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.

\[ R_{\text{MST}} = |T|^{-1} \sum_{\tau \in T} R_{\tau} \]

\[ = |T|^{-1} \sum_{\tau_0 \in T_0} \sum_{\tau_1 \in T_1} \cdots \sum_{\tau_K \in T_K} [R_{\tau_0} + R_{\tau_1} + \cdots + R_{\tau_K}] \]

\[ = |T_0|^{-1} \sum_{\tau_0 \in T_0} R_{\tau_0} + \sum_{k=1}^{K} \left( \sum_{\tau_k \in T_k} R_{\tau_k}/S_{m_k} \right). \tag{S1.1} \]

First consider the quantity \( \sum_{\tau_k \in T_k} R_{\tau_k}/S_{m_k} \). 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 \( C_k \) has \( m_k-1 \) edges. Hence, the edge between each pair of subjects in \( C_k \) appears in exactly

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

trees. Thus,

\[ \sum_{\tau_k \in T_k} R_{\tau_k}/S_{m_k} = \sum_{i,j \in C_k, i \neq j} I_{g_i \neq g_j} \frac{2S_{m_k}}{S_{m_k}} = \frac{2n_ukbn_k}{m_k}. \tag{S1.2} \]

Next consider the summation over \( T_0 \). For any \( i \in C_u, j \in C_v \), if \( (u,v) \in \tau_0^* \), then the edge \( (i,j) \) appears in

\[ \prod_{k=1}^{K} m_k^{d_{i,j}^k}/(m_vm_v) \]
S2 Hao Chen and Nancy R. Zhang

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,

$$\sum_{\tau_0 \in \mathcal{T}_0} R_{\tau_0} = \sum_{\tau_0' \in \mathcal{T}_0'} \sum_{(u,v) \in \tau_0'} \prod_{k=1}^{K} m_k^{\frac{|E_{\tau_0'}^k|}{m_u m_v}} \sum_{i \in C_u} \sum_{j \in C_v} I_{g_i \neq g_j}$$

$$= \sum_{\tau_0' \in \mathcal{T}_0'} \prod_{k=1}^{K} m_k^{\frac{|E_{\tau_0'}^k|}{m_u m_v}} \sum_{(u,v) \in \tau_0'} \frac{n_{a \in \tau_0'} + n_{a \in \tau_0'}}{m_u m_v}.$$  \hfill (S1.3)

Combining (S1.1), (S1.2) and (S1.3) gives (7). \hfill \qed

S2 Proofs for Lemmas and Theorems in Permutation Distributions

S2.1 Proof of Lemma 1

Proof. Define

$$RA = \sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in C_u} I_{g_i \neq g_j},$$

and

$$RB = \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in C_u, j \in C_v} I_{g_i \neq g_j}.$$

We have

$$E_p[R_{C_0}] = E_p[R_A] + E_p[R_B]$$

$$= \sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in C_u} P_p(g_i \neq g_j) + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in C_u, j \in C_v} P_p(g_i \neq g_j).$$

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

$$E_p[R_{C_0}] = \sum_{u=1}^{K} \frac{1}{m_u} m_u (m_u - 1) \frac{2n_a n_b}{N(N-1)} + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} m_u m_v \frac{2n_a n_b}{N(N-1)}$$

$$= (N - K + |C_0|) \frac{2n_a n_b}{N(N-1)}.$$

Now, to compute the second moment, first note that

$$E_p[R_{C_0}^2] = E_p[R_A^2] + E_p[R_B^2] + 2E_p[R_AR_B].$$
Expanding the right-hand-side in above,

\[ E_p[R_A^2] = \sum_{u,v=1}^{k} \frac{1}{m_u m_v} \sum_{i,j \in C_u, k,l \in C_v} P_p(g_i \neq g_j, g_k \neq g_l), \]

\[ E_p[R_B^2] = \sum_{(u,v) \in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k \in C_u, j,l \in C_v} P_p(g_i \neq g_j, g_k \neq g_l) \]

\[ + 2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v m_w m_y} \sum_{i,j \in C_u, k,l \in C_v, m \in C_w, n \in C_y} P_p(g_i \neq g_j, g_k \neq g_l). \]

\[ E_p[R_A R_B] = \sum_{u=1}^{K} \sum_{(v,w) \in C_0} \frac{1}{m_u m_v m_w} \sum_{i,j \in C_u, k,l \in C_v} P_p(g_i \neq g_j, g_k \neq g_l). \]

Since

\[ P_p(g_i \neq g_j, g_k \neq g_l) = \begin{cases} 0 & \text{if } i = j \text{ and/or } k = l \\ \frac{2n_u n_w}{N(N-1)} = 2p_1 & \text{if } \begin{cases} i = k, j = l, i \neq j \\ i = l, j = k, i \neq j \\ i = k, j \neq i, l \\ i = l, j \neq i, k \\ j = k, i \neq j, l \\ j = l, i \neq j, k \end{cases} \\ \frac{n_u n_w}{N(N-1)} = p_1 & \text{if } i,j,k,l \text{ are all different,} \\ \frac{4n_u(n_u-1)n_w(n_w-1)}{N(N-1)(N-2)(N-3)} = p_2 & \end{cases} \]

we have

\[ E_p[R_A^2] = \sum_{u=1}^{K} \frac{1}{m_u^2} \sum_{i,j,k,l \in C_u} P_p(g_i \neq g_j, g_k \neq g_l) + \sum_{u=1}^{K} \sum_{v \neq u} \frac{1}{m_u m_v} \sum_{i,j \in C_u, k,l \in C_v} P_p(g_i \neq g_j, g_k \neq g_l) \]

\[ = \sum_{u=1}^{K} \frac{1}{m_u^2} \left[ 2m_u(m_u-1)(2p_1) + 4m_u(m_u-1)(m_u-2)p_1 + m_u(m_u-1)(m_u-2)(m_u-3)p_2 \right] \]

\[ + \sum_{u=1}^{K} \sum_{v \neq u} \frac{1}{m_u m_v} m_u(m_u-1)m_v(m_v-1)p_2 \]

\[ = 4 \left( N - 2K + \sum_{u=1}^{K} \frac{1}{m_u} \right) p_1 + (N - K - 4)(N - K)p_2 + 6 \left( K - \sum_{u=1}^{K} \frac{1}{m_u} \right) p_2 \]

\[ E_p[R_B^2] = \sum_{(u,v) \in C_0} \frac{1}{m_u^2 m_v^2} \sum_{i,k \in C_u, j,l \in C_v} P_p(g_i \neq g_j, g_k \neq g_l) \]

\[ + \sum_{(u,v) \in C_0, w \neq u} \frac{1}{m_u m_w m_v m_y} \sum_{i,k \in C_u, j \in C_v, m \in C_w, n \in C_y} P_p(g_i \neq g_j, g_k \neq g_l) \]

\[ + \sum_{(u,v) \in C_0, w \neq u} \frac{1}{m_u m_y m_v m_w} \sum_{i \in C_u, j \in C_v} P_p(g_i \neq g_j, g_k \neq g_l) \]
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 \( P_B, E_B \) and \( \text{Var}_B \) denote respectively the probability, expectation and variance under the bootstrap null.

\[
\text{E}_B[R_{AB}] = \sum_{u=1}^{K} \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i,j,k \in C_u} \sum_{l \in C_v} P_B(g_i \neq g_j, g_k \neq g_l)
\]
\[
+ \sum_{u=1}^{K} \sum_{(v,w) \in C_0 \setminus C_u} \frac{1}{m_u m_v} \sum_{i,j \in C_u} \sum_{k,l \in C_w} P_B(g_i \neq g_j, g_k \neq g_l)
\]
\[
= \sum_{u=1}^{K} \frac{1}{m_u m_v} [2m_u (m_u - 1) m_v p_1 + m_u (m_u - 1)(m_u - 2) m_v p_2]
\]
\[
+ \sum_{u=1}^{K} \frac{1}{m_u m_v m_w} m_u (m_u - 1) m_v m_w p_2
\]
\[
= |C_0| (N - K) p_2 + 2(p_1 - p_2) \left( 2|C_0| - \frac{|C_0^C|}{m_u} \right).
\]

\( \text{Var}_B[R_{C_0}] \) follows by combining the above in computing \( \text{E}_B[R_{C_0}^2] \), and then subtracting \( \text{E}_B^2[R_{C_0}] \). \( \square \)

### 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 \( P_B, E_B \) and \( \text{Var}_B \) denote respectively the probability, expectation and variance under the bootstrap null.
Lemma 1. Assuming condition 1, under the bootstrap null distribution, the standardized statistic
\[
\frac{R_{C_0} - \mathbb{E}_B[R_{C_0}]}{\sqrt{\text{Var}_B[R_{C_0}]}},
\]
converges in distribution to \( N(0, 1) \) as \( K \to \infty \), where \( \mathbb{E}_B[R_{C_0}] \) and \( \text{Var}_B[R_{C_0}] \) are given below.

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

\[
\text{Var}_B[R_{C_0}] = 4(p_3 - p_4) \left( N - K + 2|C_0| + \sum_{u=1}^{K} \frac{|C_{C_0}^u|^2}{4m_u} - \sum_{u=1}^{K} \frac{|C_{C_0}^u|}{m_u} \right) \tag{S2.5}
\]

\[+ (6p_4 - 4p_3) \left( K - \sum_{u=1}^{K} \frac{1}{m_u} \right) + p_4 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v},\]

where \( p_3 = \frac{n_a n_b}{N^2} \), \( p_4 = \frac{4n_a^2 n_b^2}{N^4} = 4p_3^2 \). \tag{S2.6}

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

Assumption 1. [Chen and Shao, 2005, p. 17] For each \( i \in J \) there exists \( S_i \subset T_i \subset J \) such that \( \xi_i \) is independent of \( \xi_{S_i^c} \) and \( \xi_{T_i^c} \) 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
\[
\sup_{h \in \text{Lip}(1)} |E h(W) - E h(Z)| \leq \delta
\]
where \( \text{Lip}(1) = \{ h : \mathbb{R} \to \mathbb{R} \} \), \( Z \) has \( N(0,1) \) 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 \( \eta_i = \sum_{j \in S_i} \xi_j \) 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,
\[
P_B(g_i \neq g_j) = \begin{cases} 0 & \text{if } i = j \\ \frac{2n_a n_b}{N^2} = 2p_3 & \text{if } i \neq j \end{cases}
\]
and

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

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

\[R_u = \frac{2n_a n_b}{m_u}, \quad d_u = E[B[R_u]] = 2(m_u - 1)p_3,\]

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

\[R_{uv} = \frac{n_a n_b + n_a n_b}{m_u m_v}, \quad d_{uv} = E[B[R_{uv}]] = 2p_3.\]

Let \( \sigma_B^2 = \text{Var}[R_{C_0}] \), \( \xi_u, \xi_{uv} \) be the standardized mixing potentials,

\[
\xi_u = \frac{R_u - d_u}{\sigma_B}, \quad (S2.7)
\]

\[
\xi_{uv} = \frac{R_{uv} - d_{uv}}{\sigma_B}. \quad (S2.8)
\]

Finally, we define the index sets for \( \xi_u \) and \( \xi_{uv} \):

\[J_1 = \{1, \ldots, K\}, \]

\[J_2 = \{uv : u < v \text{ such that } (u, v) \in C_0\}, \]

and let \( J = J_1 \cup J_2 \). Since \( R_{C_0} = \sum_{u=1}^{K} R_u + \sum_{(u,v) \in C_0} R_{uv} \), the standardized statistic is

\[W := \sum_{i \in J} \xi_i = \sum_{u \in J_1} \frac{R_u - d_u}{\sigma_B} + \sum_{uv \in J_2} \frac{R_{uv} - d_{uv}}{\sigma_B} = \frac{R_{C_0} - E[B[R_{C_0}]]}{\sigma_B}.\]

Our notation follows those of Theorem 1 and Assumption 1. For \( u \in J_1 \), let

\[S_u = \{u\} \cup \{uw, vu : (u, v) \in C_0\}, \]

\[T_u = S_u \cup \{v, uv, vu : (u, v) \in C_0\}.\]

For \( uv \in J_2 \), let

\[S_{uv} = \{uw, u, v\} \cup \{uw, wu : (u, w) \in C_0\} \cup \{vw, vw : (v, w) \in C_0\}, \]

\[T_{uv} = S_{uv} \cup \{w, wy, yw : (u, w), (v, y) \in C_0\} \cup \{w, wy, yw : (v, w), (w, y) \in C_0\}.\]
$S_u, T_u, S_{uv}, T_{uv}$ defined in this way satisfy Assumption 1.

Since $R_u \in [0, \frac{m_u}{2}], p_3 \in [0,\frac{1}{2}], and R_{uv} \in [0,1]$, and $d_u \in [0, \frac{m_u-1}{2}], d_{uv} \in [0, \frac{1}{2}]$, and therefore $|\xi_u| \leq \frac{m_u}{2\sigma_B}, |\xi_{uv}| \leq \frac{1}{\sigma_B}$. Hence,

$$\sum_{j \in S_u} |\xi_j| \leq \frac{1}{\sigma_B} (m_u + |E_u^C|), \quad u \in J_1,$$

$$\sum_{j \in T_u} |\xi_j| \leq \frac{1}{\sigma_B} (m_u + \sum_{v \in V_u} m_v + |E_{u,2}^C|), \quad u \in J_1,$$

$$\sum_{j \in S_{uv}} |\xi_j| \leq \frac{1}{\sigma_B} (m_u + m_v + |E_u^C| + |E_v^C|), \quad uv \in J_2,$$

$$\sum_{j \in T_{uv}} |\xi_j| \leq \frac{1}{\sigma_B} (m_u + m_v + \sum_{w \in V_u \cup V_v} m_w + |E_{u,2}^C| + |E_{v,2}^C|), \quad uv \in J_2.$$

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

$$|E_B| \sum_{j \in S_i} |E_B| \sum_{k \in T_i} \xi_j \xi_k | \leq |E_B| \sum_{j \in S_i} |E_j| \sum_{k \in T_i} |\xi_j| |\xi_k|,$$

$$|E_B| \sum_{j \in S_i} |\xi_j| |E_B| \sum_{j \in T_i} |\xi_j| \leq |E_B| \sum_{j \in S_i} |E_j| \sum_{j \in T_i} |\xi_j| |\xi_j|,$$

$$|E_B| \sum_{j \in S_i} \sum_{k \in S_i} \xi_j \xi_k \leq |E_B| \sum_{j \in S_i} |E_j| \sum_{k \in S_i} |\xi_j| |\xi_k|.$$

Thus, for $i = u \in J_1$, the terms $E_B|\xi_i| \theta_i|, |E_B| \sum_{j \in S_i} |E_B| \sum_{j \in T_i} |\xi_j|, and |E_B| \sum_{j \in S_i} |E_B| \sum_{j \in T_i} |\xi_j| |\xi_j|$, and $E_B|\xi_i| \theta_i|$, are all bounded by

$$\frac{1}{\sigma_B} m_u (m_u + |E_u^C|) (m_u + \sum_{v \in V_u} m_v + |E_{u,2}^C|),$$

and for $i = uv \in J_2$, the terms $E_B|\xi_i| \theta_i|, |E_B| \sum_{j \in S_i} |E_B| \sum_{j \in T_i} |\xi_j|, and |E_B| \sum_{j \in S_i} |E_B| \sum_{j \in T_i} |\xi_j| |\xi_j|$, are all bounded by

$$\frac{1}{\sigma_B} (m_u + m_v + |E_u^C| + |E_v^C|) (m_u + m_v + \sum_{w \in V_u \cup V_v} m_w + |E_{u,2}^C| + |E_{v,2}^C|).$$

Hence,

$$\delta \leq \frac{5}{\sigma_B} \left( \sum_{u=1}^K m_u (m_u + |E_u^C|) (m_u + \sum_{v \in V_u} m_v + |E_{u,2}^C|) + \sum_{(u,v) \in C_0} (m_u + m_v + |E_u^C| + |E_v^C|) (m_u + m_v + \sum_{w \in V_u \cup V_v} m_w + |E_{u,2}^C| + |E_{v,2}^C|) \right),$$

Since $\sigma_B$ is of order $\sqrt{K}$ or higher, under condition 1, $\delta \rightarrow 0$ as $K \rightarrow \infty$. 

\[\square\]
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

\[
\left( \frac{R_{C_0} - E_B[R_{C_0}]}{\sqrt{\text{Var}_B[R_{C_0}]}} \frac{n_a^B - Np_a}{\sigma_0} \right)
\]

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

\[
a_1 \frac{R_{C_0} - E_B[R_{C_0}]}{\sqrt{\text{Var}_B[R_{C_0}]}} + a_2 \frac{n_a^B - Np_a}{\sigma_0}
\]

is asymptotic Gaussian under the bootstrap null for all \(a_1, a_2 \in \mathbb{R}, a_1a_2 \neq 0\).

Let \(\xi_i, i \in J\) be defined in the same way as in the proof of Lemma 1. Let \(J_3 = \{|J| + 1, \ldots, |J| + K\}\). For \(i \in J_3\), let

\[
\xi_i = \frac{n_{au} - p_awh}{\sigma_0}, \quad i' = i - |J|.
\]

We use Theorem 1 to show the asymptotic Gaussianity of \(\sum_{i \in J} a_1\xi_i + \sum_{i \in J_3} a_2\xi_i\). We need to redefine the neighborhood sets to satisfy Assumption 1.

For \(u \in J_1\),

\[
S_u = \{u, u + |J|\} \cup \{uv, vu : (u, v) \in C_0\},
\]

\[
T_u = S_u \cup \{v, v + |J|, uv, vu : (u, v), (v, w) \in C_0\}.
\]

For \(uv \in J_2\),

\[
S_{uv} = \{uv, u, v, u + |J|, v + |J|\} \cup \{uv, wu : (u, w) \in C_0\}
\]

\[
\cup \{vw, wv : (v, w) \in C_0\},
\]

\[
T_{uv} = S_{uv} \cup \{w, w + |J|, wy, yw : (u, w), (w, y) \in C_0\}
\]

\[
\cup \{w, w + |J|, wy, yw : (v, w), (w, y) \in C_0\}.
\]

And for \(u \in J_3\),

\[
S_u = \{u, u'\} \cup \{u'v, vu' : (u', v) \in C_0\}, \quad u' = u - |J|,
\]

\[
T_u = S_u \cup \{v, v + |J|, uv, vu : (u', v), (v, w) \in C_0\}.
\]

From the proof of Lemma 1, we have

\[|\xi_u| \leq \frac{m_u}{2\sigma_B}, \quad \forall u \in J_1; \quad |\xi_{uv}| \leq \frac{1}{\sigma_B}, \quad \forall uv \in J_2.\]
For \( u \in J_3 \),
\[
|\xi_u| \leq \frac{m_u}{\sigma_0}, \quad u' = u - |J|.
\]

Let \( \sigma = \min(\sigma_0, \sigma_0) \), then
\[
\sum_{j \in S_u} |\xi_j| \leq \frac{1}{\sigma} (2m_u + |C^e_u|), \quad u \in J_1 \cup J_3,
\]
\[
\sum_{j \in T_u} |\xi_j| \leq \frac{1}{\sigma} (2m_u + 2 \sum_{v \in V_u} m_v + |C^e_u|), \quad u \in J_1 \cup J_3,
\]
\[
\sum_{j \in S_{uv}} |\xi_j| \leq \frac{1}{\sigma} (2m_u + 2m_v + |C^e_u| + |C^e_v|), \quad uv \in J_2,
\]
\[
\sum_{j \in T_{uv}} |\xi_j| \leq \frac{1}{\sigma} (2m_u + 2m_v + 2 \sum_{w \in V_u \cup V_v} m_w + |C^e_{uv}| + |C^e_{v,2}|), \quad uv \in J_2.
\]

Thus, for \( i = u \in J_1 \cup J_3 \), the terms \( E_b|\xi_i \eta_i| \), \( |E_b(\xi_i \eta_i)|E_b|\theta_i| \), and \( E_b|\xi_i \eta_i^2| \) are all bounded by
\[
\frac{1}{\sigma^3} m_u (2m_u + |C^e_u|)(2m_u + 2 \sum_{v \in V_u} m_v + |C^e_v|),
\]
and for \( i = uv \in J_2 \), terms \( E_b|\xi_i \eta_i| \), \( |E_b(\xi_i \eta_i)|E_b|\theta_i| \), and \( E_b|\xi_i \eta_i^2| \) are all bounded by
\[
\frac{1}{\sigma^3} (2m_u + 2m_v + |C^e_u| + |C^e_v|)(2m_u + 2m_v + 2 \sum_{w \in V_u \cup V_v} m_w + |C^e_{uv}| + |C^e_{v,2}|).
\]

Define \( W_{a_1, a_2} = \sum_{i \in J} a_1 \xi_i + \sum_{i \in J_3} a_2 \xi_i \). The value of \( \delta \) in Theorem 1 has the form
\[
\delta = \frac{1}{\sqrt{E_b|W_{a_1, a_2}^2|}} \left( 2 \sum_{i \in J} (E_b|a_1 \xi_i \eta_i \theta_i| + |E_b(a_1 \xi_i \eta_i)|E_b|\theta_i|) + \sum_{i \in J} E_b|a_1 \xi_i \eta_i^2| \right.
\]
\[
+ 2 \sum_{i \in J_3} (E_b|a_2 \xi_i \eta_i \theta_i| + |E_b(a_2 \xi_i \eta_i)|E_b|\theta_i|) + \sum_{i \in J_3} E_b|a_2 \xi_i \eta_i^2| \right),
\]
where \( \eta_i = \sum_{j \in S_j} \xi_j(a_1 I_{j} + a_2 I_{j} \in J_3) \), and \( \theta_i = \sum_{j \in T_j} \xi_j(a_1 I_{j} + a_2 I_{j} \in J_3) \).
Let \( a = \max(|a_1|, |a_2|) \), we have

\[
E_b|a_1 \xi_i \eta_i \theta_i|, \quad E_b|a_2 \xi_i \eta_i \theta_i| \leq a^3 E_b|\xi_i \sum_{j \in S_i} \xi_j \sum_{k \in T_i} \xi_k| \\
\leq a^3 E_b|\xi_i \sum_{j \in S_i} |\xi_j| \sum_{k \in T_i} |\xi_k|,
\]

\[
|E_b(a_1 \xi_i \eta_i)|E_b|\theta_i|, \quad |E_b(a_2 \xi_i \eta_i)|E_b|\theta_i| \leq a^3 E_b|\xi_i \sum_{j \in S_i} |\xi_j|E_b| \sum_{j \in T_i} |\xi_j| \\
\leq a^3 E_b|\xi_i \sum_{j \in S_i} |\xi_j|E_b| \sum_{j \in T_i} |\xi_j|,
\]

\[
E_b|a_1 \xi_i \eta_i^2|, \quad E_b|a_2 \xi_i \eta_i^2| \leq a^3 E_b|\xi_i \sum_{j \in S_i} \sum_{k \in S_i} \xi_j \xi_k | \\
\leq a^3 E_b|\xi_i \sum_{j \in S_i} |\xi_j| \sum_{k \in S_i} |\xi_k|.
\]

Thus,

\[
\delta \leq \frac{40a^3}{\sigma^3 \sqrt{E_b|W_{a_1,a_2}^2|}} \left( \sum_{u=1}^{K} m_u (m_u + |E_u^C|)(m_u + \sum_{v \in V_u} m_v + |E_v^C|) \right) \\
+ \sum_{(u,v) \in C_0} (m_u + m_v + |E_u^C| + |E_v^C|)(m_u + m_v + \sum_{w \in V_u \cup V_v} m_w + |E_u^C| + |E_v^C|).
\]

Since \( \sigma_b^2 \) is at least of order \( K \) and \( \sigma_0^2 \) is of order \( N \), \( \sigma^2 \) is at least of order \( K \) by Condition 2. If \( E_b|W_{a_1,a_2}^2| \) is uniformly strictly bounded from 0 for any \( a_1 a_2 \neq 0 \), then under Condition 1, \( \delta \to 0 \) as \( K \to \infty \).

We next show that under Condition 2, \( E_b|W_{a_1,a_2}^2| \) is uniformly strictly bounded from 0 for any \( a_1 a_2 \neq 0 \).

Let \( W_1 = \sum_{i \in J} \xi_i, W_2 = \sum_{i \in J_2} \xi_i \), then

\[
E_b|W_{a_1,a_2}^2| = a_1^2 E_bW_1^2 + a_2^2 E_bW_2^2 + 2a_1a_2 E_b[W_1W_2] \\
= a_1^2 + a_2^2 + 2a_1a_2 E_b[W_1W_2]
\]

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 \( n_u/N \) to be bounded from 0 and 1, so \( p_u \) and \( p_b \) are both bounded from 0 and 1.
Correlation between $R_{C_0}$ and $n_a^B$. Observe that

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

$$= \sum_{u=1}^{K} \frac{1}{m_u} \sum_{i,j \in C_u} \left( \sum_{x=1}^{N} I_{g_x = a} \right) + \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sum_{i \in C_u, j \in C_v} \left( \sum_{x=1}^{N} I_{g_x = a} \right).$$

For any $i \neq j$,

$$E_0 \left[ I_{g_i \neq g_j} \sum_{x=1}^{N} I_{g_x = a} \right] = E_0 \left[ I_{g_i \neq g_j, g_i = a} + I_{g_i \neq g_j, g_j = a} + \sum_{x \neq i,j} I_{g_i \neq g_j, g_x = a} \right]$$

$$= P_0(g_i = a, g_j = b) + P_0(g_i = b, g_j = a) + \sum_{x \neq i,j} P_0(g_i \neq g_j, g_x = a)$$

$$= p_a p_b + p_a p_b + 2p_a p_b p_a (N - 2) = 2p_a p_b (N p_a + 1 - 2p_a).$$

Hence

$$E_0[R_{C_0}n_a^B] = (N - K + |C_0|)2p_a p_b (N p_a + 1 - 2p_a).$$

Since $E_0[R_{C_0}] = (N - K + |C_0|)2p_a p_b$ and $E_0[n_a^B] = N p_a$, we have

$$Cov_0(R_{C_0}, n_a^B) = (N - K + |C_0|)2p_a p_b (1 - 2p_a). \quad (S2.9)$$

If $p_a = 1/2$, then $Cov_0(R_{C_0}, n_a^B) = 0$. Since $Var_0[R_{C_0}]$ and $Var_0[n_a^B] = N p_a$ are positive, $Cor_0(R_{C_0}, n_a^B) = 0$, clearly bounded from 1. We consider $p_a \neq 1/2$ in the following.

$$Var_0[R_{C_0}] = 4p_a p_b (1 - 4p_a p_b) \left( N - K + 2|C_0| + \sum_{u=1}^{K} \frac{|E_u|}{4m_u} - \sum_{u=1}^{K} \frac{|E_u|}{m_u} \right)$$

$$+ 4p_a p_b (6p_a p_b - 1) \left( K - \sum_{u=1}^{K} \frac{1}{m_u} \right) + 4p_a^2 p_b^2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}$$

$$= 4p_a p_b (1 - 4p_a p_b) \left( N - 2K + 2|C_0| + \sum_{u=1}^{K} \frac{|E_u|/2 - 1}{m_u} \right)$$

$$+ 8p_a^2 p_b^2 \left( K - \sum_{u=1}^{K} \frac{1}{m_u} \right) + 4p_a^2 p_b^2 \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}.$$
we have
\[
\Var_{B}[R_{C_0}] \Var_{B}[n_B^R] \geq 4p_a^2p_b(1 - 4p_a p_b) |N - |C_0||^2 + 4p_a^3p_b^3 N \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}.
\]

Hence,
\[
|\Cor_{B}(R_{C_0}, n_B^R)| \leq \frac{1}{\sqrt{1 + \frac{p_a p_b N \sum_{(u,v) \in C_0} \frac{1}{m_u m_v}}{(1 - 4p_a p_b)|N - |C_0||^2}}}
\]

When \(N, |C_0|, \sum_{(u,v) \in C_0} \frac{1}{m_u m_v} \sim O(K)\), \(|\Cor_{B}(R_{C_0}, n_B^R)|\) is bounded by a value smaller than 1.

### S2.3 Proof of Lemma 2

Let \(\overline{G}\) be the uMST on subjects, and \(E_{\overline{G}} = \{(i, j) : (i, j) \in \overline{G}\}\). Then \(|E_{\overline{G}}| = m_u + \sum_{V_u} m_v - 1, |\overline{G}| = \sum_{u=1}^{K} m_u(m_u - 1)/2 + \sum_{(u,v) \in C_0} m_u m_v\). Since \(E_p[T_{C_0}] = |\overline{G}| 2p_1\), and the result follows.

Now, we compute the second moment.
\[
E_p[T_{C_0}^2] = \sum_{(i,j),(k,l) \in \overline{G}} P_p(g_i \neq g_j, g_k \neq g_l)
\]
\[
= \sum_{(i,j) \in \overline{G}} P_p(g_i \neq g_j) + \sum_{(i,j), (i,k) \in \overline{G}, j \neq k} P_p(g_i \neq g_j, g_i \neq g_k)
\]
\[
+ \sum_{(i,j), (k,l) \in \overline{G}, i,j,k,l \text{ all different}} P_p(g_i \neq g_j, g_k \neq g_l)
\]
\[
= |\overline{G}| 2p_1 + \sum_{i=1}^{N} |E_{\overline{G}}^i|(|E_{\overline{G}}^i| - 1)p_1 + |\overline{G}|^2 - |\overline{G}| - \sum_{i=1}^{N} |E_{\overline{G}}^i|(|E_{\overline{G}}^i| - 1)p_2
\]
\[
= (p_1 - p_2) \sum_{u=1}^{K} m_u(m_u + \sum_{v \in V_u} m_v - 1)(m_u + \sum_{v \in V_u} m_v - 2)
\]
\[
+ (p_1 - p_2/2) \left( \sum_{u=1}^{K} m_u(m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right)
\]
\[
+ p_2 \left( \sum_{u=1}^{K} m_u(m_u - 1) + 2 \sum_{(u,v) \in C_0} m_u m_v \right)^2.
\]

\(\Var_p[T_{C_0}]\) follows by \(E_p[T_{C_0}^2] - E_p^2[T_{C_0}]\).
Bibliography