Graph-Based Tests for Two-Sample Comparisons of Categorical Data

Hao Chen

Department of Statistics, Stanford University

Nancy R. Zhang

Department of Statistics, The Wharton School, University of Pennsylvania

Supplementary Material

S1 Proof of Theorem 1

Proof.

$$\begin{aligned} R_{\texttt{aMST}} &= |\mathcal{T}|^{-1} \sum_{\tau \in \mathcal{T}} R_\tau \\ &= |\mathcal{T}|^{-1} \sum_{\tau_0 \in \mathcal{T}_0} \sum_{\tau_1 \in \mathcal{T}_1} \cdots \sum_{\tau_K \in \mathcal{T}_K} [R_{\tau_0} + R_{\tau_1} + \cdots + R_{\tau_K}] \\ &= |\mathcal{T}_0|^{-1} \sum_{\tau_0 \in \mathcal{T}_0} R_{\tau_0} + \sum_{k=1}^{K} \left[\sum_{\tau_k \in \mathcal{T}_k} R_{\tau_k} / S_{m_k} \right]. \end{aligned} \tag{S1.1}$$

First consider the quantity $\sum_{\tau_k \in \mathcal{T}_k} R_{\tau_k}/S_{m_k}$. Since all pairs of subjects in a given category have the same distance ($= 0$), the edge between them should appear in the same number of trees. There are in total $m_k(m_k-1)/2$ possible pairs and each spanning tree for $\mathcal{C}_k$ has $m_k - 1$ edges. Hence, the edge between each pair of subjects in $\mathcal{C}_k$ appears in exactly

$$\frac{S_{m_k}(m_k-1)}{m_k(m_k-1)/2} = \frac{2S_{m_k}}{m_k}$$

trees. Thus,

$$\sum_{\tau_k \in \mathcal{T}_k} \frac{R_{\tau_k}}{S_{m_k}} = \sum_{i,j \in \mathcal{C}_k : i<j} I_{g_i \neq g_j} \frac{2S_{m_k}/m_k}{S_{m_k}} = \frac{2n_{ak}n_{bk}}{m_k}. \tag{S1.2}$$

Next consider the summation over $\mathcal{T}_0$. For any $i \in \mathcal{C}_u$, $j \in \mathcal{C}_v$, if $(u,v) \in \tau_0^*$, then the edge (i,j) appears in

$$\prod_{k=1}^{K} m_k^{|\mathcal{E}_k^{\tau_0^*}|}/(m_u m_v)$$

elements in $\mathcal{T}_0$, since any of the $m_u m_v$ possible edges connecting categories u and v appear in equal number of graphs in $\mathcal{T}_0$. Thus,

$$\begin{aligned}\sum_{\tau_0 \in \mathcal{T}_0} R_{\tau_0} &= \textstyle\sum_{\tau_0^* \in \mathcal{T}_0^*} \sum_{(u,v)\in\tau_0^*} \frac{\prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|}}{m_u m_v} \sum_{i\in\mathcal{C}_u}\sum_{j\in\mathcal{C}_v} I_{g_i\neq g_j} \\ &= \textstyle\sum_{\tau_0^* \in \mathcal{T}_0^*} \prod_{k=1}^K m_k^{|\mathcal{E}_k^{\tau_0^*}|} \sum_{(u,v)\in\tau_0^*} \frac{n_{au}n_{bv}+n_{av}n_{bu}}{m_u m_v}. \end{aligned} \tag{S1.3}$$

Combining (S1.1), (S1.2) and (S1.3) gives (7). □

S2 Proofs for Lemmas and Theorems in Permutation Distributions

S2.1 Proof of Lemma 1

Proof. Define

$$R_A = \sum_{u=1}^K \frac{1}{m_u} \sum_{i,j\in\mathcal{C}_u} I_{g_i\neq g_j},$$

and

$$R_B = \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} \sum_{i\in\mathcal{C}_u, j\in\mathcal{C}_v} I_{g_i\neq g_j}.$$

We have

$$\begin{aligned}\mathbf{E}_{\mathtt{P}}[R_{C_0}] &= \mathbf{E}_{\mathtt{P}}[R_A] + \mathbf{E}_{\mathtt{P}}[R_B] \\ &= \sum_{u=1}^K \frac{1}{m_u} \sum_{i,j\in\mathcal{C}_u} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j) + \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} \sum_{i\in\mathcal{C}_u, j\in\mathcal{C}_v} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j).\end{aligned}$$

Since $\mathbf{P}_{\mathtt{P}}(g_i \neq g_j) = \begin{cases} 0 & \text{if } i = j \\ \frac{2n_a n_b}{N(N-1)} & \text{if } i \neq j \end{cases}$, thus

$$\begin{aligned}\mathbf{E}_{\mathtt{P}}[R_{C_0}] &= \sum_{u=1}^K \frac{1}{m_u} m_u(m_u-1)\frac{2n_a n_b}{N(N-1)} + \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} m_u m_v \frac{2n_a n_b}{N(N-1)} \\ &= (N - K + |C_0|)\frac{2n_a n_b}{N(N-1)}.\end{aligned}$$

Now, to compute the second moment, first note that

$$\mathbf{E}_{\mathtt{P}}[R_{C_0}^2] = \mathbf{E}_{\mathtt{P}}[R_A^2] + \mathbf{E}_{\mathtt{P}}[R_B^2] + 2\mathbf{E}_{\mathtt{P}}[R_A R_B].$$

Expanding the right-hand-side in above,

$$\mathbf{E}_{\mathtt{P}}[R_A^2] = \sum_{u,v=1}^{k} \frac{1}{m_u m_v} \sum_{i,j\in\mathcal{C}_u,\ k,l\in\mathcal{C}_v} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l),$$

$$\begin{aligned}\mathbf{E}_{\mathtt{P}}[R_B^2] = &\sum_{(u,v)\in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k\in\mathcal{C}_u,\ j,l\in\mathcal{C}_v} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) \\ &+ 2 \sum_{\{(u,v),(w,y)\}\subset C_0} \frac{1}{m_u m_v m_w m_y} \sum_{i\in\mathcal{C}_u,\ j\in\mathcal{C}_v,\ k\in\mathcal{C}_w,\ l\in\mathcal{C}_y} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l),\end{aligned}$$

$$\mathbf{E}_{\mathtt{P}}[R_A R_B] = \sum_{u=1}^{K} \sum_{(v,w)\in C_0} \frac{1}{m_u m_v m_w} \sum_{i,j\in\mathcal{C}_u,\ k\in\mathcal{C}_v,\ l\in\mathcal{C}_w} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l).$$

Since

$$\mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) = \begin{cases} 0 & \text{if } i=j \text{ and/or } k=l \\ \frac{2n_a n_b}{N(N-1)} = 2p_1 & \text{if } \begin{cases} i=k, j=l, i\neq j \\ i=l, j=k, i\neq j \end{cases} \\ \frac{n_a n_b}{N(N-1)} = p_1 & \text{if } \begin{cases} i=k, j\neq i,l \\ i=l, j\neq i,k \\ j=k, i\neq j,l \\ j=l, i\neq j,k \end{cases} \\ \frac{4n_a(n_a-1)n_b(n_b-1)}{N(N-1)(N-2)(N-3)} = p_2 & \text{if } i,j,k,l \text{ are all different,} \end{cases}$$

we have

$$\begin{aligned}\mathbf{E}_{\mathtt{P}}[R_A^2] &= \sum_{u=1}^{K} \frac{1}{m_u^2} \sum_{i,j,k,l\in\mathcal{C}_u} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) + \sum_{u=1}^{k}\sum_{v\neq u} \frac{1}{m_u m_v} \sum_{i,j\in\mathcal{C}_u,\ k,l\in\mathcal{C}_v} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) \\ &= \sum_{u=1}^{K} \frac{1}{m_u^2}[2m_u(m_u-1)(2p_1) + 4m_u(m_u-1)(m_u-2)p_1 + m_u(m_u-1)(m_u-2)(m_u-3)p_2] \\ &\quad + \sum_{u=1}^{k}\sum_{v\neq u} \frac{1}{m_u m_v} m_u(m_u-1)m_v(m_v-1)p_2 \\ &= 4\left(N - 2K + \sum_{u=1}^{K}\frac{1}{m_u}\right)p_1 + (N-K-4)(N-K)p_2 + 6\left(K - \sum_{u=1}^{K}\frac{1}{m_u}\right)p_2,\end{aligned}$$

$$\begin{aligned}\mathbf{E}_{\mathtt{P}}[R_B^2] = &\sum_{(u,v)\in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k\in\mathcal{C}_u,\ j,l\in\mathcal{C}_v} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) \\ &+ \sum_{(u,v),(u,w)\in C_0,\ v\neq w} \frac{1}{m_u^2 m_v m_w} \sum_{i,k\in\mathcal{C}_u,\ j\in\mathcal{C}_v,\ l\in\mathcal{C}_w} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) \\ &+ \sum_{\substack{(u,v),(w,y)\in C_0 \\ u,v,w,y \text{ all different}}} \frac{1}{m_u m_v m_w m_y} \sum_{\substack{i\in\mathcal{C}_u,\ j\in\mathcal{C}_v \\ k\in\mathcal{C}_w,\ l\in\mathcal{C}_y}} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l)\end{aligned}$$

$$
\begin{aligned}
&= \sum_{(u,v)\in C_0} \frac{1}{m_u^2 m_v^2}[m_u m_v(2p_1) + m_u m_v(m_u+m_v-2)p_1 + m_u(m_u-1)m_v(m_v-1)p_2] \\
&\quad + \sum_{(u,v),(u,w)\in C_0,\ v\neq w} \frac{1}{m_u^2 m_v m_w}[m_u m_v m_w p_1 + m_u(m_u-1)m_v m_w p_2] \\
&\quad + \sum_{\substack{(u,v),(w,y)\in C_0 \\ u,v,w,y \text{ all different}}} \frac{1}{m_u m_v m_w m_y} m_u m_v m_w m_y p_2 \\
&= \sum_{(u,v)\in C_0} \frac{1}{m_u m_v}[(m_u+m_v)p_1 + (m_u-1)(m_v-1)p_2] \\
&\quad + \sum_{(u,v),(u,w)\in C_0,\ v\neq w} \frac{1}{m_u}[p_1 + (m_u-1)p_2] \\
&\quad + 2|\{(u,v),(w,y)\}\subset C_0 : u,v,w,y \text{ all different}|p_2 \\
&= \sum_{u=1}^{K} \frac{|\mathcal{E}_u^{C_0}|^2}{m_u}(p_1-p_2) + |C_0|^2 p_2 + \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} p_2,
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}_{\mathtt{P}}[R_A R_B] &= \sum_{u=1}^{K} \sum_{(u,v)\in \mathcal{E}_u^{C_0}} \frac{1}{m_u^2 m_v} \sum_{i,j,k\in\mathcal{C}_u,\ l\in\mathcal{C}_w} \mathbf{P}_{\mathtt{P}}(g_i\neq g_j, g_k\neq g_l) \\
&\quad + \sum_{u=1}^{K} \sum_{(v,w)\in C_0\setminus\mathcal{E}_u^{C_0}} \frac{1}{m_u m_v m_w} \sum_{i,j\in\mathcal{C}_u} \sum_{k\in\mathcal{C}_v, l\in\mathcal{C}_w} \mathbf{P}_{\mathtt{P}}(g_i\neq g_j, g_k\neq g_l) \\
&= \sum_{u=1}^{K} \sum_{(u,v)\in \mathcal{E}_u^{C_0}} \frac{1}{m_u^2 m_v}[2m_u(m_u-1)m_v p_1 + m_u(m_u-1)(m_u-2)m_v p_2] \\
&\quad + \sum_{u=1}^{K} \sum_{(v,w)\in C_0\setminus\mathcal{E}_u^{C_0}} \frac{1}{m_u m_v m_w} m_u(m_u-1)m_v m_w p_2 \\
&= |C_0|(N-K)p_2 + 2(p_1-p_2)\left(2|C_0| - \frac{|\mathcal{E}_u^{C_0}|}{m_u}\right).
\end{aligned}
$$

$\mathbf{Var}_{\mathtt{P}}[R_{C_0}]$ follows by combining the above in computing $\mathbf{E}_{\mathtt{P}}[R_{C_0}^2]$, and then subtracting $\mathbf{E}_{\mathtt{P}}^2[R_{C_0}]$. □

S2.2 Proof of Theorem 3

To prove Theorem 3, we first prove a simpler result: Asymptotic normality of the statistic under the bootstrap null, defined as the distribution obtained by sampling the group labels from the observed vector of group labels *with replacement*. Let $\mathbf{P}_{\mathtt{B}}$, $\mathbf{E}_{\mathtt{B}}$ and $\mathbf{Var}_{\mathtt{B}}$ denote respectively the probability, expectation and variance under the bootstrap null.

Lemma 1. *Assuming condition 1, under the bootstrap null distribution, the standardized statistic*

$$\frac{R_{C_0} - \mathbf{E}_B[R_{C_0}]}{\sqrt{\boldsymbol{Var}_B[R_{C_0}]}}$$

converges in distribution to $N(0,1)$ *as* $K \to \infty$*, where* $\mathbf{E}_B[R_{C_0}]$ *and* $\boldsymbol{Var}_B[R_{C_0}]$ *are given below.*

$$\mathbf{E}_B[R_{C_0}] = (N - K + |C_0|)2p_3, \tag{S2.4}$$

$$\begin{aligned}\boldsymbol{Var}_B[R_{C_0}] = 4(p_3 - p_4)\left(N - K + 2|C_0| + \sum_{u=1}^{K}\frac{|\mathcal{E}_u^{C_0}|^2}{4m_u} - \sum_{u=1}^{K}\frac{|\mathcal{E}_u^{C_0}|}{m_u}\right) \\ + (6p_4 - 4p_3)\left(K - \sum_{u=1}^{K}\frac{1}{m_u}\right) + p_4\sum_{(u,v)\in C_0}\frac{1}{m_u m_v},\end{aligned} \tag{S2.5}$$

where

$$p_3 = \frac{n_a n_b}{N^2}, \quad p_4 = \frac{4n_a^2 n_b^2}{N^4} = 4p_3^2. \tag{S2.6}$$

The proof of Lemma 1 relies on Stein's method. Consider sums of the form $W = \sum_{i\in\mathcal{J}}\xi_i$, where $\mathcal{J}$ is an index set and ξ are random variables with $E[\xi_i] = 0$, and $E[W^2] = 1$. The following assumption restricts the dependence between $\{\xi_i : i \in \mathcal{J}\}$.

Assumption 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 existing theorem.

Theorem 1. *[Chen and Shao, 2005, Theorem 3.4] Under Assumption 1, we have*

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

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

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

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 1.*

Proof of Lemma 1. The mean and variance of R_{C_0} under the bootstrap null, (S2.4) and (S2.5), can be obtained following similar steps as the proof of Lemma 1, noting that, under the bootstrap null,

$$\mathbf{P}_{\mathtt{B}}(g_i \neq g_j) = \begin{cases} 0 & \text{if } i = j \\ \frac{2n_a n_b}{N^2} = 2p_3 & \text{if } i \neq j \end{cases},$$

and

$$\mathbf{P}_{\mathtt{B}}(g_i \neq g_j, g_k \neq g_l) = \begin{cases} 0 & \text{if } i=j \text{ and/or } k=l \\ \frac{2n_a n_b}{N^2} = 2p_3 & \text{if } \begin{cases} i=k, j=l, i\neq j \\ i=l, j=k, i\neq j \end{cases} \\ \frac{n_a n_b}{N^2} = p_3 & \text{if } \begin{cases} i=k, j\neq i,l \\ i=l, j\neq i,k \\ j=k, i\neq j,l \\ j=l, i\neq j,k \end{cases} \\ \frac{4n_a^2 n_b^2}{N^4} = p_4 & \text{if } i,j,k,l \text{ are all different .} \end{cases}$$

To prove asymptotic normality, we first define more notations. For any node u of C_0, let

$$R_u = \frac{2n_{au}n_{bu}}{m_u}, \quad d_u = \mathbf{E}_{\mathtt{B}}[R_u] = 2(m_u - 1)p_3,$$

where p_3 is defined in (S2.6). Similarly, for any edge (u, v) of C_0, let

$$R_{uv} = \frac{n_{au}n_{bv} + n_{av}n_{bu}}{m_u m_v}, \quad d_{uv} = \mathbf{E}_{\mathtt{B}}[R_{uv}] = 2p_3.$$

Let $\sigma_{\mathtt{B}}^2 = \mathbf{Var}_{\mathtt{B}}[R_{C_0}]$, ξ_u, ξ_{uv} be the standardized mixing potentials,

$$\xi_u = \frac{R_u - d_u}{\sigma_{\mathtt{B}}}, \tag{S2.7}$$

$$\xi_{uv} = \frac{R_{uv} - d_{uv}}{\sigma_{\mathtt{B}}}. \tag{S2.8}$$

Finally, we define the index sets for ξ_u and ξ_{uv}:

$$\mathcal{J}_1 = \{1, \ldots, K\},$$

$$\mathcal{J}_2 = \{uv : u < v \text{ such that } (u, v) \in C_0\},$$

and let $\mathcal{J} = \mathcal{J}_1 \cup \mathcal{J}_2$. Since $R_{C_0} = \sum_{u=1}^K R_u + \sum_{(u,v)\in C_0} R_{uv}$, the standardized statistic is

$$W := \sum_{i\in\mathcal{J}} \xi_i = \sum_{u\in\mathcal{J}_1} \frac{R_u - d_u}{\sigma_{\mathtt{B}}} + \sum_{uv\in\mathcal{J}_2} \frac{R_{uv} - d_{uv}}{\sigma_{\mathtt{B}}} = \frac{R_{C_0} - \mathbf{E}_{\mathtt{B}}[R_{C_0}]}{\sigma_{\mathtt{B}}}.$$

Our notation follows those of Theorem 1 and Assumption 1. For $u \in \mathcal{J}_1$, let

$$S_u = \{u\} \cup \{uv, vu : (u, v) \in C_0\},$$
$$T_u = S_u \cup \{v, vw, wv : (u, v), (v, w) \in C_0\}.$$

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

$$S_{uv} = \{uv, u, v\} \cup \{uw, wu : (u, w) \in C_0\} \cup \{vw, wv : (v, w) \in C_0\},$$
$$T_{uv} = S_{uv} \cup \{w, wy, yw : (u, w), (w, y) \in C_0\} \cup \{w, wy, yw : (v, w), (w, y) \in C_0\}.$$

S_u, T_u, S_{uv}, T_{uv} defined in this way satisfy Assumption 1.

Since $R_u \in [0, \frac{m_u}{2}]$, $p_3 \in [0, \frac{1}{4}]$, and $R_{uv} \in [0,1]$, we have $d_u \in [0, \frac{m_u-1}{2}]$, $d_{uv} \in [0, \frac{1}{2}]$, and therefore $|\xi_u| \le \frac{m_u}{2\sigma_{\mathtt{B}}}$, $|\xi_{uv}| \le \frac{1}{\sigma_{\mathtt{B}}}$. Hence,

$$\begin{aligned}
&\sum_{j\in S_u} |\xi_j| \le \frac{1}{\sigma_{\mathtt{B}}}(m_u + |\mathcal{E}_u^{C_0}|), \quad u \in \mathcal{J}_1,\\
&\sum_{j\in T_u} |\xi_j| \le \frac{1}{\sigma_{\mathtt{B}}}(m_u + \sum_{v\in\mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|), \quad u \in \mathcal{J}_1,\\
&\sum_{j\in S_{uv}} |\xi_j| \le \frac{1}{\sigma_{\mathtt{B}}}(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|), \quad uv \in \mathcal{J}_2,\\
&\sum_{j\in T_{uv}} |\xi_j| \le \frac{1}{\sigma_{\mathtt{B}}}(m_u + m_v + \sum_{w\in\mathcal{V}_u\cup\mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|), \quad uv \in \mathcal{J}_2.
\end{aligned}$$

As in Theorem 1, let $\eta_i = \sum_{j\in S_i}\xi_j$ and $\theta_i = \sum_{j\in T_i}\xi_j$. Then

$$\begin{aligned}
\mathbf{E}_{\mathtt{B}}|\xi_i\eta_i\theta_i| = \mathbf{E}_{\mathtt{B}}|\xi_i \sum_{j\in S_i}\xi_j \sum_{k\in T_i}\xi_k| &\le \mathbf{E}_{\mathtt{B}}|\xi_i| \sum_{j\in S_i}|\xi_j| \sum_{k\in T_i}|\xi_k|,\\
|\mathbf{E}_{\mathtt{B}}(\xi_i\eta_i)|\mathbf{E}_{\mathtt{B}}|\theta_i| \le \mathbf{E}_{\mathtt{B}}|\xi_i \sum_{j\in S_i}\xi_j|\mathbf{E}_{\mathtt{B}}| \sum_{j\in T_i}\xi_j| &\le \mathbf{E}_{\mathtt{B}}|\xi_i| \sum_{j\in S_i}|\xi_j|\mathbf{E}_{\mathtt{B}} \sum_{j\in T_i}|\xi_j|,\\
\mathbf{E}_{\mathtt{B}}|\xi_i\eta_i^2| = \mathbf{E}_{\mathtt{B}}|\xi_i \sum_{j\in S_i}\sum_{k\in S_i}\xi_j\xi_k| &\le \mathbf{E}_{\mathtt{B}}|\xi_i| \sum_{j\in S_i}|\xi_j| \sum_{k\in S_i}|\xi_k|.
\end{aligned}$$

Thus, for $i = u \in \mathcal{J}_1$, the terms $\mathbf{E}_{\mathtt{B}}|\xi_i\eta_i\theta_i|$, $|\mathbf{E}_{\mathtt{B}}(\xi_i\eta_i)|\mathbf{E}_{\mathtt{B}}|\theta_i|$, and $\mathbf{E}_{\mathtt{B}}|\xi_i\eta_i^2|$ are all bounded by

$$\frac{1}{\sigma_{\mathtt{B}}^3} m_u(m_u + |\mathcal{E}_u^{C_0}|)(m_u + \sum_{v\in\mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|),$$

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

$$\frac{1}{\sigma_{\mathtt{B}}^3}(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v|)(m_u + m_v + \sum_{w\in\mathcal{V}_u\cup\mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|).$$

Hence,

$$\begin{aligned}
\delta \le \frac{5}{\sigma_{\mathtt{B}}^3}\Bigg(&\sum_{u=1}^{K} m_u(m_u + |\mathcal{E}_u^{C_0}|)(m_u + \sum_{v\in\mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|)\\
&+ \sum_{(u,v)\in C_0}(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|)(m_u + m_v + \sum_{w\in\mathcal{V}_u\cup\mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|)\Bigg).
\end{aligned}$$

Since $\sigma_{\mathtt{B}}$ is of order $\sqrt{K}$ or higher, under condition 1, $\delta \to 0$ as $K \to \infty$.

□

Proof of Theorem 3. To show the asymptotic normality of the standardized statistic under the permutation null, we only need to show that (R_{C_0}, n_a^B) converges to a bivariate Gaussian distribution under the bootstrap null, where n_a^B is the number of observations that belong to group a in the bootstrap sample. Then asymptotic normality of R_{C_0} under the permutation null follows from the fact that its distribution is equal to the conditional distribution of R_{C_0} given $n_a^B = n_a$. The standardized bivariate vector is

$$\left(\frac{R_{C_0} - \mathbf{E}_{\mathtt{B}}[R_{C_0}]}{\sqrt{\mathbf{Var}_{\mathtt{B}}[R_{C_0}]}}, \frac{n_a^B - Np_a}{\sigma_0} \right)$$

with $p_a = n_a/N, \sigma_0^2 = Np_a(1-p_a)$. By the Cramér-Wold device, we only need to show that

$$a_1 \frac{R_{C_0} - \mathbf{E}_{\mathtt{B}}[R_{C_0}]}{\sqrt{\mathbf{Var}_{\mathtt{B}}[R_{C_0}]}} + a_2 \frac{n_a^B - Np_a}{\sigma_0}$$

is asymptotic Gaussian under the bootstrap null for all $a_1, a_2 \in \mathbb{R}, a_1 a_2 \neq 0$.

Let $\xi_i, i \in \mathcal{J}$ be defined in the same way as in the proof of Lemma 1. Let $\mathcal{J}_3 = \{|\mathcal{J}|+1, \ldots, |\mathcal{J}|+K\}$. For $i \in \mathcal{J}_3$, let

$$\xi_i = \frac{n_{ai'} - p_a m_{i'}}{\sigma_0}, \quad i' = i - |\mathcal{J}|.$$

We use Theorem 1 to show the asymptotic Gaussianity of $\sum_{i\in\mathcal{J}} a_1\xi_i + \sum_{i\in\mathcal{J}_3} a_2\xi_i$. We need to redefine the neighborhood sets to satisfy Assumption 1.

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

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

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

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

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

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

From the proof of Lemma 1, we have

$$|\xi_u| \leq \frac{m_u}{2\sigma_{\mathtt{B}}}, \quad \forall u \in \mathcal{J}_1; \quad |\xi_{uv}| \leq \frac{1}{\sigma_{\mathtt{B}}}, \quad \forall uv \in \mathcal{J}_2.$$

For $u \in \mathcal{J}_3$,

$$|\xi_u| \le \frac{m_{u'}}{\sigma_0}, \quad u' = u - |\mathcal{J}|.$$

Let $\sigma = \min(\sigma_{\mathtt{B}}, \sigma_0)$, then

$$\begin{aligned}
&\sum_{j \in S_u} |\xi_j| \le \frac{1}{\sigma}(2m_u + |\mathcal{E}_u^{C_0}|), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,\\
&\sum_{j \in T_u} |\xi_j| \le \frac{1}{\sigma}(2m_u + 2\sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|), \quad u \in \mathcal{J}_1 \cup \mathcal{J}_3,\\
&\sum_{j \in S_{uv}} |\xi_j| \le \frac{1}{\sigma}(2m_u + 2m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|), \quad uv \in \mathcal{J}_2,\\
&\sum_{j \in T_{uv}} |\xi_j| \le \frac{1}{\sigma}(2m_u + 2m_v + 2\sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|), \quad uv \in \mathcal{J}_2.
\end{aligned}$$

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

$$\frac{1}{\sigma^3} m_u (2m_u + |\mathcal{E}_u^{C_0}|)(2m_u + 2\sum_{v \in \mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|),$$

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

$$\frac{1}{\sigma^3}(2m_u + 2m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v|)(2m_u + 2m_v + 2\sum_{w \in \mathcal{V}_u \cup \mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|).$$

Define $W_{a_1,a_2} = \sum_{i \in \mathcal{J}} a_1\xi_i + \sum_{i \in \mathcal{J}_3} a_2\xi_i$. The value of δ in Theorem 1 has the form

$$\begin{aligned}
\delta = \frac{1}{\sqrt{\mathbf{E}_{\mathtt{B}}[W_{a_1,a_2}^2]}} \Bigg(& 2\sum_{i \in \mathcal{J}} (\mathbf{E}_{\mathtt{B}}|a_1\xi_i\eta_i\theta_i| + |\mathbf{E}_{\mathtt{B}}(a_1\xi_i\eta_i)|\mathbf{E}_{\mathtt{B}}|\theta_i|) + \sum_{i \in \mathcal{J}} \mathbf{E}_{\mathtt{B}}|a_1\xi_i\eta_i^2|\\
& + 2\sum_{i \in \mathcal{J}_3} (\mathbf{E}_{\mathtt{B}}|a_2\xi_i\eta_i\theta_i| + |\mathbf{E}_{\mathtt{B}}(a_2\xi_i\eta_i)|\mathbf{E}_{\mathtt{B}}|\theta_i|) + \sum_{i \in \mathcal{J}_3} \mathbf{E}_{\mathtt{B}}|a_2\xi_i\eta_i^2| \Bigg),
\end{aligned}$$

where $\eta_i = \sum_{j \in S_i} \xi_j(a_1 I_{j \in \mathcal{J}} + a_2 I_{j \in \mathcal{J}_3})$, and $\theta_i = \sum_{j \in T_i} \xi_j(a_1 I_{j \in \mathcal{J}} + a_2 I_{j \in \mathcal{J}_3})$.

Let $a = \max(|a_1|, |a_2|)$, we have

$$\begin{aligned}
\mathbf{E}_{\mathtt{B}}|a_1\xi_i\eta_i\theta_i|,\ \mathbf{E}_{\mathtt{B}}|a_2\xi_i\eta_i\theta_i| &\le a^3\mathbf{E}_{\mathtt{B}}|\xi_i\sum_{j\in S_i}\xi_j\sum_{k\in T_i}\xi_k| \\
&\le a^3\mathbf{E}_{\mathtt{B}}|\xi_i|\sum_{j\in S_i}|\xi_j|\sum_{k\in T_i}|\xi_k|, \\
|\mathbf{E}_{\mathtt{B}}(a_1\xi_i\eta_i)|\mathbf{E}_{\mathtt{B}}|\theta_i|,\ |\mathbf{E}_{\mathtt{B}}(a_2\xi_i\eta_i)|\mathbf{E}_{\mathtt{B}}|\theta_i| &\le a^3\mathbf{E}_{\mathtt{B}}|\xi_i\sum_{j\in S_i}\xi_j|\mathbf{E}_{\mathtt{B}}|\sum_{j\in T_i}\xi_j| \\
&\le a^3\mathbf{E}_{\mathtt{B}}|\xi_i|\sum_{j\in S_i}|\xi_j|\mathbf{E}_{\mathtt{B}}\sum_{j\in T_i}|\xi_j|, \\
\mathbf{E}_{\mathtt{B}}|a_1\xi_i\eta_i^2|,\ \mathbf{E}_{\mathtt{B}}|a_2\xi_i\eta_i^2| &\le a^3\mathbf{E}_{\mathtt{B}}|\xi_i\sum_{j\in S_i}\sum_{k\in S_i}\xi_j\xi_k| \\
&\le a^3\mathbf{E}_{\mathtt{B}}|\xi_i|\sum_{j\in S_i}|\xi_j|\sum_{k\in S_i}|\xi_k|.
\end{aligned}$$

Thus,

$$\begin{aligned}
\delta \le \frac{40a^3}{\sigma^3\sqrt{\mathbf{E}_{\mathtt{B}}[W_{a_1,a_2}^2]}}\Bigg(&\sum_{u=1}^{K} m_u(m_u + |\mathcal{E}_u^{C_0}|)(m_u + \sum_{v\in\mathcal{V}_u} m_v + |\mathcal{E}_{u,2}^{C_0}|) \\
&+ \sum_{(u,v)\in C_0}(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|)(m_u + m_v + \sum_{w\in\mathcal{V}_u\cup\mathcal{V}_v} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|)\Bigg).
\end{aligned}$$

Since $\sigma_{\mathtt{B}}^2$ is at least of order K and σ_0^2 is of order N, σ^2 is at least of order K by Condition 2. If $\mathbf{E}_{\mathtt{B}}[W_{a_1,a_2}^2]$ is uniformly strictly bounded from 0 for any $a_1a_2 \neq 0$, then under Condition 1, $\delta \to 0$ as $K \to \infty$.

We next show that under Condition 2, $\mathbf{E}_{\mathtt{B}}[W_{a_1,a_2}^2]$ is uniformly strictly bounded from 0 for any $a_1a_2 \neq 0$.

Let $W_1 = \sum_{i\in\mathcal{J}}\xi_i, W_2 = \sum_{i\in\mathcal{J}_3}\xi_i$, then

$$\begin{aligned}
\mathbf{E}_{\mathtt{B}}[W_{a_1,a_2}^2] &= a_1^2\mathbf{E}_{\mathtt{B}}W_1^2 + a_2^2\mathbf{E}_{\mathtt{B}}W_2^2 + 2a_1a_2\mathbf{E}_{\mathtt{B}}[W_1W_2] \\
&= a_1^2 + a_2^2 + 2a_1a_2\mathbf{E}_{\mathtt{B}}[W_1W_2]
\end{aligned}$$

Thus, we only need to show that the absolute correlation between W_1 and W_2 is uniformly strictly bounded from 1. Notice that, in the theorem, we require n_a/N to be bounded from 0 and 1, so p_a and p_b are both bounded from 0 and 1.

Correlation between R_{C_0} and n_a^B: Observe that

$$R_{C_0} n_a^B = \left[\sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u} I_{g_i \neq g_j} + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} I_{g_i \neq g_j} \right] \sum_{x=1}^{N} I_{g_x = a}$$
$$= \sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in \mathcal{C}_u} \left(I_{g_i \neq g_j} \sum_{x=1}^{N} I_{g_x = a} \right) + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in \mathcal{C}_u, j \in \mathcal{C}_v} \left(I_{g_i \neq g_j} \sum_{x=1}^{N} I_{g_x = a} \right).$$

For any $i \neq j$,

$$\mathbf{E}_{\mathtt{B}} \left[I_{g_i \neq g_j} \sum_{x=1}^{N} I_{g_x = a} \right] = \mathbf{E}_{\mathtt{B}} \left[I_{g_i \neq g_j, g_i = a} + I_{g_i \neq g_j, g_j = a} + \sum_{x \neq i,j} I_{g_i \neq g_j, g_x = a} \right]$$
$$= \mathbf{P}_{\mathtt{B}}(g_i = a, g_j = b) + \mathbf{P}_{\mathtt{B}}(g_i = b, g_j = a) + \sum_{x \neq i,j} \mathbf{P}_{\mathtt{B}}(g_i \neq g_j, g_x = a)$$
$$= p_a p_b + p_a p_b + 2 p_a p_b p_a (N-2) = 2 p_a p_b (N p_a + 1 - 2 p_a).$$

Hence

$$\mathbf{E}_{\mathtt{B}}[R_{C_0} n_a^B] = (N - K + |C_0|) 2 p_a p_b (N p_a + 1 - 2 p_a).$$

Since $\mathbf{E}_{\mathtt{B}}[R_{C_0}] = (N - K + |C_0|) 2 p_a p_b$ and $\mathbf{E}_{\mathtt{B}}[n_a^B] = N p_a$, we have

$$\mathbf{Cov}_{\mathtt{B}}(R_{C_0}, n_a^B) = (N - K + |\mathcal{E}_{C_0}|) 2 p_a p_b (1 - 2 p_a). \tag{S2.9}$$

If $p_a = 1/2$, then $\mathbf{Cov}_{\mathtt{B}}(R_{C_0}, n_a^B) = 0$. Since $\mathbf{Var}_{\mathtt{B}}[R_{C_0}]$ and $\mathbf{Var}_{\mathtt{B}}[n_a^B] = N p_a p_b$ are positive, $\mathbf{Cor}_{\mathtt{B}}(R_{C_0}, n_a^B) = 0$, clearly bounded from 1. We consider $p_a \neq 1/2$ in the following.

$$\mathbf{Var}_{\mathtt{B}}[R_{C_0}] = 4 p_a p_b (1 - 4 p_a p_b) \left(N - K + 2|C_0| + \sum_{u=1}^{K} \frac{|\mathcal{E}_u^{C_0}|^2}{4 m_u} - \sum_{u=1}^{K} \frac{|\mathcal{E}_u^{C_0}|}{m_u} \right)$$
$$+ 4 p_a p_b (6 p_a p_b - 1) \left(K - \sum_{u=1}^{K} \frac{1}{m_u} \right) + 4 p_a^2 p_b^2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}$$
$$= 4 p_a p_b (1 - 4 p_a p_b) \left(N - 2K + 2|C_0| + \sum_{u=1}^{K} \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u} \right)$$
$$+ 8 p_a^2 p_b^2 \left(K - \sum_{u=1}^{K} \frac{1}{m_u} \right) + 4 p_a^2 p_b^2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}.$$

Since

$$N \sum_{u=1}^{K} \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u} = \sum_{u=1}^{K} m_u \sum_{u=1}^{K} \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u} \geq \left(\sum_{u=1}^{K} \sqrt{m_u \frac{(|\mathcal{E}_u^{C_0}|/2 - 1)^2}{m_u}} \right)^2$$
$$= \left(\sum_{u=1}^{K} ||\mathcal{E}_u^{C_0}|/2 - 1| \right)^2 \geq \left(\sum_{u=1}^{K} (|\mathcal{E}_u^{C_0}|/2 - 1) \right)^2 = (|C_0| - K)^2,$$

we have

$$\mathbf{Var}_{\mathtt{B}}[R_{C_0}]\mathbf{Var}_{\mathtt{B}}[n_a^B] \geq 4p_a^2p_b^2(1-4p_ap_b)[N-K+|C_0|]^2 + 4p_a^3p_b^3N\sum_{(u,v)\in C_0}\frac{1}{m_um_v}.$$

Hence,

$$|\mathbf{Cor}_{\mathtt{B}}(R_{C_0}, n_a^B)| \leq \frac{1}{\sqrt{1+\frac{p_ap_bN\sum_{(u,v)\in C_0}\frac{1}{m_um_v}}{(1-4p_ap_b)[N-K+|C_0|]^2}}}.$$

When $N, |C_0|, \sum_{(u,v)\in C_0}\frac{1}{m_um_v} \sim \mathcal{O}(K)$, $|\mathbf{Cor}_{\mathtt{B}}(R_{C_0}, n_a^B)|$ is bounded by a value smaller than 1.

□

S2.3 Proof of Lemma 2

Let $\overline{G}$ be the uMST on subjects, and $\mathcal{E}_i^{\overline{G}} = \{(i,j) : (i,j) \in \overline{G}\}$. Then $|\mathcal{E}_i^{\overline{G}}| = m_u + \sum_{\mathcal{V}_u} m_v - 1$, $|\overline{G}| = \sum_{u=1}^K m_u(m_u-1)/2 + \sum_{(u,v)\in C_0} m_um_v$. Since $\mathbf{E}_{\mathtt{P}}[T_{C_0}] = |\overline{G}|2p_1$, and the result follows.

Now, we compute the second moment.

$$\begin{aligned}
\mathbf{E}_{\mathtt{P}}[T_{C_0}^2] &= \sum_{(i,j),(k,l)\in\overline{G}} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) \\
&= \sum_{(i,j)\in\overline{G}} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j) + \sum_{(i,j),(i,k)\in\overline{G}, j\neq k} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_i \neq g_k) \\
&\quad + \sum_{\substack{(i,j),(k,l)\in\overline{G} \\ i,j,k,l \text{ all different}}} \mathbf{P}_{\mathtt{P}}(g_i \neq g_j, g_k \neq g_l) \\
&= |\overline{G}|2p_1 + \sum_{i=1}^N |\mathcal{E}_i^{\overline{G}}|(|\mathcal{E}_i^{\overline{G}}|-1)p_1 + (|\overline{G}|^2 - |\overline{G}| - \sum_{i=1}^N |\mathcal{E}_i^{\overline{G}}|(|\mathcal{E}_i^{\overline{G}}|-1))p_2 \\
&= (p_1-p_2)\sum_{u=1}^K m_u(m_u + \sum_{v\in\mathcal{V}_u} m_v - 1)(m_u + \sum_{v\in\mathcal{V}_u} m_v - 2) \\
&\quad + (p_1 - p_2/2)\left(\sum_{u=1}^K m_u(m_u-1) + 2\sum_{(u,v)\in C_0} m_um_v\right) \\
&\quad + p_2\left(\sum_{u=1}^K m_u(m_u-1) + 2\sum_{(u,v)\in C_0} m_um_v\right)^2.
\end{aligned}$$

$\mathbf{Var}_{\mathtt{P}}[T_{C_0}]$ follows by $\mathbf{E}_{\mathtt{P}}[T_{C_0}^2] - \mathbf{E}_{\mathtt{P}}^2[T_{C_0}]$.

□

Bibliography

L.H.Y. Chen and Q.M. Shao. Stein's method for normal approximation. *An introduction to Stein's method*, 4:1–59, 2005.