: Boxplots for the differences between the asymptotic p-value and the permutation p-value based on 100 simulation runs under each setting for $S_{(a)}, S_{(u)}, R_{w,(a)}$ and $R_{w,(u)}$.

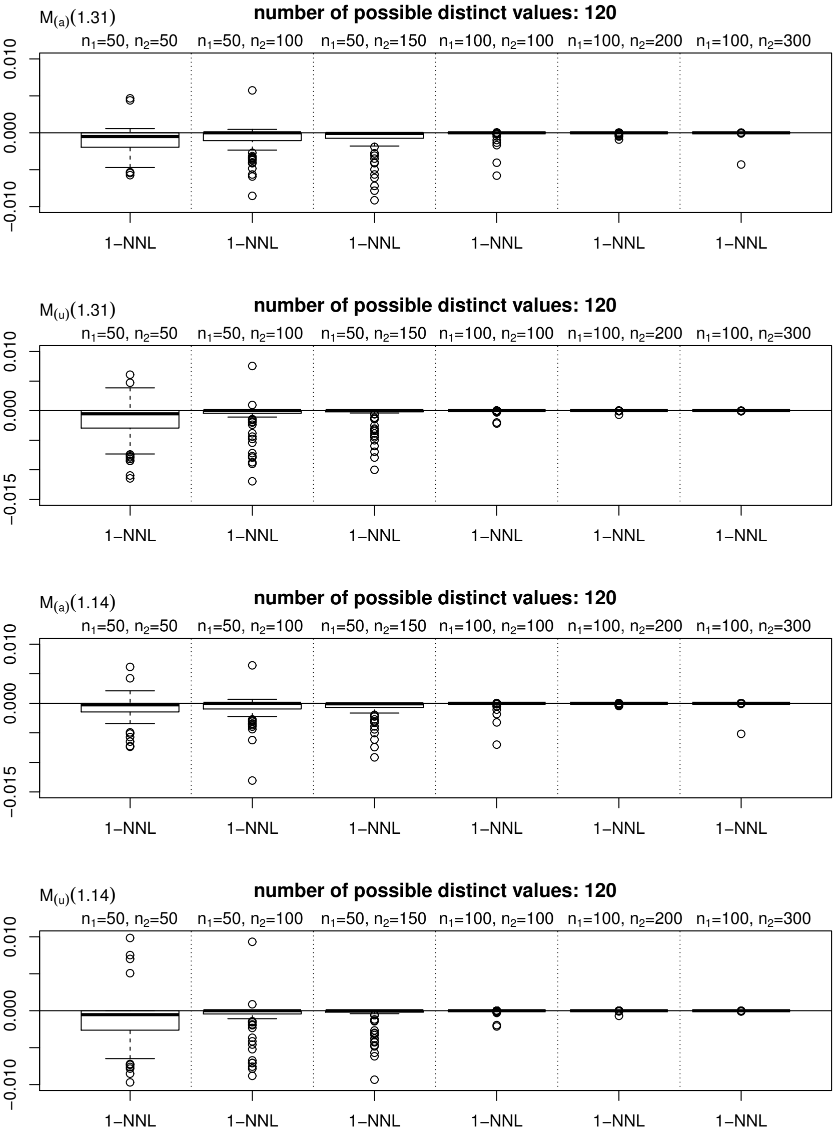


Figure 5: Boxplots for the differences between the asymptotic p-value and the permutation p-value based on 100 simulation runs under each setting for $M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$ with $\kappa = 1.31, 1.14$.

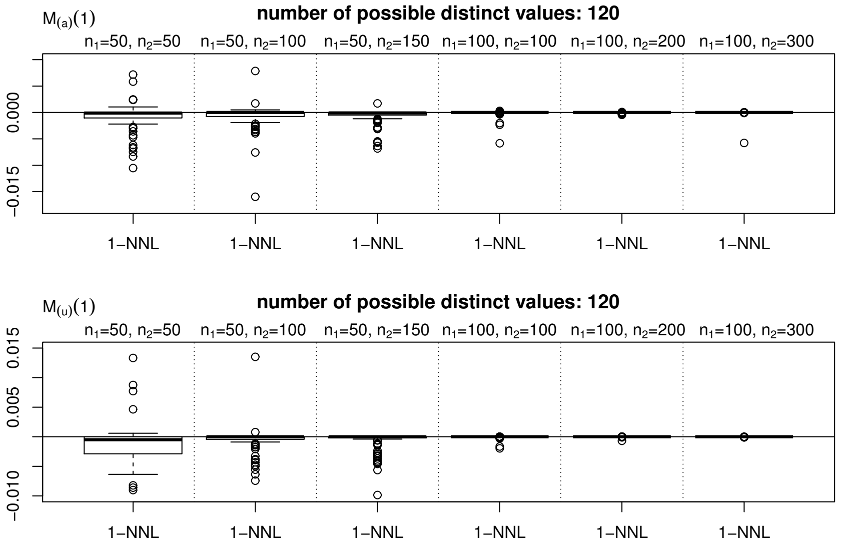


Figure 6: Boxplots for the differences between the asymptotic p-value and the permutation p-value based on 100 simulation runs under each setting for $M_{(a)}(1)$ and $M_{(u)}(1)$.

Phone-call network data

Table 4: The p-value obtained from the asymptotic results (Asym.) and from doing 10,000 random permutations (Perm.) for different statistics.

p-value	Asym.	Perm.	p-value	Asym.	Perm.
$S_{(a)}$	0.040	0.042	$S_{(u)}$	0.082	0.086
$R_{w,(a)}$	0.007	0.013	$R_{w,(u)}$	0.017	0.024
$M_{(a)}(1.31)$	0.009	0.014	$M_{(u)}(1.31)$	0.022	0.026
$M_{(a)}(1.14)$	0.013	0.019	$M_{(u)}(1.14)$	0.032	0.034
$M_{(a)}(1)$	0.022	0.025	$M_{(u)}(1)$	0.050	0.049

We check the analytic p-values obtained based on asymptotic results with those based on 10,000 random permutations and the results are shown in Table 4. We can see that the asymptotic p-values and the permutation p-values are quite close for all test statistics.

Bibliography

Chen, H., X. Chen, and Y. Su (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association* **113**(523), 1146–1155.

Chen, H. and J. H. Friedman (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association* **112**(517), 397–409.

Chen, L. H. and Q.-M. Shao (2005). Steins method for normal approximation. *An Introduction to Steins Method* **4**, 1–59.

Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika* **44**(1/2), 114–130.