A ROBUST FRAMEWORK FOR GRAPH-BASED TWO-SAMPLE TESTS USING WEIGHTS

Iowa State University

Supplementary Material

In this Supplementary Material, we provide additional simulations, extra figures, and proofs for Lemma 1, Remark 1, Theorem 1, Theorem 2 and Theorem 3.

S1 Additional Simulations

S1.1 imulation results of SHP test and cross-match test

In this section, we explore the performance of the Shortest Hamiltonian path (SHP)-based test (Biswas et al., 2014) and the cross-match test based on non-bipartite matching (Rosenbaum, 2005) in the high-dimensional setting.

Observations are simulated under distributional changes. Specifically, the simulation settings are as follows:

- Mean change only. Observations are generated from multivariate normal distributions: $X \sim \mathcal{N}(1_d, I_d)$, $Y \sim \mathcal{N}(\sqrt{1.5log(d)/d}1_d, I_d)$, where d denotes the dimension. $n_1 = n_2 = 100$.

- Scale change only. Observations are generated from multivariate normal distributions: $X \sim \mathcal{N}(1_d, I_d)$, $Y \sim \mathcal{N}(1_d, (1+1.5log(d)/d)I_d)$, where d denotes the dimension. $n_1 = n_2 = 100$.

The SHP-based test and the cross-match test are designed using a similar rationale as the original graph-based test proposed by Friedman and Rafsky (Friedman and Rafsky, 1979). As such, these tests focus on the between-sample edge counts in the test statistic, which can encounter problems detecting general changes as the dimension d increases (Chen and Friedman, 2017). We compare their performances to the robust edge-count tests S_R and M_R (introduced in Section 3 in the paper). From Table 1, we can see the SHP-based test and cross-match test have reasonable power when $d = 500$ and $d = 800$ for mean change, but its power starts to decay as d increases. Under scale change, both have lower power than the robust edge-count tests; the cross-match test in particular seems to struggle in this setting. As d goes to 2000, both robust edge-count tests demonstrate superior power.

Table 1: Number of trials with significance less than 5% for comparison of robust graph-based test S_R, M_R, SHP-based test and cross-match test with mean change and scale change.

	mean change				scale change			
d	SHP	cross-match	S_R	M_R	SHP	cross-match	S_R	M_R
500	95	83	**100**	**100**	76	37	**100**	**100**
800	92	84	98	**100**	67	24	**99**	**99**
1100	77	67	95	**97**	55	20	**97**	95
1400	68	62	**93**	92	43	15	94	**97**
1700	66	57	91	**92**	35	16	**93**	92
2000	71	55	92	**96**	35	24	**88**	86

S1.2 Simulation results of robust edge-count tests under imbalanced sample sizes

We carry out simulations to demonstrate the performance of the tests under imbalanced sample sizes. The data are simulated using the same settings as those in Simulation III in Section 5:

$$\mathbf{X} \sim \exp(\mathcal{N}(\mathbf{1}_d, 0.6\mathbf{I}_d))$$

$$\mathbf{Y} \sim \exp(\mathcal{N}((1 + \sqrt{0.01 log(d)/d})\mathbf{1}_d, (0.6 + 1.8 log(d)/d)\mathbf{I}_d)),$$

where d denotes the dimension. We investigate two unbalanced settings with different sample sizes of the two samples. As shown in Table 2 and 3, the robust edge-count tests S_R and M_R still retain good performance across all imbalanced settings, and demonstrate improvement compared to the edge-count tests S and M. When the sample sizes are not too unbalanced (Table 2), most of the graph-based tests are on equal footing. However, when the imbalance between samples becomes more severe (Table 3), all tests have diminished power. We observe that the hubness phenomenon is not exacerbated by the imbalanced sample size - both settings have max node degrees of similar sizes (142 and 138, when $d = 2000$, respectively). However, hubness is still clearly a problem here, since the new proposed tests tend to have better (or comparable) power across all settings. When the sample sizes are severely unbalanced (Table 3), we see the new proposed robust tests are still performing quite well.

Table 2: Number of trials with significance less than 5%. $n_1 = 50$, $n_2 = 150$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	124	52	26	96	96	97	97	**99**	**99**
800	130	49	11	90	90	90	**97**	**97**	**97**
1100	132	35	8	81	78	80	86	92	**95**
1400	137	35	3	66	78	70	87	90	**91**
1700	138	36	2	69	77	69	**82**	80	**82**
2000	142	31	6	71	68	72	**83**	81	**83**

Table 3: Number of trials with significance less than 5%. $n_1 = 15$, $n_2 = 185$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	128	24	0	70	74	69	77	**80**	**80**
800	133	20	0	59	58	64	64	**68**	64
1100	132	18	0	53	52	51	54	54	**56**
1400	138	15	1	47	48	52	52	**56**	53
1700	140	18	0	39	41	38	41	**47**	46
2000	138	16	0	43	44	43	51	48	**53**

S2 Extra Figures

S2.1 Hubness phenomenon in high-dimensional data using 5-NN

The maximum and 95th percentile of node degrees in the similarity graph constructed using 5-NN are shown in Figure 1. The hubness phenomenon is similar to what we can see using the 5-MST as the similarity graph. The maximum node degrees are over three

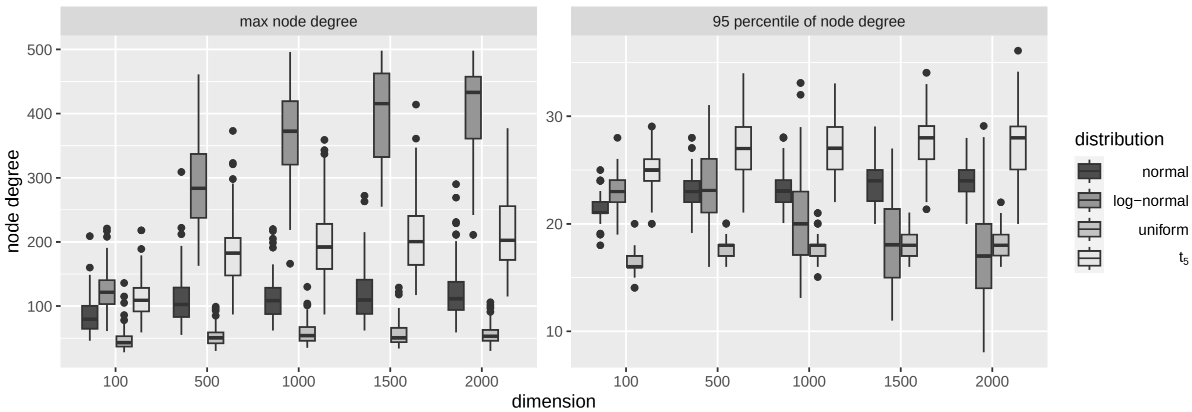

Figure 1: Boxplot of maximum and 95th percentiles of node degrees for different dimensions. Results are from 100 simulations with $n = 500$, where observations are drawn from multivariate normal, log-normal, uniform, and t distributions.

times as much as the 95th percentiles.

S3 Proof of Lemma 1

The mean and variance of R_1^w under the permutation null distribution can be derived as follows:

$$\mu_1^w = \sum_{(i,j)\in G} w_{ij} P(J_{(i,j)} = 1) = \sum_{(i,j)\in G} w_{ij} \frac{n_1(n_1-1)}{N(N-1)},$$

$$\begin{aligned}
E((R_1^w)^2) = &\sum_{(i,j),(k,l)\in G} w_{ij} w_{kl} P(J_{(i,j)} = 1, J_{(k,l)} = 1) \\
=& S_1 \frac{n_1(n_1-1)}{N(N-1)} + S_2' \frac{n_1(n_1-1)(n_1-2)}{N(N-1)(N-2)} + \\
& S_3' \frac{n_1(n_1-1)(n_1-2)(n_1-3)}{N(N-1)(N-2)(N-3)},
\end{aligned}$$

$$\Sigma_{11} = E((R_1^w)^2) - E^2(R_1^w),$$

where $S_1 = \sum_{(i,j)\in G} w_{ij}^2$, $S_2' = \sum_{\substack{(i,j),(i,k)\in G \\ k,l \text{ are different}}} w_{ij}w_{ik}$, and

$S_3' = \sum_{\substack{(i,j),(k,l)\in G \\ i,j,k,l \text{ all different}}} w_{ij}w_{kl}$.

Similarly, we can get the mean and variance of R_2^w under the permutation null distribution:

$$\begin{aligned}
\mu_2^w &= \sum_{(i,j)\in G} w_{ij} P(J_{(i,j)} = 2) = \sum_{(i,j)\in G} w_{ij} \frac{n_2(n_2-1)}{N(N-1)}, \\
E((R_2^w)^2) =& S_1 \frac{n_2(n_2-1)}{N(N-1)} + S_2' \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} + \\
& S_3' \frac{n_2(n_2-1)(n_2-2)(n_2-3)}{N(N-1)(N-2)(N-3)} \\
\Sigma_{22} =& E((R_2^w)^2) - E^2(R_2^w).
\end{aligned}$$

The covariance of R_1^w and R_2^w under the permutation null distribution can be derived as follows:

$$\begin{aligned}
E(R_1^w R_2^w) &= \sum_{(i,j),(k,l)\in G} w_{ij} w_{kl} P(J_{(i,j)} = 1, J_{(k,l)} = 2) \\
&= S_3' \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}, \\
\Sigma_{12} &= E(R_1^w R_2^w) - E(R_1^w)E(R_2^w).
\end{aligned}$$

Note :

$$\begin{aligned}
\sum_{\substack{(i,j),(k,l)\in G \\ i,j,k,l \text{ all different}}} w_{ij}w_{kl} &= \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl} - \sum_{\substack{(i,j),(i,k)\in G \\ k,l \text{ are different}}} w_{ij}w_{ik} - \sum_{(i,j)\in G} w_{ij}^2, \\
\sum_{\substack{(i,j),(i,k)\in G \\ k,l \text{ are different}}} w_{ij}w_{ik} &= \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik} - \sum_{(i,j)\in G} w_{ij}^2.
\end{aligned}$$

The variance and covariance can be simplified as

$$\begin{aligned}
\Sigma_{11} &= D_N \left\{ \frac{N-3}{n_2-1} S_1 + \frac{n_1-2}{n_2-1} S_2 + \frac{6(n_2-1)-4n_1(N-3)}{N(N-1)(n_2-1)} S_3 \right\} \\
&= D_N \left\{ -S_2 + \frac{2(2N-3)}{N(N-1)} S_3 + \frac{N-3}{n_2-1}(S_1+S_2) - \frac{4(N-3)}{N(n_2-1)} S_3 \right\}, \\
\Sigma_{12} &= D_N \left\{ -S_2 + \frac{2(2N-3)}{N(N-1)} S_3 \right\}, \\
\Sigma_{22} &= D_N \left\{ \frac{N-3}{n_1-1} S_1 + \frac{n_2-2}{n_1-1} S_2 + \frac{6(n_1-1)-4n_2(N-3)}{N(N-1)(n_1-1)} S_3 \right\} \\
&= D_N \left\{ -S_2 + \frac{2(2N-3)}{N(N-1)} S_3 + \frac{N-3}{n_1-1}(S_1+S_2) - \frac{4(N-3)}{N(n_1-1)} S_3 \right\},
\end{aligned}$$

where $S_1 = \sum_{(i,j)\in G} w_{ij}^2$, $S_2 = \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik}$, $S_3 = \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}$ and $D_N = [n_1n_2(n_1-1)(n_2-1)]/[N(N-1)(N-2)(N-3)]$..

S4 Proof of Remark 1

$$\begin{aligned}
\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik} =& \sum_{i=1}^{N} (\sum_{\{j,\text{s.t.}(i,j)\in G\}} w_{ij})^2 \\
\geq& \frac{1}{N} (\sum_{i=1}^{N} \sum_{\{j,\text{s.t.}(i,j)\in G\}} w_{ij})^2 \\
=& \frac{4}{N} (\sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}).
\end{aligned}$$

$$\begin{aligned}
\text{Var}(R_1^w - R_2^w) > 0 \Leftrightarrow& \sum_{\{j\in G_i\}} w_{ij} \text{ are not all equal for all } i \in [1, N], \\
\text{Var}(q_w R_1^w + p_w R_2^w) > 0 \Leftrightarrow& (N-3)S_1 - S_2 + \frac{2}{N-1} S_3 > 0.
\end{aligned}$$

S5 Proof of Theorem 1

Let $\mathbf{R} = \begin{pmatrix} R_1^w \\ R_2^w \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} 1 & -1 \\ q & p \end{pmatrix}$, $R_{\text{diff}}^w = R_1^w - R_2^w$ and $R_w^w = qR_1^w + pR_2^w$.

$$\begin{aligned} S &= (\mathbf{R} - E(\mathbf{R}))^T \mathbf{\Sigma}^{-1} (\mathbf{R} - E(\mathbf{R})) \\ &= (\mathbf{R} - E(\mathbf{R}))^T \mathbf{C}^T (\mathbf{C}^T)^{-1} \mathbf{\Sigma}^{-1} \mathbf{C}^{-1} \mathbf{C} (\mathbf{R} - E(\mathbf{R})) \\ &= (\mathbf{C}(\mathbf{R} - E(\mathbf{R})))^T (\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T)^{-1} (\mathbf{C}(\mathbf{R} - E(\mathbf{R}))), \end{aligned}$$

$$\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T = \mathbf{C} \begin{pmatrix} \text{Var}(R_1^w) & \text{Cov}(R_1^w, R_2^w) \\ \text{Cov}(R_1^w, R_2^w) & \text{Var}(R_2^w) \end{pmatrix} \mathbf{C}^T,$$

$$\mathbf{C}\mathbf{\Sigma}\mathbf{C}^T = \begin{pmatrix} \text{Var}(R_{\text{diff}}^w) & C_1 \\ C_1 & \text{Var}(R_w^w) \end{pmatrix},$$

where

$$\begin{aligned} \text{Var}(R_{\text{diff}}^w) &= \text{Var}(R_1^w) - 2\text{Cov}(R_1^w, R_2^w) + \text{Var}(R_2^w), \\ \text{Var}(R_w^w) &= q^2\text{Var}(R_1^w) + 2pq\text{Cov}(R_1^w, R_2^w) + p^2\text{Var}(R_2^w), \\ C_1 &= q\text{Var}(R_1^w) + (p - q)\text{Cov}(R_1^w, R_2^w) - p\text{Var}(R_2^w) \\ &= D_N \left\{ \frac{(N-3)(n_2-1)}{(N-2)(n_2-1)} \left(S_1 + S_2 - \frac{4}{N} S_3 \right) - \right. \\ &\quad \left. \frac{(N-3)(n_1-1)}{(N-2)(n_1-1)} \left(S_1 + S_2 - \frac{4}{N} S_3 \right) \right\} \\ &= 0. \end{aligned}$$

So $S_R = \frac{(R_{\text{diff}}^w - E(R_{\text{diff}}^w))^2}{\text{Var}(R_{\text{diff}}^w)} + \frac{(R_w^w - E(R_w^w))^2}{\text{Var}(R_w^w)}$, and the robust test statistic S_R can be decomposed as $S_R = (Z_{\text{diff}}^R)^2 + (Z_w^R)^2$ and $\mathbf{Cov}(Z_{\text{diff}}^R, Z_w^R) = 0$.

S6 Proof of Theorem 2

For $s = 1, 2$, $R_j^w = \sum_{(i,j)\in G} w_{ij} I_{J_{(i,j)}=s} > \min(w_{ij}) \sum_{(i,j)\in G} I_{J_{(i,j)}=s}$.

Then $\min(w_{ij})$ is asymptotically bounded below by $1/|G|$ and $\sum_{(i,j)\in G} I_{J_{(i,j)}=s} = O(|G|)$ since $\sum_{(i,j)\in G} I_{J_{(i,j)}=s}/N$ converge to a constant related to the densities of the two samples according to Theorem 2 in Henze and Penrose (1999).

So $\lim_{N\to\infty} \min(w_{ij}) \sum_{(i,j)\in G} I_{J_{(i,j)}=s} > 0$, $s = 1, 2$.

S7 Proof of Theorem 3

We will use the bootstrap null distribution to prove Theorem 3. Under the bootstrap null, the probability of an observation assigned to sample $\boldsymbol{X}$ is $\frac{n_X}{N}$, and the probability of an observation assigned to sample $\boldsymbol{Y}$ is $1 - \frac{n_X}{N}$. When $n_x = n_1$, the bootstrap null distribution is equivalent to the permutation null. We use subscripts to denote statistics under the bootstrap null.

First, we introduce Theorem 1 to help prove Theorem 3.

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

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

Under Assumption 1, we have $\sup_{h\in Lip(1)} |Eh(W) - Eh(Z)| \leq \delta$, *where* $Lip(1) = \{h : R \to R\}$, Z *has* $\mathcal{N}(0,1)$ *distribution and* $\delta = 2\sum_{i\in J}(E|\xi_i\eta_i\theta_i| + |E(\xi_i\eta_i)|E|\theta_i|) + \sum_{i\in J}|E|\xi_i\eta_i^2|$, *with* $\eta_i = \sum_{j\in K_i}\xi_j$ *and* $\theta_i = \sum_{j\in L_i}\xi_j$, *where* K_i *and* L_i *are defined in Assumption 1.*

Let $p_n = \frac{n_1}{N}$, $q_n = 1 - \frac{n_1}{N} = \frac{n_2}{N}$,

$$\mathrm{E}_B(R_1^w) = \sum_{(i,j)\in G} w_{ij} P(J_{(i,1)=1}) = \sum_{(i,j)\in G} w_{ij} p_n^2 := \mu_1^B,$$

$$\mathrm{E}_B(R_2^w) = \sum_{(i,j)\in G} w_{ij} P(J_{(i,1)=2}) = \sum_{(i,j)\in G} w_{ij} q_n^2 := \mu_2^B,$$

$$\begin{aligned}
\mathrm{Var}_B(R_1^w) = &\sum_{(i,j)\in G} w_{ij}^2 p_n^2 + \sum_{\substack{(i,j),(i,k)\in G\\ j\neq k}} w_{ij} w_{ik} p_n^3 + \\
&\sum_{\substack{(i,j),(k,l)\in G\\ i,j,k,l \text{ all different}}} w_{ij} w_{kl} p_n^4 - (\sum_{(i,j)\in G} w_{ij})^2 p_n^4 \\
= &\sum_{(i,j)\in G} w_{ij}^2 (p_n^2 - p_n^4) + \sum_{\substack{(i,j),(i,k)\in G\\ j\neq k}} w_{ij} w_{ik} (p_n^3 - p_n^4) \\
= &\sum_{(i,j)\in G} w_{ij}^2 (p_n^2 - p_n^4) + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} (p_n^3 - p_n^4) - \\
&\sum_{(i,j)\in G} w_{ij}^2 (p_n^3 - p_n^4) \\
= &\sum_{(i,j)\in G} w_{ij}^2 p_n^2 q_n + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} p_n^3 q_n \\
:= &(\sigma_1^B)^2.
\end{aligned}$$

Similarly,

$$\mathrm{Var}_B(R_2^w) = \sum_{(i,j)\in G} w_{ij}^2 q_n^2 p_n + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} q_n^3 p_n := (\sigma_2^B)^2,$$

$$\begin{aligned}
\mathrm{Cov}_B(R_1^w, R_2^w) &= \mathrm{E}_B(R_1^w R_2^w) - \mathrm{E}_B(R_1^w)\mathrm{E}_B(R_2^w) \\
&= \sum_{(i,j)\in G} w_{ij} \sum_{\substack{(k,l)\in G\\ i,j,k,l \text{ all different}}} w_{kl} p_n^2 q_n^2 - \sum_{(i,j)\in G} w_{ij} p_n^2 \sum_{(i,j)\in G} w_{ij} q_n^2 \\
&= -\sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} p_n^2 q_n^2 := (\sigma_{12}^B)^2.
\end{aligned}$$

Let $R^w_{\text{diff}} = R^w_1 - R^w_2$, we have

$$\mathrm{E}_B(R^w_{\text{diff}}) = \sum_{(i,j)\in G} w_{ij}(p_n - q_n) := \mu^B_{diff},$$

$$\begin{aligned}
\mathrm{Var}_B(R^w_{\text{diff}}) &= \mathrm{Var}_B(R^w_1) + \mathrm{Var}_B(R^w_2) - 2\mathrm{Cov}_B(R^w_1, R^w_2) \\
&= p_n q_n \sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik}(p_n^3 q_n + q_n^3 p_n + 2p_n^2 q_n^2) \\
&= p_n q_n \Big(\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik}\Big) \\
&:= (\sigma^B_{diff})^2.
\end{aligned}$$

Let $R^w_w = qR^w_1 + pR^w_2$, we have

$$\mathrm{E}_B(R^w_w) = \sum_{(i,j)\in G} w_{ij} \frac{n_2^2(n_1-1) + n_1^2(n_2-1)}{N^2(N-2)} := \mu^B_w,$$

$$\begin{aligned}
\mathrm{Var}_B(R^w_w) &= q^2\mathrm{Var}_B(R^w_1) + p^2\mathrm{Var}_B(R^w_2) + 2pq\mathrm{Cov}_B(R^w_1, R^w_2) \\
&= \frac{n_1 n_2 (n_1 - n_2)^2}{N^4(N-2)^2} \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} + \\
&\quad \frac{n_1 n_2\{n_1 n_2 (N-4) + N\}}{N^3(N-2)^2} \sum_{(i,j)\in G} w_{ij}^2 \\
&:= (\sigma^B_w)^2.
\end{aligned}$$

Let,

$$\begin{aligned}
W^B_1 &= \frac{R^w_w - E_B(R^w_w)}{\sqrt{\mathrm{Var}_B(R^w_w)}}, \\
W^B_2 &= \frac{R^w_{\text{diff}} - E_B(R^w_{\text{diff}})}{\sqrt{\mathrm{Var}_B(R^w_{\text{diff}})}}, \\
W^B_3 &= \frac{n_X - n}{\sqrt{Np_n(1-p_n)}}.
\end{aligned}$$

Lemma 1. *Under conditions*

(i) $|G| = \mathcal{O}(N^\alpha), 1 \leq \alpha < 1.25,$

(ii)
$$\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik} - \frac{4}{N}\sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}$$
$$=\mathcal{O}(\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik}),$$

(iii) $\sum_{(i,j)\in G}(w_{ij}|A_{(i,j)}|)^2 = o(\sum_{(i,j)\in G} w_{ij}^2 N^{0.5}),$

(iv) $\sum_{(i,j)\in G} w_{ij} \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} \sum_{(i'',j'')\in B_{(i,j)}} w_{i''j''} = o(\sum_{(i,j)\in G} w_{ij}^2)^{1.5},$

and under the bootstrap null, (W_1^B, W_2^B, W_3^B) *is multivariate normal.*

Lemma 2. *We have*

- $\frac{Var_B(R_w^w)}{Var(R_w^w)} \to c_1,$
- $\frac{Var_B(R_{diff}^w)}{Var(R_{diff}^w)} \to c_2,$
- $\frac{E_B(R_w^w)-E(R_w^w)}{\sqrt{Var(R_w^w)}} \to 0,$
- $\frac{E_B(R_{diff}^w)-E(R_{diff}^w)}{\sqrt{Var(R_{diff}^w)}} \to 0,$
- $\lim_{N\to\infty} Cov(Z_w, Z_{diff}) = 0,$

where c_1 *and* c_2 *are constant.*

From Lemma 1, $(W_1^B, W_2^B|W_3^B)$ is multivariate normal under the bootstrap null. Since conditioning on $W_3^B = 0$, $(W_1^B, W_2^B|W_3^B = 0)$ and (W_1^B, W_2^B) under the permutation distribution have the same distribution, and

$$\begin{aligned}
Z_w^R &= \frac{\sqrt{\mathrm{Var}_B(R_w^w)}}{\sqrt{\mathrm{Var}(R_w^w)}}(W_1^B + \frac{\mathrm{E}_B(R_w^w) - \mathrm{E}(R_w^w)}{\mathrm{Var}_B(R_w^w)}),\\
Z_{\mathrm{diff}}^R &= \frac{\sqrt{\mathrm{Var}_B(R_{\mathrm{diff}}^w)}}{\sqrt{\mathrm{Var}(R_{\mathrm{diff}}^w)}}(W_2^B + \frac{\mathrm{E}_B(R_{\mathrm{diff}}^w) - \mathrm{E}(R_{\mathrm{diff}}^w)}{\mathrm{Var}_B(R_{\mathrm{diff}}^w)}),
\end{aligned}$$

with Lemma 2, we conclude that Z_w^R and Z_{diff}^R are Gaussian under the permutation distribution.

S7.1 Proof of Lemma 1

We first show (W_1^B, W_2^B, W_3^B) is multivariate Gaussian under the bootstrap null distribution, which is equivalent to showing that $W = a_1 W_1^B + a_2 W_2^B + a_3 W_3^B$ is asymptotically Gaussian distributed for each $(a_1, a_2, a_3) \in \mathbb{R}^3$ such that $\text{Var}_B(W) > 0$ by Cramer-Wold theorem.

Let the index set $J = \{(i,j) \in G\} \bigcup \{1, 2, \ldots, N\}$,

$$\begin{aligned}
\xi_{(i,j)} =& a_1 \left(\frac{w_{ij} \frac{m-1}{N-2} I(J_{(i,j)=1}) + w_{ij} \frac{n-1}{N-2} I(J_{(i,j)=2})}{\sigma_w^B} - \frac{w_{ij} \frac{n^2(m-1)+m^2(n-1)}{N^2(N-2)}}{\sigma_w^B} \right) + \\
& a_2 \frac{w_{ij} I(J_{(i,j)=1}) - w_{ij} I(J_{(i,j)=2}) - (w_{ij}(p_n - q_n))}{\sigma_{diff}^B}, \\
\xi_i =& a_3 \frac{I(g_i = 0) - p_n}{\sqrt{N p_n (1 - p_n)}}.
\end{aligned}$$

Let, $a = max(|a_1|, |a_2|, |a_3|)$, $\sigma = \min(\sigma_w^B, \sigma_{diff}^B)$, $\sigma_0 = \sqrt{N p_n (1 - p_n)}$. σ^2 is at least of order $\sum\limits_{(i,j) \in G} w_{ij}^2$, $\sigma_0 = O(N^{0.5})$. Then $|\xi_{(i,j)}| \leq \frac{2a}{w_{ij}\sigma}$, $|\xi_i| \leq \frac{a}{\sigma_0}$ and $W = \sum_{j \in J} \xi_j$.

For $(i,j) \in J$, let

$$\begin{aligned}
A_{(i,j)} &= \{(i,j)\} \cup \{(i',j') \in G : (i',j') \text{ and } (i,j) \text{ share a node}\}, \\
B_{(i,j)} &= A_{(i,j)} \cup \{(i'',j'') \in G : \exists (i',j') \in A_{(i,j)}, \\
&\qquad \text{s.t. } (i',j') \text{ and } (i'',j'') \text{ share a node}\}, \\
K_{(i,j)} &= A_{(i,j)} \cup \{i, j\}, \\
L_{(i,j)} &= B_{(i,j)} \cup \{\text{nodes in } A_{(i,j)}\}.
\end{aligned}$$

For $i \in \{1, 2, ..., N\}$, let

$$
\begin{aligned}
&G_i = \{(i, j) \in G\},\\
&G_{i,2} = \{(i, j) \in G\} \cup \{(i'', j'') \in G : \exists (i', j') \in G_i,\\
&\qquad\qquad\qquad \text{s.t. } (i', j') \text{ and } (i'', j'') \text{ share a node}\},\\
&K_i = G_i \cup \{i\},\\
&L_j = G_{i,2} \cup \{\text{nodes in } G_i\}.
\end{aligned}
$$

For $j \in J$, let $\eta_j = \sum_{k \in K_j} \xi_k$ and $\theta_j = \sum_{k \in L_j} \xi_k$.

$$sup_{h \in Lip(1)} |E_B h(W) - Eh(Z)| \le \delta \text{ for } Z \sim N(0, 1),$$

where $\delta = \frac{1}{\sqrt{\mathrm{Var}_B(W)}} \big(2 \sum_{j \in J} (E_B |\xi_j \eta_j \theta_j| + E_B(\xi_j \eta_j) E_B |\theta_j|) + \sum_{j \in J} E_B |\xi_j \eta_j^2| \big)$, according to Theorem 1. For $j \in \{1, 2, ..., N\}$,

$$
\begin{aligned}
\eta_j &= \sum_{k \in K_j} \xi_k = \xi_i + \sum_{(i',j') \in G_i} \xi_{(i',j')} \le \frac{a}{\sigma_0} + \frac{2a}{\sigma} \sum_{(i',j') \in G_i} w_{i'j'},\\
\theta_j &= \sum_{k \in L_j} \xi_k = \sum_{\text{nodes in } G_i} \xi_i + \sum_{(i',j') \in G_{i,2}} \xi_{(i',j')} \le 2\frac{a|G_i|}{\sigma_0} + \frac{2a}{\sigma} \sum_{(i',j') \in G_{i,2}} w_{i'j'}.
\end{aligned}
$$

So,

$$
\begin{aligned}
&2 \sum_{j \in \{1,2,...,N\}} (E_B |\xi_j \eta_j \theta_j| + E_B(\xi_j \eta_j) E_B |\theta_j|) + \sum_{j \in \{1,2,...,N\}} E_B |\xi_j \eta_j^2|\\
&\le 5 \frac{a^3}{\sigma_0} \Big(\frac{1}{\sigma_0} + \frac{2}{\sigma} \sum_{(i',j') \in G_i} w_{i'j'}\Big)\Big(2\frac{|G_i|}{\sigma_0} + \frac{2}{\sigma} \sum_{(i',j') \in G_{i,2}} w_{i'j'}\Big).
\end{aligned}
$$

For $(i,j) \in G$,

$$\begin{aligned}
\eta_{(i,j)} &= \sum_{k \in K_{(i,j)}} \xi_k = \xi_i + \xi_j + \sum_{(i',j') \in A_{(i,j)}} \xi_{(i',j')} \\
&\le \frac{2a}{\sigma_0} + \frac{2a}{\sigma} \sum_{(i',j') \in A_{(i,j)}} w_{i'j'}, \\
\theta_{(i,j)} &= \sum_{k \in L_{(i,j)}} \xi_k = \sum_{\text{nodes in } A_{(i,j)}} \xi_i + \sum_{(i',j') \in B_{(i,j)}} \xi_{(i',j')} \\
&\le 2\frac{a|A_{(i,j)}|}{\sigma_0} + \frac{2a}{\sigma} \sum_{(i',j') \in B_{(i,j)}} w_{i'j'}.
\end{aligned}$$

So,

$$\begin{aligned}
&2 \sum_{(i,j) \in G} \left(E_B|\xi_{(i,j)}\eta_{(i,j)}\theta_{(i,j)}| + E_B(\xi_{(i,j)}\eta_{(i,j)})E_B|\theta_{(i,j)}| \right) \\
&+ \sum_{(i,j) \in G} E_B|\xi_{(i,j)}\eta^2_{(i,j)}| \\
\le\, & 5\frac{2aw_{ij}}{\sigma}\left(\frac{2a}{\sigma_0} + \frac{2a}{\sigma} \sum_{(i',j') \in A_{(i,j)}} w_{i'j'}\right)\left(2\frac{a|A_{(i,j)}|}{\sigma_0} + \frac{2a}{\sigma} \sum_{(i',j') \in B_{(i,j)}} w_{i'j'}\right) \\
=\, & 40\frac{a^3 w_{ij}}{\sigma}\left(\frac{1}{\sigma_0} + \frac{1}{\sigma} \sum_{(i',j') \in A_{(i,j)}} w_{i'j'}\right)\left(\frac{|A_{(i,j)}|}{\sigma_0} + \frac{1}{\sigma} \sum_{(i',j') \in B_{(i,j)}} w_{i'j'}\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
\delta \le & \Big[\sum_{(i,j) \in G} 40\frac{a^3 w_{ij}}{\sigma}\left(\frac{1}{\sigma_0} + \frac{1}{\sigma} \sum_{(i',j') \in A_{(i,j)}} w_{i'j'}\right)\left(\frac{|A_{(i,j)}|}{\sigma_0} + \frac{1}{\sigma} \sum_{(i',j') \in B_{(i,j)}} w_{i'j'}\right) + \\
& \sum_{i=1}^{N} 5\frac{a^3}{\sigma_0}\left(\frac{1}{\sigma_0} + \frac{2}{\sigma} \sum_{(i',j') \in G_i} w_{i'j'}\right)\left(2\frac{|G_i|}{\sigma_0} + \frac{2}{\sigma} \sum_{(i',j') \in G_{i,2}} w_{i'j'}\right)\Big] \frac{1}{\sqrt{\mathrm{Var}_B(W)}}.
\end{aligned}$$

If we want $\delta \to 0$ as $N \to \infty$, we need the following conditions to hold:

(1) $\displaystyle \sum_{i=1}^{N} \sum_{(i',j') \in G_i} w_{i'j'} \sum_{(i'',j'') \in G_{i,2}} w_{i''j''} = o\Big(\sum_{(i,j) \in G} w_{ij}^2 N^{0.5}\Big),$

(2) $\sum_{i=1}^{N} \sum_{(i',j') \in G_i} w_{i'j'} |G_i| = o\big((\sum_{(i,j) \in G} w_{ij}^2)^{0.5} N\big),$

(3) $\sum_{i=1}^{N} \sum_{(i',j')\in G_{i,2}} w_{i'j'} = o((\sum_{(i,j)\in G} w_{ij}^2)^{0.5} N)$,

(4) $\sum_{i=1}^{N} |G_i| = o(N^{1.5})$,

(5) $\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}| = o((\sum_{(i,j)\in G} w_{ij}^2)^{0.5} N)$,

(6) $\sum_{(i,j)\in G} w_{ij} \sum_{(i',j')\in B_{(i,j)}} w_{i'j'} = o(\sum_{(i,j)\in G} w_{ij}^2 N^{0.5})$,

(7) $$\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}| \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} = o(\sum_{(i,j)\in G} w_{ij}^2 N^{0.5}),$$

(8) $$\sum_{(i,j)\in G} w_{ij} \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} \sum_{(i'',j'')\in B_{(i,j)}} w_{i''j''} = o(\sum_{(i,j)\in G} w_{ij}^2)^{1.5}.$$

We need conditions:

(i) $|G| = \mathcal{O}(N^{\alpha}), 1 \le \alpha < 1.25$,

(ii) $\sum_{(i,j)\in G} (w_{ij}|A_{(i,j)}|)^2 = o(\sum_{(i,j)\in G} w_{ij}^2 N^{0.5})$,

(iii) $\sum_{(i,j)\in G} w_{ij} = o((\sum_{(i,j)\in G} w_{ij}^2)^{0.5} N)$,

(iv) $$\sum_{(i,j)\in G} w_{ij} \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} \sum_{(i'',j'')\in B_{(i,j)}} w_{i''j''} = o(\sum_{(i,j)\in G} w_{ij}^2)^{1.5},$$

(v) $$\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik} - \frac{4}{N} \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}$$
$$= \mathcal{O}(\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik}).$$

$\sum_{i=1}^{N} |G_i| = 2|G|$, so condition (4) holds according to (i). Since

$$\sum_{(i',j')\in A_{(i,j)}} w_{i'j'} \le |A_{(i,j)}| \max_{(i',j')\in A_{(i,j)}} w_{i'j'} =: |A_{(i,j)}| w_{\max} = |A_{(i,j)}| w_{ij} \frac{w_{\max}}{w_{ij}},$$

$\sum_{(i',j')\in A_{(i,j)}} w_{i'j'} = \mathcal{O}(|A_{(i,j)}|w_{ij})$, and condition (7) holds according to (ii). Besides,

$$\begin{aligned}
&\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}| \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} \\
=&\mathcal{O}(\sum_{(i,j)\in G} w_{ij}^2 |A_{(i,j)}|^2) \\
=&\mathcal{O}(\sum_{i=1}^{N} \sum_{(i',j')\in G_i} w_{i'j'}^2 |A_{(i',j')}|^2) \\
=&\mathcal{O}(\sum_{i=1}^{N} \sum_{(i',j')\in G_i} w_{i'j'}^2 |G_i|^2) \\
\le&\mathcal{O}(\sum_{(i,j)\in G} w_{ij}^2) N^{2\alpha-2} \\
=&o(\sum_{(i,j)\in G} w_{ij}^2 N^{0.5}).
\end{aligned}$$

So $2\alpha - 2 \le 0.5$, $\alpha \le 1.25$.

Let γ_{G_i} denotes the vertex set of $G_i/\{i\}$,

$$\begin{aligned}
\sum_{i=1}^{N} \sum_{(i',j')\in G_i} w_{i'j'} \sum_{(i'',j'')\in G_{i,2}} w_{i''j''} &\le \sum_{i=1}^{N} \sum_{(i',j')\in G_i} w_{i'j'} \sum_{j\in\gamma_{G_i}} \sum_{(i'',j'')\in G_j} w_{i''j''} \\
&= \sum_{i=1}^{N} \sum_{j\in\gamma_{G_i}} \sum_{(i',j')\in G_i} w_{i'j'} \sum_{(i'',j'')\in G_j} w_{i''j''} \\
&= 2 \sum_{(i,j)\in G} \sum_{(i',j')\in G_i} w_{i'j'} \sum_{(i'',j'')\in G_j} w_{i''j''} \\
&\le 2 \sum_{(i,j)\in G} (\sum_{(i',j')\in A_{(i,j)}} w_{i'j'})^2 \\
&= \mathcal{O}(\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}| \sum_{(i',j')\in A_{(i,j)}} w_{i'j'}).
\end{aligned}$$

So condition (7) implies condition (1).

By Cauchy-Schwarz inequality and (ii)

$$\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}| \leq \sqrt{\sum_{(i,j)\in G} w_{ij}^2 |A_{(i,j)}|^2 |G|}$$
$$= o((\sum_{(i,j)\in G} w_{ij}^2)^{0.5} N^{0.25})|G|^{0.5}.$$

So (i) ensures that condition (5) holds.

$$\sum_{(i,j)\in G} \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} = \mathcal{O}(\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}|)$$

$$\begin{aligned}\sum_{(i,j)\in G} \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} &= \sum_{(i,j)\in G} (\sum_{(i',j')\in G_i} w_{i'j'} + \sum_{(i'',j'')\in G_j} w_{i''j''} - w_{ij}) \\ &= \sum_{i=1}^{N} \sum_{j\in\gamma_{G_i}} \sum_{(i',j')\in G_j} w_{i'j'} - \sum_{(i,j)\in G} w_{ij} \\ &= \sum_{i=1}^{N} \sum_{(i',j')\in G_i} w_{i'j'}|G_i| - \sum_{(i,j)\in G} w_{ij}.\end{aligned}$$

According to conditions (5) and (iii), condition (2) holds.

$$G_{i,2} \subset \bigcup_{j\in\gamma_{G_i}} G_j,$$

$$\begin{aligned}\sum_{i=1}^{N} \sum_{(i',j')\in G_{i,2}} w_{i'j'} &\leq \sum_{i=1}^{N} \sum_{j\in\gamma_{G_i}} \sum_{(i',j')\in G_j} w_{i'j'} \\ &= \sum_{(i,j)\in G} (\sum_{(i',j')\in G_i} w_{i'j'} + \sum_{(i'',j'')\in G_j} w_{i''j''}) \\ &\leq 2 \sum_{(i,j)\in G} \sum_{(i',j')\in A_{(i,j)}} w_{i'j'} = \mathcal{O}(\sum_{(i,j)\in G} w_{ij}|A_{(i,j)}|).\end{aligned}$$

So condition (5) implies condition (3).

$$\sum_{(i,j)\in G} w_{ij} \sum_{(i',j')\in B_{(i,j)}} w_{i'j'} \leq \sum_{(i,j)\in G} w_{ij} \sum_{(i',j')\in A_{(i,j)}} \sum_{(i'',j'')\in A_{(i',j')}} w_{i''j''}$$
$$= \sum_{(i,j)\in G} w_{ij} (\sum_{(i',j')\in A_{(i,j)}} w_{i'j'})^2$$
$$= \mathcal{O}(\sum_{(i,j)\in G} w_{ij} |A_{(i,j)}| \sum_{(i',j')\in A_{(i,j)}} w_{i'j'}).$$

So condition (7) implies condition (6).

Finally, since $(\sum_{(i,j)\in G} w_{ij})^2 \leq |G| \sum_{(i,j)\in G} w_{ij}^2$,

$$\sum_{(i,j)\in G} w_{ij} = o(|G|^{0.5}(\sum_{(i,j)\in G} w_{ij}^2)^{0.5}) = o((\sum_{(i,j)\in G} w_{ij}^2)^{0.5} N),$$

if condition (i) is satisfied.

So we need conditions (i), (iii), (iv), (v).

S7.2 Proof of Lemma 2

$$\begin{aligned} &\mathrm{Var}_B(R_w^w) \\ =&\frac{n_1 n_2 (n_1 - n_2)^2}{N^4 (N-2)^2} \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} + \frac{n_1 n_2 \{n_1 n_2 (N-4) + N\}}{N^3 (N-2)^2} \sum_{(i,j)\in G} w_{ij}^2 \\ =&\mathcal{O}(\sum_{(i,j)\in G} w_{ij}^2), \end{aligned}$$

since

$$\sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} \leq (\sum_{(i,j)\in G} w_{ij})^2 = o(\sum_{(i,j)\in G} w_{ij}^2 N^2).$$

$$\begin{aligned}
\mathrm{Var}(R_w^w) = \frac{n_1 n_2 (n_1-1)(n_2-1)}{N(N-1)(N-2)(N-3)} \{ \sum_{(i,j)\in G} w_{ij}^2 - \\
\frac{1}{N-2} (\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} - \frac{4}{N} \sum_{(i,j),(k,l)\in G} w_{ij} w_{kl}) - \\
\frac{2}{N(N-1)} \sum_{(i,j),(k,l)\in G} w_{ij} w_{kl} \} \\
= \mathcal{O}(\sum_{(i,j)\in G} w_{ij}^2).
\end{aligned}$$

So, $\lim_{N\to\infty} \frac{\mathrm{Var}_B(R_w^w)}{\mathrm{Var}(R_w^w)} = c_1$, where c_1 is a constant.

$$\begin{aligned}
\lim_{N\to\infty} \frac{\mathrm{Var}_B(R_{\text{diff}}^w)}{\mathrm{Var}(R_{\text{diff}}^w)} &= \lim_{N\to\infty} \left(\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} \right) / \\
&\left(\sum_{(i,j)\in G} w_{ij}^2 + \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik} - \frac{4}{N} \sum_{(i,j),(k,l)\in G} w_{ij} w_{kl} \right) \\
&= c_2,
\end{aligned}$$

where c_2 is a constant, according to condition (v).

Since $E_B(R_w^w) - E(R_w^w) = \frac{n_1 n_2}{N^2(N-1)} \sum_{(i,j)\in G} w_{ij}$,

$$\lim_{N\to\infty} \frac{E_B(R_w^w) - E(R_w^w)}{\sqrt{\mathrm{Var}(R_w^w)}} = \lim_{N\to\infty} \frac{1}{N} \frac{\sum\limits_{(i,j)\in G} w_{ij}}{c_3 \sqrt{\sum\limits_{(i,j)\in G} w_{ij}^2}},$$

where c_3 is a constant.

From condition (iii) $\sum_{(i,j)\in G} w_{ij} = o((\sum_{(i,j)\in G} w_{ij}^2)^{0.5} N)$,

$$\lim_{N\to\infty} \frac{E_B(R_w^w) - E(R_w^w)}{\sqrt{\mathrm{Var}(R_w^w)}} = 0.$$

Since $E_B(R_{\text{diff}}^w) - E(R_{\text{diff}}^w) = 0$,

$$\lim_{N\to\infty} \frac{E_B(R_{\text{diff}}^w) - E(R_{\text{diff}}^w)}{\sqrt{\mathrm{Var}(R_{\text{diff}}^w)}} = 0.$$

We still need to show $\lim_{N\to\infty} \mathrm{Cov}(Z_w, Z_{\text{diff}}) = 0$.

$$\begin{aligned}
\mathrm{Cov}(Z_w, Z_{\text{diff}}) &= \frac{E(R_w^w R_{\text{diff}}^w) - E(R_w^w)E(R_{\text{diff}}^w)}{\sqrt{\mathrm{Var}(R_w^w)\mathrm{Var}(R_{\text{diff}}^w)}}, \\
E(R_w^w R_{\text{diff}}^w) &= S_3[q\frac{n_1^2(n_1-1)^2}{N^2(N-1)^2} - p\frac{n_2^2(n_2-1)^2}{N^2(N-1)^2} + \\
&\quad (p-q)\frac{n_1 n_2(n_1-1)(n_2-1)}{N^2(N-1)^2}] \\
&= \frac{(n_1-1)(n_2-1)(n_1-n_2)}{N(N-1)(N-2)}S_3, \\
E(R_w^w)E(R_{\text{diff}}^w) &= S_3[(\frac{n_1-n_2}{N})(\frac{n_1n_2-N+1}{(N-1)(N-2)})],
\end{aligned}$$

where $S_3 = \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}$.

$$\begin{aligned}
\lim_{N\to\infty} E(R_w^w R_{\text{diff}}^w) &= \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}p_nq_n(p_n-q_n), \\
\lim_{N\to\infty} E(R_w^w)E(R_{\text{diff}}^w) &= \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}p_nq_n(p_n-q_n).
\end{aligned}$$

So $\lim_{N\to\infty}(E(R_w^w R_{\text{diff}}^w) - E(R_w^w)E(R_{\text{diff}}^w)) = 0$.

Bibliography

Biswas, M., M. Mukhopadhyay, and A. K. Ghosh (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika 101*(4), 913–926.

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). Stein's method for normal approximation. *An introduction to Stein's method 4*, 1–59.

Friedman, J. H. and L. C. Rafsky (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics 7*(4), 697–717.

Henze, N. and M. D. Penrose (1999). On the multivariate runs test. *The Annals of Statistics 27*(1), 290–298.

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(4), 515–530.