Graph-based two-sample tests for data with repeated observations

Journal:	*Statistica Sinica*
Manuscript ID	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Zhang, Jingru; Peking University, Beijing International Center for Mathematical Research Chen, Hao ; University of California, Davis, Statistics
Keywords:	Ties in distance, Nonparametric test, Non-Euclidean data

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

Jingru Zhang and Hao Chen

Peking University

and University of California, Davis

Abstract: In the regime of two-sample comparison, tests based on a graph constructed on observations by utilizing similarity information among them is gaining attention due to their flexibility and good performances for high-dimensional/non-Euclidean data. However, when there are repeated observations, these graph-based tests could be problematic as they are versatile to the choice of the similarity graph. We propose extended graph-based test statistics to resolve this problem. We also study asymptotic properties of these extended statistics and provide analytic formulas to approximate the p-values of the tests under finite samples, facilitating the application of the new tests in practice. The new tests are illustrated in the analysis of a phone-call network dataset. All tests are implemented in an `R` package `gTests`.

Key words and phrases: High-dimensional data; Network data; Non-Euclidean data; Nonparametric test; Similarity graph; Ties in distance.

1. Introduction

Two-sample comparison is a fundamental problem in statistics and has been extensively studied for univariate data and low-dimensional data. The testing problem for high-dimensional data and non-Euclidean data, such as network data, is gaining more and more attention in this big-data era. In the parametric domain, for multivariate data, many endeavors have been made in testing whether the means are the same (see for examples Bai and Saranadasa (1996); Srivastava and Du (2008); Chen et al. (2010); Cai et al. (2014); Xu et al. (2016)) and whether the covariance matrices are the same (see for examples Schott (2007); Srivastava and Yanagihara (2010); Li and Chen (2012); Cai et al. (2013); Xia et al. (2015)). These parametric methods provide useful tools when the data follow their assumptions, but they are often restrictive and not robust enough if model assumptions are violated.

In the nonparametric domain, efforts had been made in extending the Kolmogorov-Smirnov test, the Wilcoxon rank test, and the Wald-Wolfowitz runs test to high-dimensional data (see Chen and Friedman (2017) for a review). Among these efforts, the first practical test was proposed by Friedman and Rafsky (1979) as an extension of the Wald-Wolfowitz runs test to multivariate data. They pool the observations from the two samples to-

gether and construct a minimum spanning tree (MST), which is a spanning tree that connects all observations with the sum of the distances of the edges in the tree minimized. They then count the number of edges in the MST that connects observations from different samples, and reject the null hypothesis of equal distribution if this count is significantly *smaller* than its expectation under the null hypothesis. This test later has been extended to other similarity graphs where observations closer in distance are more likely to be connected than those farther in distance, such as the minimum distance pairing (MDP) in Rosenbaum (2005) and the nearest neighbor graph (NNG) in Schilling (1986) and Henze (1988). We call this type of tests the *edge-count test* for easy reference. Recently, a generalized edge-count test and a weighted edge-count test were proposed to address the problems of the edge-count test under scale alternatives and under unequal sample sizes (Chen and Friedman, 2017; Chen et al., 2018). Since these tests and the edge-count test are all based on a similarity graph, we call them the *graph-based tests*. These tests have many advantages: They can be applied to data with arbitrary dimension and to non-Euclidean data, and exhibit high power in detecting a variety of differences in distribution – they have higher power than likelihood-based tests when the dimension of the data is moderate to high for practical sample sizes, from hundreds to millions.

However, the graph-based tests could be problematic for data with repeated observations. All these tests rely on a similarity graph constructed on the observations. When there are repeated observations, the similarity graph is not uniquely defined based on common optimization criteria, such as the MST or the MDP. Indeed, several graphs could be equally "optimal" in terms of the criterion. The results of the graph-based tests can vary a lot under the different similarity graphs, leading to conflicting conclusions (see Table 1 for a snapshot of the results of the generalized and weighted edge-count tests on a network data set; details in Supplement 2.1).

Table 1: Test statistics and their corresponding p-values (in parentheses, bold if < 0.05) of the generalized edge-count test (S) and the weighted edge-count test (Z_w) under four 9-MSTs on the phone-call network data.

MST	#1	#2	#3	#4
S	6.86 (**0.032**)	3.92 (0.141)	7.89 (**0.019**)	3.90 (0.142)
Z_w	2.61 (**0.004**)	1.95 (**0.025**)	-1.13 (0.871)	0.26 (0.396)

In this work, we seek ways to effectively summarize the tests over these equally "optimal" similarity graphs. As we will show in Section 2.2, it is easy to have more than a million equally optimal similarity graphs when there are repeat observations, so manually examining the results from each of these graphs is usually not feasible. Chen and Zhang (2013) studied the problem of extending the original edge-count test to deal with repeated

observations. However, due to the quadratic terms in the generalized edge-count test statistic, directly extending the statistic to deal with repeated observations following the approach in Chen and Zhang (2013) is not feasible (details seen in Section 3). On the other hand, we could first extend the basic quantities in these graph-based test statistics so that they can handle repeated observations and then define extended generalized/weighted edge-count test statistics in a way similar to how they were designed at the first place for continuous data. Following this line, we show the following results in the paper:

(1) The extended weighted edge-count test statistic constructed in this way adopts the same weights as the weighted edge-count test to resolve the variance boosting problem of the edge-count test when the sample sizes of the two samples are different;

(2) The extended generalized edge-count test statistic can be well defined in this way, and we further show that it can be decomposed into the summation of squares of two asymptotically independent normal random variables, allowing for fast approximate p-value computation.

Based on (2), we further study an extended max-type edge-count test that builds upon the two asymptotically independent normal random variables.

The tests are all implemented in an `R` package `gTests`.

The rest of the paper is organized as follows. Section 2 provides notations used in the paper and preliminary setups. Section 3 discusses in details the extended weighted, generalized, and max-type edge-count tests. The performance of these new tests is examined in Section 4 and their asymptotic properties are studied in Section 5. Section 6 illustrates the new tests in the analysis of the phone-call network dataset.

2. Notations and preliminary setups

2.1 Basic notations

For data with repeated observations, assume that there are K distinct values and we index them by $1, 2, \cdots, K$. In the following, we use the notations summarized in Table 2.

Table 2: Data with repeated observations summarized by distinct values.

Distinct value index	1	2	$\cdots$	K	Total
Sample 1	n_{11}	n_{12}	$\cdots$	n_{1K}	n_1
Sample 2	n_{21}	n_{22}	$\cdots$	n_{2K}	n_2
Total	m_1	m_2	$\cdots$	m_K	N

Here, $m_i = n_{1i} + n_{2i},\ i = 1, \cdots, K;\ n_i = \sum_{k=1}^{K} n_{ik},\ i = 1, 2;\ N = n_1 + n_2$.

Let $\{d(i, j) : i, j = 1, \cdots, K\}$ be the distance matrix on the distinct values, with $d(i, j)$ being the distance between values indexed by i and j,

respectively. For an undirected graph G, let $|G|$ be the number of edges in G. For any node i in the graph G, $\mathcal{E}_i^G$ denotes the set of edge(s) in G that contains node i.

We do not impose any distributional assumption on the data and work under the permutation null distribution, which places $n_1!n_2!/N!$ probability on each of the $N!/(n_1!n_2!)$ ways of assigning the sample labels such that sample 1 has n_1 observations. Without further specification, we use E, Var, Cov, Cor to denote the expectation, variance, covariance and correlation under the permutation null distribution.

2.2 Similarity graphs on observations

Let C_0 be a similarity graph constructed on the distinct values. It could be the MST, the MDP, or the NNG on the distinct values if it can be uniquely defined. If the common optimization rules do not result in an unique solution, we adopt the same treatment as in Chen and Zhang (2013) by using the union of all MSTs. Figure 1 is a simple example. It can be shown that this union of all MSTs on the distinct values can be obtained through Algorithm 1. For example, for the data in Figure 1, distinct values **a** and **b**, **a** and **c**, **b** and **c**, **d** and **e** are connected in the first step, then **b** and **d**, **c** and **e** are connected in the second step. We call this graph

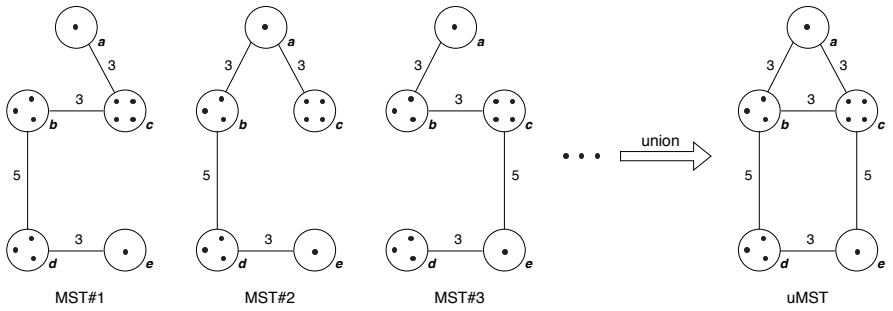


Figure 1: There are five distinct values (**a**, **b**, **c**, **d**, **e**) denoted by the circles. For some distinct values, there are more than one observations, denoted by the more than one point in the circle. The distance between the distinct values are denoted on the edges. It is clear that there are six MSTs on the distinct values (three of them are presented on the left) and the last plot is the union of the six MSTs on the distinct values.

the nearest neighbor link (NNL) for easy reference. If one wants denser graphs, k-NNL could be considered, which is the union of the 1st,$\cdots$,kth NNLs, where the jth ($j > 1$) NNL is a graph generated by Algorithm 1 subject to the constraint that this graph does not contain any edge in the 1st,$\cdots$,$(j-1)$th NNLs.

Then, a graph on observations initiated from C_0 can be defined in the following way: First, for each pair of edges $(i, j) \in C_0$, randomly choose an observation with value indexed by i and another observation with value indexed by j, connect these two observations; then, for each i, if there are more than one observation with value indexed by i, connect these observa-

Algorithm 1 Generate a NNL

For each distinct value indexed by i $(i = 1, \cdots, K)$, let $d_{\min}(i) = \min\{d(i,j) : j \neq i\}$ and $N(i) = \{j : d(i,j) = d_{\min}(i)\}$. Connect i to each element in $N(i)$.

while Not all distinct values are in one component **do**

Let $\mathcal{U}$ be one component, let $d_{\min}(\mathcal{U}) = \min\{d(i,j) : i \in \mathcal{U}, j \notin \mathcal{U}\}$ and $N(\mathcal{U}) = \{(i,j) : d(i,j) = d_{\min}(\mathcal{U}), i \in \mathcal{U}, j \notin \mathcal{U}\}$. Connect each pair of distinct values indexed by i and j if $(i,j) \in N(\mathcal{U})$.

end while

tions by a spanning tree (any edge in such a spanning tree has distance 0). Let $\mathcal{G}_{C_0}$ be the set of all graphs initiated from C_0.

For the example in Figure 1, since the MST on the distinct values is not uniquely defined, let C_0 be the NNL. There are only 5 distinct values and 6 edges on C_0. However, there are $15,552 (= 1^2 \cdot 3^3 \cdot 4^3 \cdot 3^2 \cdot 1^2)$ different ways in assigning the 6 edges in C_0 to corresponding observations in each circle. In addition, by Cayley's lemma, for the observations belonging to the same distinct value, there are 1, 3, 16, 3 and 1 spanning trees, respectively. Therefore, we have $2,239,488 (= 15,552 \times 3 \times 16 \times 3)$ graphs in $\mathcal{G}_{C_0}$. Figure 2 plots four of these graphs for illustration.

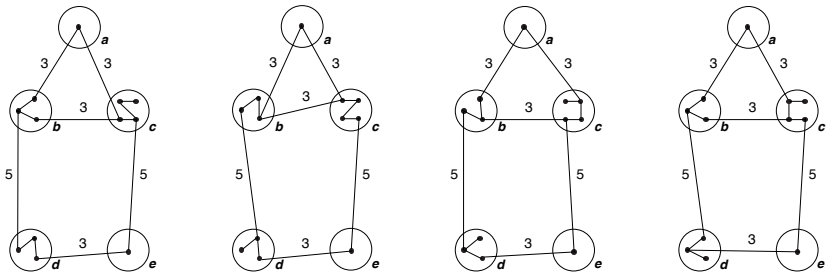


Figure 2: Four graphs, out of 2,239,488, on observations initiated from the NNL on distinct values.

2.3 A brief review of generalized and weighted edge-count tests

For any graph G, let $R_{0,G}$ be the number of edges in G that connect observations from different samples, $R_{1,G}$ be the number of edges in G that connect observations from sample 1, and $R_{2,G}$ be that for sample 2. Here, $R_{0,G}$ is the test statistic for the original edge-count test. In Chen and Friedman (2017), the authors noticed that, the edge-count test ($R_{0,G}$) has low or even no power for scale alternatives when the dimension is moderate to high unless the sample size is extremely large due to the curse-of-dimensionality. To solve this problem, they considered the numbers of within-sample edges of the two samples separately and proposed the following generalized edge-

count statistic

$$S_G = \begin{pmatrix} R_{1,G} - \mathsf{E}(R_{1,G}) \\ R_{2,G} - \mathsf{E}(R_{2,G}) \end{pmatrix}^T \boldsymbol{\Sigma}_G^{-1} \begin{pmatrix} R_{1,G} - \mathsf{E}(R_{1,G}) \\ R_{2,G} - \mathsf{E}(R_{2,G}) \end{pmatrix}, \qquad (2.1)$$

where $\boldsymbol{\Sigma}_G = \mathsf{Var}(\binom{R_{1,G}}{R_{2,G}})$.

Both the edge-count test and the generalized edge-count test are suggested to perform on a similarity graph that is denser than the MST, such as 5-MST, to boost their power (Friedman and Rafsky, 1979; Chen and Friedman, 2017). Here, a k-MST is the union of the 1st,$\cdots$, kth MSTs, where the 1st MST is the MST and the jth ($j > 1$) MST is a spanning tree that connects all observations such that the sum of the edges in the tree is minimized under the constraint that it does not contain any edge in the 1st,$\cdots$, $(j-1)$th MSTs. However, Chen et al. (2018) found that, for k-MST ($k > 1$), the edge-count test ($R_{0,G}$) behaves weirdly when the two sample sizes are different. For example, consider the testing problem that the two underlying distributions are $\mathcal{N}_d(0, \mathbf{I}_d)$ vs $\mathcal{N}_d(\mu, \mathbf{I}_d)$ (e.g., $\|\mu\|_2 = 1.3,\ d = 50$), and two scenarios (i) $n_1 = n_2 = 50$ and (ii) $n_1 = 50,\ n_2 = 100$. The edge-count test has lower power in (ii) compared to that in (i) even though there are more observations in (ii). This is due to a variance boosting issue under unbalanced sample sizes (details seen in Chen et al. (2018)). To solve this issue, Chen et al. (2018) proposed a

weighted edge-count test by inversely weighting the within-sample edges by the sample sizes

$$R_{w,G} = \frac{n_2 - 1}{n_1 + n_2 - 2} R_{1,G} + \frac{n_1 - 1}{n_1 + n_2 - 2} R_{2,G} \tag{2.2}$$

with the reasoning that the sample with a larger number of observations is more likely to be connected within the sample if all other conditions are the same and thus shall be down-weighted. This weighted edge-count test statistic successfully addressed the variance boosting issue and works well for unequal sample sizes. Indeed, $\mathsf{Var}(R_{w,G}) \leq \mathsf{Var}((1-p)R_{1,G} + pR_{2,G})$ for any $p \in [0, 1]$.

2.4 Extended basic quantities in the graph-based framework

In Chen and Zhang (2013), the authors considered two ways to summarize the test statistics for $R_{0,G}$:

(1) averaging

$R_{0,(a)} = \frac{1}{|\mathcal{G}_{C_0}|} \sum_{G \in \mathcal{G}_{C_0}} R_{0,G}$ where $|\mathcal{G}_{C_0}|$ is the number of graphs in $\mathcal{G}_{C_0}$;

(2) union

$R_{0,(u)} = R_{0,\bar{G}_{C_0}}$ where $\bar{G}_{C_0} = \cup\{G \in \mathcal{G}_{C_0}\}$, i.e., if observations u and v are connected in at least one of the graphs in $\mathcal{G}_{C_0}$, then these two

observations are connected in $\bar{G}_{C_0}$. In the following, we sometimes use $\bar{G}$ instead of $\bar{G}_{C_0}$ when there is no confusion for simplicity.

Since it is easy to have a lot of graphs in $\mathcal{G}_{C_0}$, it is many times not feasible to compute these two quantities directly. Chen and Zhang (2013) derived analytic expressions for computing these two quantities in terms of the summary quantities in Table 2 and C_0:

$$R_{0,(a)} = \sum_{k=1}^{K} \frac{2n_{1k}n_{2k}}{m_k} + \sum_{(u,v)\in C_0} \frac{n_{1u}n_{2v} + n_{1v}n_{2u}}{m_u m_v},$$

$$R_{0,(u)} = \sum_{k=1}^{K} n_{1k}n_{2k} + \sum_{(u,v)\in C_0} (n_{1u}n_{2v} + n_{1v}n_{2u}).$$

Similarly, we could defined $R_{1,(a)}$, $R_{1,(u)}$, $R_{2,(a)}$ and $R_{2,(u)}$, and their analytic expressions in terms of the summary quantities in Table 2 and C_0 are given in Lemma 1.

Lemma 1. *The analytic expressions for $R_{1,(a)}$, $R_{1,(u)}$, $R_{2,(a)}$ and $R_{2,(u)}$ are:*

$$\begin{aligned}
R_{1,(a)} &\equiv \frac{1}{|\mathcal{G}_{C_0}|} \sum_{G\in\mathcal{G}_{C_0}} R_{1,G} &&= \sum_{u=1}^{K} \frac{n_{1u}(n_{1u}-1)}{m_u} + \sum_{(u,v)\in C_0} \frac{n_{1u}n_{1v}}{m_u m_v}, \\
R_{1,(u)} &\equiv R_{1,\bar{G}_{C_0}} &&= \sum_{u=1}^{K} \frac{n_{1u}(n_{1u}-1)}{2} + \sum_{(u,v)\in C_0} n_{1u}n_{1v}, \\
R_{2,(a)} &\equiv \frac{1}{|\mathcal{G}_{C_0}|} \sum_{G\in\mathcal{G}_{C_0}} R_{2,G} &&= \sum_{u=1}^{K} \frac{n_{2u}(n_{2u}-1)}{m_u} + \sum_{(u,v)\in C_0} \frac{n_{2u}n_{2v}}{m_u m_v}, \\
R_{2,(u)} &\equiv R_{2,\bar{G}_{C_0}} &&= \sum_{u=1}^{K} \frac{n_{2u}(n_{2u}-1)}{2} + \sum_{(u,v)\in C_0} n_{2u}n_{2v}.
\end{aligned}$$

These analytic expressions can be obtained through similar arguments in Chen and Zhang (2013) and the proof is omitted here.

3. Extended graph-based tests

Since the generalized edge-count test could cover a wider range of alternatives than the original edge-count test (Chen and Friedman, 2017), we would like to have the generalized edge-count test statistic well defined when there are repeated observations. For the generalized edge-count test statistic:

$$S_G = \begin{pmatrix} R_{1,G} - \mathsf{E}(R_{1,G}) \\ R_{2,G} - \mathsf{E}(R_{2,G}) \end{pmatrix}^T \boldsymbol{\Sigma}_G^{-1} \begin{pmatrix} R_{1,G} - \mathsf{E}(R_{1,G}) \\ R_{2,G} - \mathsf{E}(R_{2,G}) \end{pmatrix},$$

one straightforward way of defining the average statistic would be

$$\frac{1}{|\mathcal{G}_{C_0}|} \sum_{G \in \mathcal{G}_{C_0}} S_G.$$

However, $\boldsymbol{\Sigma}_G$ varies for different G's in $\mathcal{G}_{C_0}$, making the averaging over S_G's difficult to move forward. Even consider the simplified version that Σ_G is fixed over G's in $\mathcal{G}_{C_0}$, the quadratic terms in S_G also make the averaging analytically intractable. To view the problem more straightforwardly, notice that S_G can be written as

$$S_G = \left(\frac{R_{w,G} - \mathsf{E}(R_{w,G})}{\sqrt{\mathsf{Var}(R_{w,G})}} \right)^2 + \left(\frac{R_{d,G} - \mathsf{E}(R_{d,G})}{\sqrt{\mathsf{Var}(R_{d,G})}} \right)^2,$$

where $R_{w,G} = \frac{n_2-1}{N-2}R_{1,G} + \frac{n_1-1}{N-2}R_{2,G}$, $R_{d,G} = R_{1,G} - R_{2,G}$, and the two components are asymptotically independent under mild conditions (Chu and Chen, 2019). Let $\mathsf{E}_{\mathcal{G}_{C_0}}$ and $\mathsf{Var}_{\mathcal{G}_{C_0}}$ be the expectation and variance defined on the sample space $\mathcal{G}_{C_0}$ that places probability $1/|\mathcal{G}_{C_0}|$ on each $G \in \mathcal{G}_{C_0}$. Using the first component as an example: the averaging over all $G \in \mathcal{G}_{C_0}$ is essentially $\mathsf{E}_{\mathcal{G}_{C_0}}\left(\left(\frac{R_{w,G}-\mathsf{E}(R_{w,G})}{\sqrt{\mathsf{Var}(R_{w,G})}}\right)^2\right) = \left(\mathsf{E}_{\mathcal{G}_{C_0}}\left(\frac{R_{w,G}-\mathsf{E}(R_{w,G})}{\sqrt{\mathsf{Var}(R_{w,G})}}\right)\right)^2 + \mathsf{Var}_{\mathcal{G}_{C_0}}\left(\frac{R_{w,G}-\mathsf{E}(R_{w,G})}{\sqrt{\mathsf{Var}(R_{w,G})}}\right)$. Here,

$$\mathsf{Var}(R_{w,G}) = \frac{n_1 n_2 (n_1-1)(n_2-1)}{N(N-1)(N-2)(N-3)}\left(|G| - \frac{\sum_{i=1}^{N}|\mathcal{E}_i^G|^2}{N-2} + \frac{2|G|^2}{(N-1)(N-2)}\right)$$

contains $\sum_{i=1}^{N}|\mathcal{E}_i^G|^2$, which varies across different G's in $\mathcal{G}_{C_0}$. So it is already difficult to derive analytic tractable expression even only for $\mathsf{E}_{\mathcal{G}_{C_0}}\left(\frac{R_{w,G}-\mathsf{E}(R_{w,G})}{\sqrt{\mathsf{Var}(R_{w,G})}}\right)$. To get around the issues, we extend the generalized and weighted edge-count test based on how they were introduced in Chen et al. (2018) and Chen and Friedman (2017), respectively, through the extended quantities derived in Section 2.4. In the following, we first discuss the extended weighted edge-count test, and then the extended generalized edge-count test. The key components in the extended generalized edge-count test further compose the extended max-type edge-count test.

3.1 Extended weighted edge-count tests

As mentioned in Section 2.3, for data without repeated observations, there is a variance boosting problem for the edge-count test under unbalanced sample sizes. To solve the issue, Chen et al. (2018) proposed a weighted edge-count test $R_{w,G}$ (see definition in (2.2)). When there are repeated observations, the above problem also exists for the extended edge-count test (see Supplement 2.2). Following the similar idea, we could weight $R_{1,(a)}$ and $R_{2,(a)}$, and $R_{1,(u)}$ and $R_{2,(u)}$ to solve the problem. Under the union approach, the statistics $R_{1,(u)}$ and $R_{2,(u)}$ are simplified versions of R_1 and R_2 defined on $\bar{G}$, so the weights should be the same, i.e.,

$$R_{w,(u)} = (1-\hat{p})R_{1,(u)} + \hat{p}R_{2,(u)} \text{ with } \hat{p} = \frac{n_1 - 1}{N-2}. \tag{3.1}$$

However, for the average approach, the weights are not this straightforward. The following theorem shows that the weights for the average approach should also be the same.

Theorem 1. *For all test statistics of the form* $aR_{1,(a)} + bR_{2,(a)}$, $a+b=1$, $a, b > 0$, *we have* $\mathsf{Var}(aR_{1,(a)} + bR_{2,(a)}) \geq \mathsf{Var}(R_{w,(a)})$, *where* $R_{w,(a)} = (1-\hat{p})R_{1,(a)} + \hat{p}R_{2,(a)}$ *with* $\hat{p} = \frac{n_1-1}{N-2}$.

Proof. It is not hard to see that the minimum is achieved at

$$\hat{p} = \frac{\mathsf{Var}(R_{1,(a)}) - \mathsf{Cov}(R_{1,(a)}, R_{2,(a)})}{\mathsf{Var}(R_{1,(a)}) + \mathsf{Var}(R_{2,(a)}) - 2\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})}. \tag{3.2}$$

Plugging $\mathsf{Var}(R_{1,(a)})$, $\mathsf{Var}(R_{2,(a)})$ and $\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})$ provided in Supplement 1.4 into (3.2), we have $\hat{p} = \frac{n_1-1}{N-2}$. □

In the following lemma, we provide exact analytic formulas to the expectation and variance of $R_{w,(u)}$ and $R_{w,(a)}$, respectively, so that both extended weighted edge-count tests can be standardized easily.

Lemma 2. *The expectation and variance of $R_{w,(u)}$ and $R_{w,(a)}$ under the permutation null are:*

$$\mathsf{E}(R_{w,(u)}) = |\bar{G}|\frac{(n_1-1)(n_2-1)}{(N-1)(N-2)},$$

$$\mathsf{Var}(R_{w,(u)}) = \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{|\bar{G}| - \frac{1}{N-2}\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2 + \frac{2}{(N-1)(N-2)}|\bar{G}|^2\right\},$$

$$\mathsf{E}(R_{w,(a)}) = (N - K + |C_0|)\frac{(n_1-1)(n_2-1)}{(N-1)(N-2)},$$

$$\mathsf{Var}(R_{w,(a)}) = \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{-\frac{4}{N-2}\left(\sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \frac{(|C_0|-K)^2}{N}\right)\right.$$
$$\left. + 2(K - \sum_u \frac{1}{m_u}) + \sum_{(u,v)\in C_0}\frac{1}{m_u m_v} - \frac{2}{N(N-1)}(|C_0| + N - K)^2\right\},$$

where $|\mathcal{E}_i^{\bar{G}}| = m_u - 1 + \sum_{\mathcal{V}_u} m_v$ if observation i is of value indexed by u, and $|\bar{G}| = \sum_{u=1}^{K} m_u(m_u-1)/2 + \sum_{(u,v)\in C_0} m_u m_v$. Here, $\mathcal{V}_u^{C_0}$ is the set of distinct values that connect to the distinct value indexed by u in C_0.

This lemma can be proved straightforwardly by 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 $\mathsf{Cov}(R_{1,(u)}, R_{2,(u)})$ provided in Supplement 1.4.

3.2 Extended generalized edge-count tests

As we discussed earlier, it is technically intractable to derive the analytic expression for the average of S_G's for $G \in \mathcal{G}_{C_0}$. Here, we define extended generalized edge-count test statistic based on how the statistic was introduced in Chen and Friedman (2017) through the extended basic quantities:

$$S_{(a)} = \begin{pmatrix} R_{1,(a)} - \mathsf{E}(R_{1,(a)}) \\ R_{2,(a)} - \mathsf{E}(R_{2,(a)}) \end{pmatrix}^T \boldsymbol{\Sigma}_{(a)}^{-1} \begin{pmatrix} R_{1,(a)} - \mathsf{E}(R_{1,(a)}) \\ R_{2,(a)} - \mathsf{E}(R_{2,(a)}) \end{pmatrix}, \tag{3.3}$$

$$S_{(u)} = \begin{pmatrix} R_{1,(u)} - \mathsf{E}(R_{1,(u)}) \\ R_{2,(u)} - \mathsf{E}(R_{2,(u)}) \end{pmatrix}^T \boldsymbol{\Sigma}_{(u)}^{-1} \begin{pmatrix} R_{1,(u)} - \mathsf{E}(R_{1,(u)}) \\ R_{2,(u)} - \mathsf{E}(R_{2,(u)}) \end{pmatrix}, \tag{3.4}$$

where $\Sigma_{(a)} = \mathsf{Var}(\binom{R_{1,(a)}}{R_{2,(a)}})$, $\Sigma_{(u)} = \mathsf{Var}(\binom{R_{1,(u)}}{R_{2,(u)}})$. With similar arguments in Chen and Friedman (2017), $S_{(a)}$ and $S_{(u)}$ defined in this way could deal with location and scale alternatives. More studies on the performance of the tests are in Section 4. Similar to S_G, $S_{(a)}$ and $S_{(u)}$ defined above can also be decomposed to components that are asymptotically independent under mild conditions, respectively (details see Theorems 3 and 4).

Theorem 2. *The extended generalized edge-count test statistics can be ex-*

pressed as

$$S_{(a)} = \left(\frac{R_{w,(a)} - E(R_{w,(a)})}{\sqrt{Var(R_{w,(a)})}} \right)^2 + \left(\frac{R_{d,(a)} - E(R_{d,(a)})}{\sqrt{Var(R_{d,(a)})}} \right)^2, \quad (3.5)$$

$$S_{(u)} = \left(\frac{R_{w,(u)} - E(R_{w,(u)})}{\sqrt{Var(R_{w,(u)})}} \right)^2 + \left(\frac{R_{d,(u)} - E(R_{d,(u)})}{\sqrt{Var(R_{d,(u)})}} \right)^2, \quad (3.6)$$

where $R_{w,(a)}$, $E(R_{w,(a)})$, $Var(R_{w,(a)})$, $R_{w,(u)}$, $E(R_{w,(u)})$ *and* $Var(R_{w,(u)})$ *are provided in Section 3.1, and* $R_{d,(a)} = R_{1,(a)} - R_{2,(a)}$, $R_{d,(u)} = R_{1,(u)} - R_{2,(u)}$ *with their expectations and variances provided below.*

$$E(R_{d,(a)}) = (N - K + |C_0|)\tfrac{n_1 - n_2}{N},$$

$$Var(R_{d,(a)}) = \tfrac{4n_1n_2}{N(N-1)} \left\{ \sum_u \tfrac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \tfrac{(|C_0|-K)^2}{N} \right\},$$

$$E(R_{d,(u)}) = |\bar{G}|\tfrac{n_1 - n_2}{N},$$

$$Var(R_{d,(u)}) = \tfrac{n_1n_2}{N(N-1)} \left\{ \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|^2 - \tfrac{4}{N}|\bar{G}|^2 \right\}.$$

Theorem 2 is proved in Supplement 1.1.

3.3 Extended max-type edge-count test statistics

Let $Z_{w,(a)} = \frac{R_{w,(a)} - \mathsf{E}(R_{w,(a)})}{\sqrt{\mathsf{Var}(R_{w,(a)})}}$, $Z_{d,(a)} = \frac{R_{d,(a)} - \mathsf{E}(R_{d,(a)})}{\sqrt{\mathsf{Var}(R_{d,(a)})}}$, $Z_{w,(u)} = \frac{R_{w,(u)} - \mathsf{E}(R_{w,(u)})}{\sqrt{\mathsf{Var}(R_{w,(u)})}}$, and $Z_{d,(u)} = \frac{R_{d,(u)} - \mathsf{E}(R_{d,(u)})}{\sqrt{\mathsf{Var}(R_{d,(u)})}}$. Under some mild conditions, $Z_{w,(a)}$ and $Z_{d,(a)}$ are asymptotically independent with their joint distribution bivariate normal, and same for $Z_{w,(u)}$ and $Z_{d,(u)}$ (details see Theorems 3 and 4). Here,

we define the extended max-type edge-count statistics:

$$M_{(a)}(\kappa) = \max(\kappa Z_{w,(a)}, |Z_{d,(a)}|), \text{ and } M_{(u)}(\kappa) = \max(\kappa Z_{w,(u)}, |Z_{d,(u)}|).$$

As the following arguments hold the same for the averaging and the union statistics, we omit subscripts (a) and (u) for simplicity. From the definition of the extended max-type edge-count test statistic, we can see that it makes use of both Z_w and Z_d, and would be similar to S_G and effective to both location and scale alternatives. Also, the introduction of κ in the definition makes it more flexible than S_G.

We here briefly discuss the choice of κ. It is easy to see that the rejection region $\{M(\kappa) \geq \beta\}$ is equivalent to $\{Z_w \geq \frac{\beta}{\kappa}$ or $|Z_d| \geq \beta\}$. Let $\mathsf{P}(Z_w \geq \beta_w) = \alpha_1$ and $\mathsf{P}(|Z_d| \geq \beta_d) = \alpha_2$, and define $\gamma = \frac{\alpha_1}{\alpha_2}$. Based on the asymptotic distribution of $(Z_w, Z_d)^T$ derived in Section 5, the relationship between γ and κ with the overall type I error rate controlled at 0.05 is shown in Table 3.

Table 3: Relationship between γ and κ.

γ	8	4	2	1	1/2	1/4	1/8
κ	1.63	1.47	1.31	1.14	1	0.88	0.79

To check how the choice of κ affects the performance of the test, we examine the test on 100-dimensional multivariate normal distributions $\mathcal{N}_d(\mu_1, \Sigma_1)$

and $\mathcal{N}_d(\mu_2, \Sigma_2)$ that are different in mean and/or variance. Three scenarios are considered and the detailed results are presented in Supplement 3.2. Based on the simulation results, if there is no prior knowledge about the type of difference between the two distributions, we recommend $\kappa = \{1.31, 1.14, 1\}$ for $M(\kappa)$.

4. Performance of the extended test statistics

In this section, we study the performance of various tests through the ranking problems, 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. We consider the following two data generating mechanisms.

(i) Data are genearated from the probability model shown in Section 3.1

$$\mathsf{P}_{\theta,\eta}(\zeta) = \frac{1}{\psi(\theta)} \exp\{-\theta d(\zeta, \eta)\}, \quad \zeta, \eta \in \Xi, \quad \theta \in \mathbf{R}, \tag{4.1}$$

where Ξ be the set of all permutations of the set {1,2,3,4,5,6} and $d(\cdot, \cdot)$ is a distance function such as Kendall's or Spearman's distance. The two samples are generated from $\mathsf{P}_{\theta_1,\eta_1}(\cdot)$ and $\mathsf{P}_{\theta_2,\eta_2}(\cdot)$, respectively.

(ii) Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two different subsets of all possible rankings. The two sample are generated from the uniform distribution on $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively.

When Kendall's or Spearman's distance is used for $d(\cdot,\cdot)$, there are in general ties in the distance matrix, which lead to non-unique MSTs. Hence, we apply 3-NNL to construct the graph on distinct values. The results for Kendall's and Spearman's distance are very similar, so we present the results based on the Spearman's distance in the following.

We compare the following statistics: $R_{0,(a)}$, $R_{0,(u)}$, $S_{(a)}$, $S_{(u)}$, $R_{w,(a)}$, $R_{w,(u)}$, $M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$ ($\kappa = 1.31, 1.14, 1$) in six scenarios (Scenarios 1–3 under (i) and Scenarios 4–6 under (ii)) with balanced and unbalanced sample sizes. The settings with both different θ and different η under (i) are also considered and the results can be found in Supplement 3.1. In each scenario, the specific parameters under each scenario are chosen such that the tests have moderate power to be comparable.

- Scenario 1 (Only η differs) : $\eta_1 = \{1,2,3,4,5,6\}$, $\eta_2 = \{1,2,5,4,3,6\}$, $\theta_1 = \theta_2 = 5$ with balanced ($n_1 = n_2 = 100$) and unbalance ($n_1 = 100, n_2 = 400$) sample sizes.

- Scenario 2 (Only θ differs with $\theta_1 > \theta_2$) : $\eta_1 = \eta_2 = \{1,2,3,4,5,6\}$, $\theta_1 = 5.5$, $\theta_2 = 4$ with balanced ($n_1 = n_2 = 300$) and unbalance ($n_1 = 300, n_2 = 600$) sample sizes.

- Scenario 3 (Only θ differs with $\theta_1 < \theta_2$) : $\eta_1 = \eta_2 = \{1,2,3,4,5,6\}$,

$\theta_1 = 4$, $\theta_2 = 5.5$ with balanced ($n_1 = n_2 = 300$) and unbalance ($n_1 = 300, n_2 = 600$) sample sizes.

- Scenario 4 (Different supports): $\mathcal{D}_1 = \{\zeta \in \Xi : \zeta$ does not begin with No.6$\}$, $\mathcal{D}_2 = \{\zeta \in \Xi : \zeta$ does not end with No.1$\}$ with balanced ($n_1 = n_2 = 150$) and unbalance ($n_1 = 150, n_2 = 250$) sample sizes.

- Scenario 5 (Different supports): $\mathcal{D}_1 = \{\zeta \in \Xi : \zeta$ ranks No.1 before No.5$\}$, $\mathcal{D}_2 = \{\zeta \in \Xi : \zeta$ ranks No.1 before No.6$\}$ with balanced ($n_1 = n_2 = 150$) and unbalance ($n_1 = 150, n_2 = 250$) sample sizes.

- Scenario 6 (Different supports): $\mathcal{D}_1 = \{\zeta \in \Xi : \zeta$ does not begin with No.6 and does not end with No.1$\}$, $\mathcal{D}_2 = \{\zeta \in \Xi : \zeta$ ranks No.1 or No.2 in top 3$\}$ with balanced ($n_1 = n_2 = 150$) and unbalance ($n_1 = 150, n_2 = 250$) sample sizes.

The results are presented in Tables 4 and 5. Each table lists the fraction of trials (out of 1000) that the test reject the null hypothesis at 0.05 significance level. Those above 95 percentage of the best power under each setting are in bold.

Table 4 provides results for the data generated by mechanism (i). We see that $S_{(u)}$ and $M_{(u)}$ work well for all scenarios, while the others show obvious strengthes and weaknesses for different settings. For example, un-

Table 4: Results of the extended test statistics for the data generated by mechanism (i).

Scenario 1. $n_1 = n_2 = 100$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.857**	0.750	**0.857**	0.831	0.813	0.780
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.888**	0.791	**0.888**	**0.861**	0.840	0.818

Scenario 1. $n_1 = 100$, $n_2 = 400$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.641	0.889	**0.949**	**0.940**	**0.935**	0.915
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.871	**0.951**	**0.977**	**0.969**	**0.961**	**0.959**

Scenario 2. $n_1 = n_2 = 300$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.265	0.172	0.265	0.239	0.223	0.194
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.438	**0.796**	0.438	0.767	**0.797**	**0.828**

Scenario 2. $n_1 = 300$, $n_2 = 600$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.525	0.325	0.310	0.348	0.334	0.318
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0	**0.899**	0.566	**0.887**	**0.912**	**0.929**

Scenario 3. $n_1 = n_2 = 300$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.279	0.181	0.279	0.250	0.231	0.208
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.413	0.755	0.413	0.730	**0.781**	**0.806**

Scenario 3. $n_1 = 300$, $n_2 = 600$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.061	0.378	0.355	0.393	0.393	0.386
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.954**	0.899	0.545	0.874	**0.909**	**0.922**

Table 5: Results of the extended test statistics for the data generated by mechanism (ii).

Scenario 4. $n_1 = n_2 = 150$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.745**	0.557	**0.745**	0.695	0.646	0.594
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.670	0.503	0.670	0.626	0.580	0.528

Scenario 4. $n_1 = 150, n_2 = 250$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.826	0.744	**0.881**	0.834	0.804	0.767
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.782	0.637	0.783	0.746	0.714	0.668

Scenario 5. $n_1 = n_2 = 150$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.620**	0.447	**0.620**	0.573	0.528	0.468
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.502	0.387	0.502	0.470	0.450	0.415

Scenario 5. $n_1 = 150, n_2 = 250$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.840**	0.743	**0.880**	**0.841**	0.815	0.790
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.834	0.661	0.698	0.692	0.683	0.647

Scenario 6. $n_1 = n_2 = 150$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.886**	0.763	**0.886**	**0.858**	0.828	0.788
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.814	0.681	0.814	0.774	0.745	0.708

Scenario 6. $n_1 = 150$, $n_2 = 250$

Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	**0.943**	**0.916**	**0.962**	**0.944**	**0.938**	**0.928**
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.888	0.821	**0.917**	0.895	0.885	0.852

der the unbalanced setting ($n_1 = 300, n_2 = 600$), $R_{0,(u)}$ has no power under Scenario 2, $R_{0,(a)}$ has very low power under Scenario 3, and both $R_{w,(a)}$ and $R_{w,(u)}$ do not perform well when only θ differs (Scenarios 2 and 3). Overall, $M_{(u)}(\kappa)$ perform best among all the tests. When θ differs, $S_{(a)}$ and $S_{(u)}$ provide similar results to $M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$, respectively, but they perform worse than $M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$, respectively, when only η differs (Scenario 1). In general, the tests based on "union" are slightly better than their "averaging" counterparts (except for some cases for R_0).

Table 5 provides results for data generated by mechanism (ii). We see that the tests perform similarly well with those based on "averaging" slightly better than their "union" counterparts.

Remark 1. For either the "averaging" statistics and the "union" statistics, their relationships can be represented by the following schematic plots on the reject regions in terms of Z_w and Z_d. In general, Z_w aims for detecting location alternative and Z_d aims for detecting scale alternative, so the extended generalized edge-count test and the extended max-type edge-count test are effective on both alternatives. On the other hand, if we know in prior that the difference is only in mean, then the extended weighted edge-count tests are preferred.

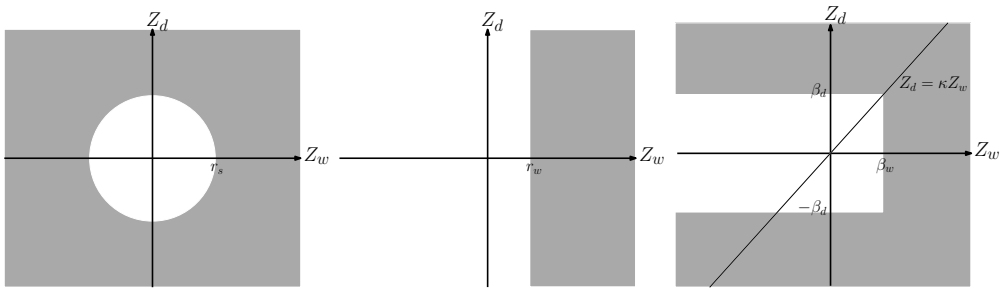


Figure 3: Rejection regions (in gray) of S_G, $R_{w,G}$, $M(\kappa)$. Left: $\{S_G \geq r_s^2\}$; middle: $\{Z_w \geq r_w\}$; right: $\{M(\kappa) \geq \beta\}$.

5. Asymptotics

In this section, we provide the asymptotic distributions of new test statistics described in Section 3. This provides us theoretical bases for obtaining analytic p-value approximation. The examination of how well these approximations work for finite samples is provided in Supplement 3.3. In the following, we use $a = O(b)$ to denote that a and b are of the same order and $a = o(b)$ to denote that a is of a smaller order than b. Let $\mathcal{E}_{i,2}^G$ be the set of edges in G that contain at least one node in $\mathcal{V}_i^G$.

5.1 Statistics based on averaging

To derive the asymptotic behavior of the statistics based on averaging $(R_{w,(a)}, S_{(a)}, M_{(a)}(\kappa))$, we work under the following conditions:

Condition 1.

$$|C_0|, \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} = O(N),$$
$$K, \sum_u \frac{1}{m_u} = O(N^\alpha), \quad \alpha \le 1.$$

Condition 2.

$$\sum_u m_u(m_u + |\mathcal{E}_u^{C_0}|)(m_u + \sum_{v\in\mathcal{V}_u^{C_0}} m_v + |\mathcal{E}_{u,2}^{C_0}|) = o(N^{3/2}),$$

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

Condition 3.

$$\sum_u \frac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \frac{(|C_0|-K)^2}{N} = O(N).$$

Remark 2. One special case for Condition 1 is

$$|C_0|, \sum_{(u,v)\in C_0} \frac{1}{m_u m_v}, K, \sum_u \frac{1}{m_u} = O(N). \tag{5.1}$$

Conditions (5.1) and 2 are the same conditions stated in Chen and Zhang (2013) in obtaining the asymptotic properties of $R_{0,(a)}$ and $R_{0,(u)}$. Condition 1 is easy to be satisfied and Condition 2 sets constrains on the number of repeated observations and the degrees of nodes in the graph C_0 such that they cannot be too large.

When $m_u \equiv m$ for all u, Condition 2 can be simplified to

$$\sum_u |\mathcal{E}_u^{C_0}||\mathcal{E}_{u,2}^{C_0}| = o(N^{3/2})$$

and

$$\sum_{(u,v)\in C_0} (|\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|)(|\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|) = o(N^{3/2}).$$

The additional condition (Condition 3) makes sure that $(R_1, R_2)^T$ does not degenerate asymptotically. When $m_u \equiv m$ for all u, Condition 3 becomes

$$\frac{1}{4m}\sum_u |\mathcal{E}_u^{C_0}|^2 - \frac{|C_0|^2}{mK} = \frac{1}{4m}\sum_u (|\mathcal{E}_u^{C_0}| - \frac{2|C_0|}{K})^2 = O(N),$$

which is the variance of the degrees of nodes in C_0. When there is not enough variety in the degrees of nodes in C_0, the correlation between R_1 and R_2 tends to 1. (A similar condition is needed for the continuous counterpart (Chen and Friedman, 2017).)

Theorem 3. *Under Conditions 1, 2 and 3, as $N \to \infty$,*

$$\begin{pmatrix} Z_{w,(a)} \\ Z_{d,(a)} \end{pmatrix} \xrightarrow{D} \mathcal{N}_2(0, \mathbf{I}_2)$$

under the permutation null distribution.

The proof of this theorem is in Supplement 1.2. Based on Theorem 3, it is easy to obtain the asymptotic distributions of $S_{(a)}$ and $M_{(a)}(\kappa)$.

Corollary 1. *Under Conditions 1, 2 and 3, as $N \to \infty$, $S_{(a)} \xrightarrow{D} \mathcal{X}_2^2$ under the permutation null distribution.*

Corollary 2. *Under Conditions 1, 2 and 3, the asymptotic cumulative distribution function of* $M_{(a)}(\kappa)$ *is* $\Phi(\frac{x}{\kappa})(2\Phi(x)-1)$ *under the permutation null distribution, where* $\Phi(x)$ *denotes the cumulative distribution function of the standard normal distribution.*

5.2 Statistics based on taking union

To derive the asymptotic behavior of the statistics based on taking union $(R_{w,(u)}, S_{(u)}, M_{(u)}(\kappa))$, we work under the following conditions:

Condition 4.

$$|\bar{G}| = O(N).$$

Condition 5.

$$\sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|^2 - \frac{4}{N}|\bar{G}|^2 = O(N).$$

Condition 6.

$$\sum_{u=1}^{K} m_u^3 (m_u + \sum_{v \in \mathcal{V}_u^{C_0}} m_v) \sum_{v \in \{u\} \cup \mathcal{V}_u^{C_0}} m_v (m_v + \sum_{w \in \mathcal{V}_v^{C_0}} m_w) = o(N^{3/2}),$$

$$\sum_{(u,v) \in C_0} m_u m_v \left[m_u (m_u + \sum_{w \in \mathcal{V}_u^{C_0}} m_w) + m_v (m_v + \sum_{w \in \mathcal{V}_v^{C_0}} m_w) \right] \cdot \left[\sum_{\substack{w \in \{u\} \cup \{v\} \cup \mathcal{V}_u^{C_0} \cup \mathcal{V}_v^{C_0} \\ y \in \mathcal{V}_w^{C_0}}} m_w (m_w + m_y) \right] = o(N^{3/2}).$$

Remark 3. Condition 4 is easy to satisfy. Condition 5 was mentioned in Chen and Friedman (2017) in the continuous version. When $m_u \equiv m$ for all u, Condition 5 could be rewritten as

$$\sum_{u=1}^{K} |\mathcal{E}_u^{C_0}|^2 - \frac{4}{K}|C_0|^2 = O(K).$$

If C_0 is the k-MST, $k = O(1)$, constructed under Euclidean distance, the above condition always holds based on results in Chen and Friedman (2017).

When $m_u \equiv m$ for all u, Condition 6 becomes

$$\sum_{u} |\mathcal{E}_u^{C_0}||\mathcal{E}_{u,2}^{C_0}| = o(N^{3/2})$$

and

$$\sum_{(u,v)\in C_0} (|\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|)(|\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|) = o(N^{3/2}),$$

which are the same as the simplified form in Remark 2. These conditions restrict the degrees of nodes in graph C_0.

Theorem 4. *Under Conditions 4, 5 and 6, as $N \to \infty$,*

$$\begin{pmatrix} Z_{w,(u)} \\ Z_{d,(u)} \end{pmatrix} \xrightarrow{D} \mathcal{N}_2(0, \mathbf{I}_2),$$

under the permutation null distribution.

The proof of this theorem is in Supplement 1.3. Based on Theorem 4, it is easy to obtain the asymptotic distributions of $S_{(u)}$ and $M_{(u)}(\kappa)$.

Corollary 3. *Under Conditions 4, 5 and 6, as* $N \to \infty$, $S_{(u)} \xrightarrow{D} \mathcal{X}_2^2$ *under the permutation null distribution.*

Corollary 4. *Under Conditions 4, 5 and 6, the asymptotic cumulative distribution function of* $M_{(u)}(\kappa)$ *is* $\Phi(\frac{x}{\kappa})(2\Phi(x)-1)$ *under the permutation null distribution, where* $\Phi(x)$ *denotes the cumulative distribution function of the standard normal distribution.*

6. Phone-call network data analysis

In this section, we analyze the phone-call network data mentioned in Section 1 in details. We first present the test results of various statistics, and then examine the analytic p-value approximations through this real data example.

The MIT Media Laboratory conducted a study following 106 subjects, including students and staffs in an institute, who used mobile phones with pre-installed software that can record call logs. The study lasted from July 2004 to June 2005 (Eagle et al. (2009)). Given the richness of this dataset, many problems can be studied. One question of interest is whether phone call patterns on weekdays are different from those on weekends. The phone calls on weekdays and weekends can be viewed as representations of professional relationship and personal relationship, respectively.

We bin the phone calls by day and, for each day, construct a directed phone-call network with the 106 subjects as nodes and a directed edge pointing from person i to person j if person i made one call to person j on that day. We encode the directed network of each day by an adjacency matrix, with 1 for element $[i, j]$ if there is a directed edge pointing from subject i to subject j, and 0 otherwise.

The original dataset was sorted in the calendar order with 236 weekdays and 94 weekends. Among the 330 (236+94) networks, there are 285 distinct values and 11 of them have more than one observations. We denote the distinct values as matrices $B_1, \cdots, B_{285}$. We adopt the distance measure used in Chen and Friedman (2017) and Chen et al. (2018), which is defined as the number of different entries, i.e.,

$$d(B_i, B_j) = \|B_i - B_j\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix. Besides the repeated observations, there are many equal distances among distinct values. We set C_0 to be the 3-NNL, which has similar density as the 9-MST recommended in Chen et al. (2018).

Table 6 lists the results. In particular, we list the values, expectation (Mean) and standard deviations (SD) of $R_{1,(a)}$, $R_{1,(u)}$, $R_{2,(a)}$, $R_{2,(u)}$, $(R_{1,(a)} + R_{2,(a)})/2$, $(R_{1,(u)} + R_{2,(u)})/2$, $R_{w,(a)}$, $R_{w,(u)}$, $R_{d,(a)}$ and $R_{d,(u)}$, as well as the

Table 6: Breakdown statistics of the phone-call network data.

	Value	Mean	Value-Mean	SD
$R_{1,(a)}$	2800.26	2669.56	130.70	143.33
$R_{2,(a)}$	409.18	420.80	-11.62	57.75
$(R_{1,(a)} + R_{2,(a)})/2$	1604.72	1545.18	59.54	44.74
$R_{w,(a)}$	1087.14	1058.40	28.73	11.79
$R_{d,(a)}$	2391.08	2248.76	142.32	199.37

	Value	Mean	Value-Mean	SD
$R_{1,(u)}$	7163.00	6860.35	302.65	381.50
$R_{2,(u)}$	1008.00	1081.38	-73.38	151.66
$(R_{1,(u)} + R_{2,(u)})/2$	4085.50	3970.86	114.64	116.22
$R_{w,(u)}$	2753.17	2719.93	33.24	15.65
$R_{d,(u)}$	6155.00	5778.97	376.03	532.03

		Value	p-Value			Value	p-Value
$Z_{0,(a)}$		-1.33	0.092	$Z_{0,(u)}$		-0.99	0.162
$S_{(a)}$		6.45	0.040	$S_{(u)}$		5.01	0.082
$Z_{w,(a)}$		2.44	0.007	$Z_{w,(u)}$		2.12	0.017
$\lvert Z_{d,(a)}\rvert$		0.71	0.475	$\lvert Z_{d,(u)}\rvert$		0.71	0.480
$M_{(a)}(\kappa)$	$\kappa = 1.31$	3.19	0.009	$M_{(u)}(\kappa)$	$\kappa = 1.31$	2.78	0.022
	$\kappa = 1.14$	2.78	0.013		$\kappa = 1.14$	2.42	0.032
	$\kappa = 1$	2.44	0.022		$\kappa = 1$	2.12	0.050

values and p-values of $Z_{0,(a)}$, $Z_{0,(u)}$, $S_{(a)}$, $S_{(u)}$, $Z_{w,(a)}$, $Z_{w,(u)}$, $|Z_{d,(a)}|$, $|Z_{d,(u)}|$, $M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$, where $Z_{0,(a)}$ and $Z_{0,(u)}$ are standardizations for $R_{0,(a)}$ and $R_{0,(u)}$, respectively. Note that the tests based on $(R_{1,(a)} + R_{2,(a)})/2$, and $(R_{1,(u)} + R_{2,(u)})/2$ are equivalent to those based on $R_{0,(a)}$ and $R_{0,(u)}$, respectively.

We first check results based on "averaging". We can see that $R_{1,(a)}$ is much higher than its expectation, while $R_{2,(a)}$ is smaller than its expecta-

tion. The original edge-count test $R_{0,(a)}$ is equivalent to adding $R_{1,(a)}$ and $R_{2,(a)}$ directly, so the signal in $R_{1,(a)}$ is diluted by $R_{2,(a)}$. In addition, due to the variance boosting issue, it fails to reject the null hypothesis at 0.05 significance level. On the other hand, the weighted edge-count test chooses the proper weight to minimize the variance and performs well. Since $S_{(a)}$ and $M_{(a)}(\kappa)$ consider the weighted edge-count statistic and the difference of two with-in sample edge-counts simultaneously, these tests all reject the null at 0.05 significance level. The larger the κ is, the more similar the max-type test ($M_{(a)}(\kappa)$) and the weighted test ($R_{w,(a)}$) are. So the p-values of $M_{(a)}(\kappa)$ are very close to that of $R_{w,(a)}$, when κ is large. The results on the "union" counterparts are similar, except that $S_{(u)}$ cannot reject the null at 0.05 significance level. Based on the information in the table, it is clear that there is mean difference between the two samples, while no significant scale difference between the two samples.

We also check the analytic p-values obtained based on asymptotical results with those based on 10,000 random permutations and the results are shown in Table 7. We can see that the asymptotic p-values and the permutation p-values are quite close for all test statistics.

Table 7: 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

7. Conclusion

The generalized edge-count test and the weighted edge-count test are useful tools in two-sample testing regime. Both tests rely on a similarity graph constructed on the pooled observations from the two samples and can be applied to various data types as long as a reasonable similarity measure on the sample space can be defined. However, they are problematic when the similarity graph is not uniquely defined, which is common for data with repeated observations. In this work, we extend them as well as a max-type statistic, to accommodate scenarios when the similarity graph cannot be uniquely defined. The extended test statistics are equipped with easy-to-evaluate analytic expressions, making them easy to compute in real data analysis. The asymptotic distributions of the extended test statistics are also derived and simulation studies show that the p-values obtained based on asymptotic distributions are quite accurate under sample sizes in hundreds

and beyond, making these tests easy-off-the-shelf tools for large data sets.

Among the extended edge-count tests, the extended weighted edge-count tests aim for location alternatives, and the extended generalized/max-type edge-count tests aim for more general alternatives. When these tests do not reach a consensus, a detailed analysis illustrated by the phone-call network data in Section 6 is recommended.

Supplementary Materials

The supplementary material contains proofs of lemmas and theorems, and some additional results.

Acknowledgements

Jingru Zhang is supported in part by the CSC scholarship. Hao Chen is supported in part by NSF award DMS-1513653.

References

Bai, Z. and H. Saranadasa (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 311–329.

Cai, T. T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association 108*(501), 265–277.

Cai, T. T., W. Liu, and Y. Xia (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(2), 349–372.

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, H. and N. R. Zhang (2013). Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica*, 1479–1503.

Chen, S. X., Y.-L. Qin, et al. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics 38*(2), 808–835.

Chu, L. and H. Chen (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of Statistics 47*(1), 382–414.

Eagle, N., A. S. Pentland, and D. Lazer (2009). Inferring friendship net-

work structure by using mobile phone data. *Proceedings of the National Academy of Sciences 106*(36), 15274–15278.

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

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 772–783.

Li, J. and S. X. Chen (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics 40*(2), 908–940.

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.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association 81*(395), 799–806.

Schott, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Computational Statistics & Data Analysis 51*(12), 6535–6542.

Srivastava, M. S. and M. Du (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis 99*(3), 386–402.

Srivastava, M. S. and H. Yanagihara (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *Journal of Multivariate Analysis 101*(6), 1319–1329.

Xia, Y., T. Cai, and T. T. Cai (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika 102*(2), 247–266.

Xu, G., L. Lin, P. Wei, and W. Pan (2016). An adaptive two-sample test for high-dimensional means. *Biometrika 103*(3), 609–624.

Beijing International Center for Mathematical Research, Five Yiheyuan Road, Beijing, 100871, P.R.China

E-mail: (jingruzhang@pku.edu.cn)

Department of Statistics, University of California, Davis, One Shields Avenue, Davis, California 95616, USA

E-mail: (hxchen@ucdavis.edu)

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

Jingru Zhang and Hao Chen

Beijing International Center for Mathematical Research and University of California, Davis

Supplementary Material

1. Proofs for lemmas and theorems

1.1 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:

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

$$\begin{pmatrix} \mathsf{Var}(R_{1,(a)}) & \mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) \\ \mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) & \mathsf{Var}(R_{2,(a)}) \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}} \frac{n_2-1}{N-2} & \frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} \\ \frac{1}{\sqrt{\mathsf{Var}(R_{w,(a)})}} \frac{n_1-1}{N-2} & -\frac{1}{\sqrt{\mathsf{Var}(R_{d,(a)})}} \end{pmatrix}$$

$$\triangleq \begin{pmatrix} ① & ② \\ ② & ③ \end{pmatrix},$$

where

$$① = \frac{\mathsf{Var}(R_{1,(a)})(n_2-1)^2 + 2\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})(n_1-1)(n_2-1) + \mathsf{Var}(R_{2,(a)})(n_1-1)^2}{\mathsf{Var}(R_{w,(a)})(N-2)^2} = 1,$$

$$② = \frac{(n_2-1)(\mathsf{Var}(R_{1,(a)}) - \mathsf{Cov}(R_{1,(a)}, R_{2,(a)})) + (n_1-1)(\mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) - \mathsf{Var}(R_{2,(a)}))}{(N-2)\sqrt{\mathsf{Var}(R_{d,(a)})\mathsf{Var}(R_{w,(a)})}}$$

$$= 0, \text{ and}$$

$$③ = \frac{\mathsf{Var}(R_{1,(a)}) - 2\mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) + \mathsf{Var}(R_{2,(a)})}{\mathsf{Var}(R_{d,(a)})} = 1.$$

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 Appendix 1.4, we can obtain the expectations and variances of $R_{d,(a)}$ and $R_{d,(u)}$. □

1.2 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 1.1. [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 1.1,*

we have

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

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

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

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 1.1.*

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 $\mathsf{P_B}, \mathsf{E_B}, \mathsf{Var_B}$ 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 Supplement 1.4, we have

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

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 $\mathsf{Var}_{\mathsf{B}}(W_3^B) = 1$, 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 $W = a_1 W_1^B + a_2 W_2^B + a_3 W_3^B$ is asymptotically Gaussian distributed for any combination of a_1, a_2, a_3 such that $\mathsf{Var}_{\mathsf{B}}(W) > 0$.

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,

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

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

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

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

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

$$\xi_i = \begin{cases} a_1\xi_u^{(1)} + a_2\xi_u^{(2)}, & \text{if } i = u \in \mathcal{J}_1, \\ a_1\xi_{uv}^{(1)} + a_2\xi_{uv}^{(2)}, & \text{if } i = uv \in \mathcal{J}_2, \\ a_3\xi_u^{(3)}, & \text{if } i = u \in \mathcal{J}_3. \end{cases} \tag{1.1}$$

We introduce following index sets to satisfy Assumption 1.1.

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

$$S_u = \{u, u + |\mathcal{J}_0|\} \cup \{uv, vu : (u, v) \in C_0\},$$

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

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

$$S_{uv} = \{uv, u, v, u + |\mathcal{J}_0|, v + |\mathcal{J}_0|\} \cup \{uw, wu : (u, w) \in C_0\}$$
$$\cup \{vw, wv : (v, w) \in C_0\},$$

$$T_{uv} = S_{uv} \cup \{w, w + |\mathcal{J}_0|, wy, yw : (u, w), (w, y) \in C_0\}$$
$$\cup \{w, w + |\mathcal{J}_0|, wy, yw : (v, w), (w, y) \in C_0\}.$$

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

$$S_u = \{u, u'\} \cup \{u'v, vu' : (u', v) \in C_0\}, \quad u' = u - |\mathcal{J}_0|,$$

$$T_u = S_u \cup \{v, v + |\mathcal{J}_0|, vw, wv : (u', v), (v, w) \in C_0\}.$$

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 (1.1),

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

Hence,

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

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

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

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

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

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

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

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

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

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 $b_1, b_3, b_4 \in (0,\infty)$; $b_2, b_5 \in [0,1]$, $b_2 \le b_5$. 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,

$$\mu_1^B - \mu_1 = (N - K + |C_0|)(p_4 - p_1) = (N - K + |C_0|)\frac{n_1 n_2}{N^2(N-1)},$$

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

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

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

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

□

1.3 Proof of Theorem 4

Proof. We will use Assumption 1.1 and Theorem 1 in Proof 1.2. The proof is similar to Proof 1.2, 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 Supplement 1.4, 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

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

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 1.2. 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 $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_3$,

$$\tilde{\xi}_i = \begin{cases} a_1\tilde{\xi}_u^{(1)} + a_2\tilde{\xi}_u^{(2)}, & \text{if } i = u \in \mathcal{J}_1, \\ a_1\tilde{\xi}_{uv}^{(1)} + a_2\tilde{\xi}_{uv}^{(2)}, & \text{if } i = uv \in \mathcal{J}_2, \\ a_3\tilde{\xi}_u^{(3)}, & \text{if } i = u \in \mathcal{J}_3. \end{cases} \tag{1.2}$$

The definition of index sets $(S_u, T_u, S_{uv}, T_{uv})$ are same as those in Proof 1.2.

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

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

Similarly,

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

Since $n_{1u'} \in [0, m_{u'}], p_a \in [0,1]$, we have $|\tilde{\xi}_u^{(3)}| \le \frac{m_{u'}}{\delta_0} \le \frac{m_{u'}}{\sigma}$ with $u \in \mathcal{J}_3, u' = u - |\mathcal{J}_0|$. Plugging these inequations into (1.2),

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

Hence,

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

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

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

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

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

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

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

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

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

$$(\delta_1^B)^2 = \frac{n_1^2 n_2^2}{N^4}\left\{|\bar{G}| + \frac{n_1}{n_2}\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2\right\}.$$

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 $N \to \infty$,

$$\frac{(\delta_1)^2}{N} \longrightarrow p^2q^2\left\{d_1 + \frac{p}{q}d_2\right\},$$

$$\frac{(\delta_1^B)^2}{N} \longrightarrow p^2q^2\left\{d_1 + \frac{p}{q}(4d_1^2 + d_2)\right\},$$

$$\frac{\delta_1^B}{\delta_1} \longrightarrow \sqrt{\frac{d_1q + (4d_1^2+d_2)p}{d_1q + d_2p}} = \sqrt{1 + \frac{4d_1^2p}{d_1q + d_2p}}.$$

Similarly, we have

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

Also,

$$\nu_1^B - \nu_1 = |\bar{G}|(p_4 - p_1) = |\bar{G}|\frac{n_1 n_2}{N^2(N-1)},$$

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 $N \to \infty$,

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

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

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

Last, with some computation as in the last part of Proof 1.2, we obtain

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

□

1.4 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*

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

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

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

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

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

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

we have

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

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.

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

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

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

$\mathsf{Var}(R_{1,(a)})$ 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

$$\begin{aligned} &\mathsf{Cov}(R_{1,(a)}, R_{2,(a)}) \\ =&\mathsf{Cov}(X_{1A}+X_{1B}, X_{2A}+X_{2B}) \\ =&\mathsf{Cov}(X_{1A}, X_{2A}) + \mathsf{Cov}(X_{1B}, X_{2B}) \\ &+ \mathsf{Cov}(X_{1B}, X_{2A}) + \mathsf{Cov}(X_{1B}, X_{2B}). \end{aligned} \tag{1.3}$$

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 (1.3). 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

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

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

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

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

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

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

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

Plugging these equations into (1.3), 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*

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

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

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

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

$$\mathsf{Cov}(R_{1,(u)}, R_{2,(u)}) = f_1\left[|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}| - 1)\right] - p_1q_1|\bar{G}|^2.$$

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

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

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

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

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

$$\begin{aligned}
\mathsf{Cov}(R_{1,(u)}, R_{2,(u)}) &= \mathsf{E}(R_{1,(u)}R_{2,(u)}) - \mathsf{E}(R_{1,(u)})\mathsf{E}(R_{2,(u)}) \\
&= f_1\left[|\bar{G}|^2 - |\bar{G}| - \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|(|\mathcal{E}_i^{\bar{G}}|-1)\right] - p_1q_1|\bar{G}|^2.
\end{aligned}$$

□

2. Issues of existing graph-based tests

2.1 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).

2.2 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

3. Additional results

3.1 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	**0.848**	0.754	**0.848**	0.821	0.805	0.778
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.884**	**0.865**	**0.884**	**0.883**	**0.879**	**0.863**

$n_1 = 100, n_2 = 300$						
Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.790	0.888	**0.948**	**0.940**	**0.925**	0.912
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	0.493	**0.952**	**0.970**	**0.965**	**0.965**	**0.954**

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	**0.888**	0.778	**0.888**	0.854	0.834	0.805
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.917**	**0.873**	**0.917**	**0.898**	**0.890**	0.870

$n_1 = 100, n_2 = 300$						
Statistic	$R_{0,(a)}$	$S_{(a)}$	$R_{w,(a)}$	$M_{(a)}(1.31)$	$M_{(a)}(1.14)$	$M_{(a)}(1)$
Estimated Power	0.813	0.917	**0.962**	**0.954**	**0.947**	0.935
Statistic	$R_{0,(u)}$	$S_{(u)}$	$R_{w,(u)}$	$M_{(u)}(1.31)$	$M_{(u)}(1.14)$	$M_{(u)}(1)$
Estimated Power	**0.996**	**0.993**	**0.985**	**0.986**	**0.986**	**0.989**

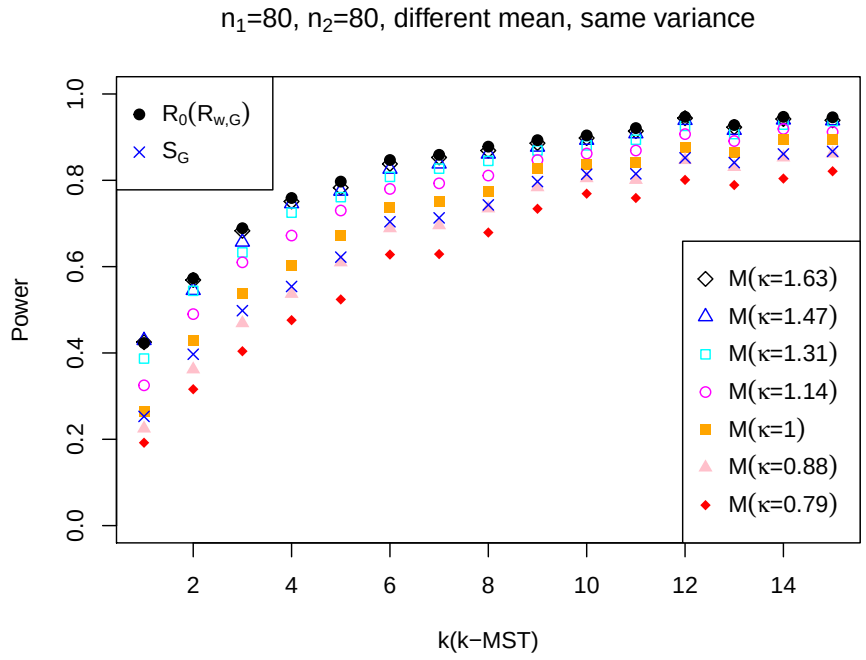
3.2 Choice of κ for max-type edge-count test statistics

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.



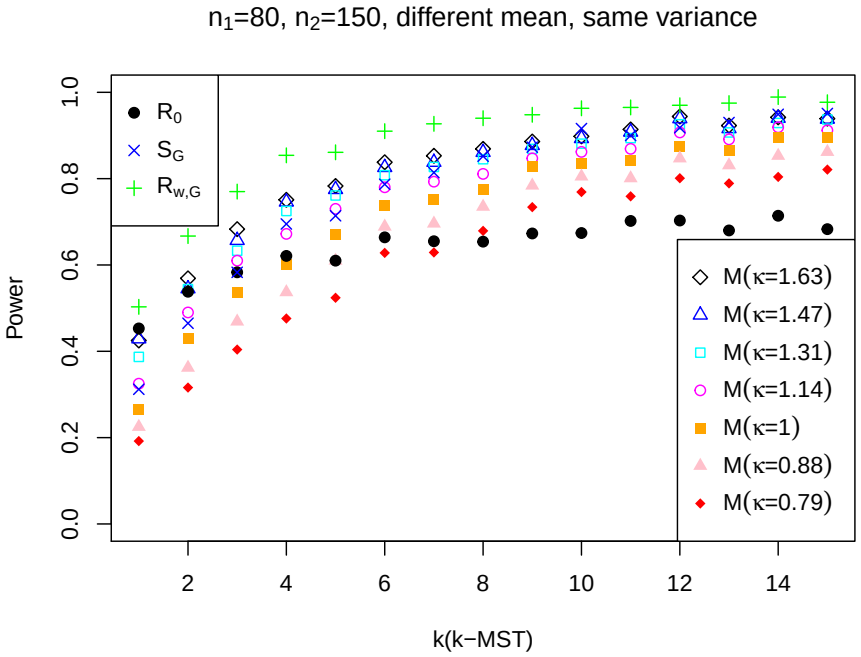
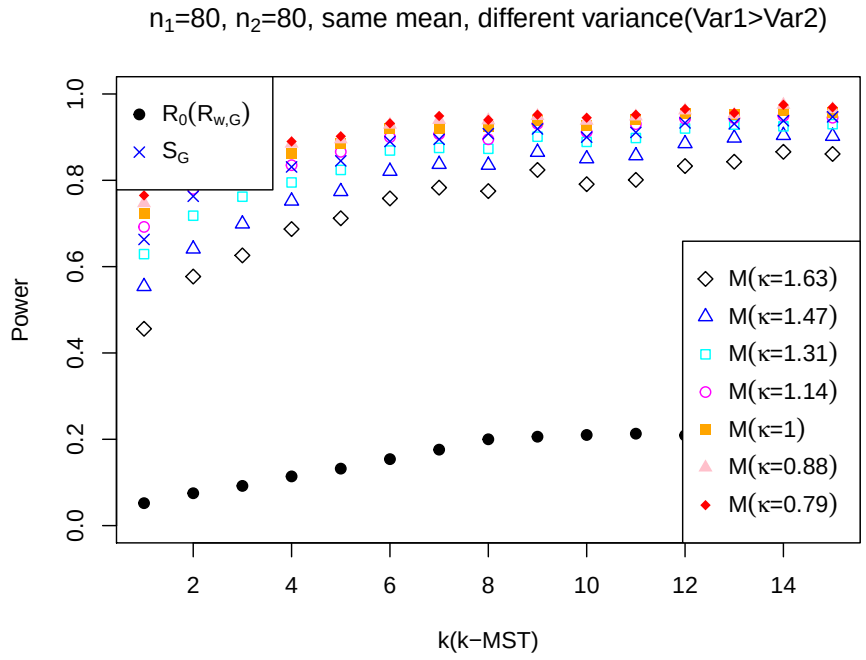


Figure 1: Under Scenario 1, the fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level.



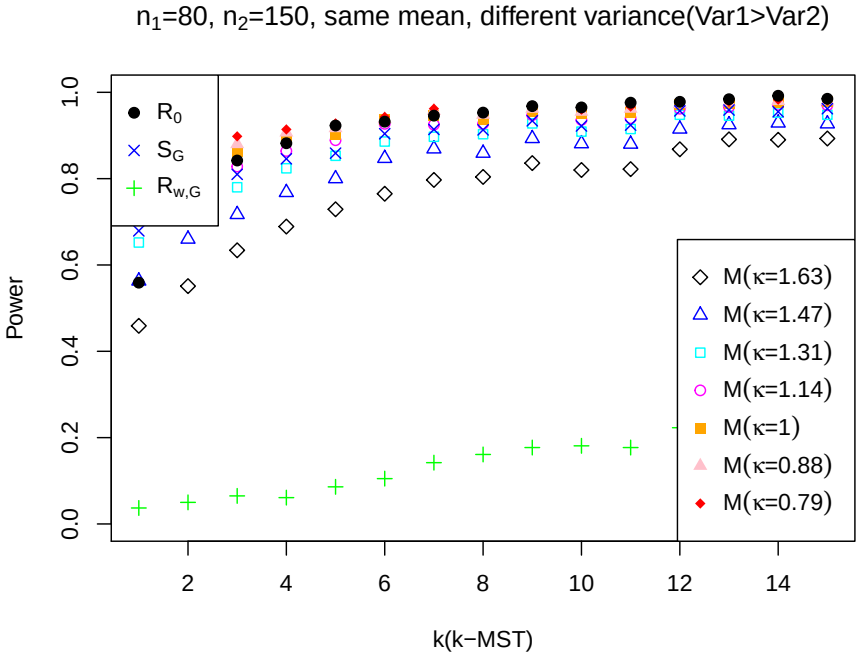
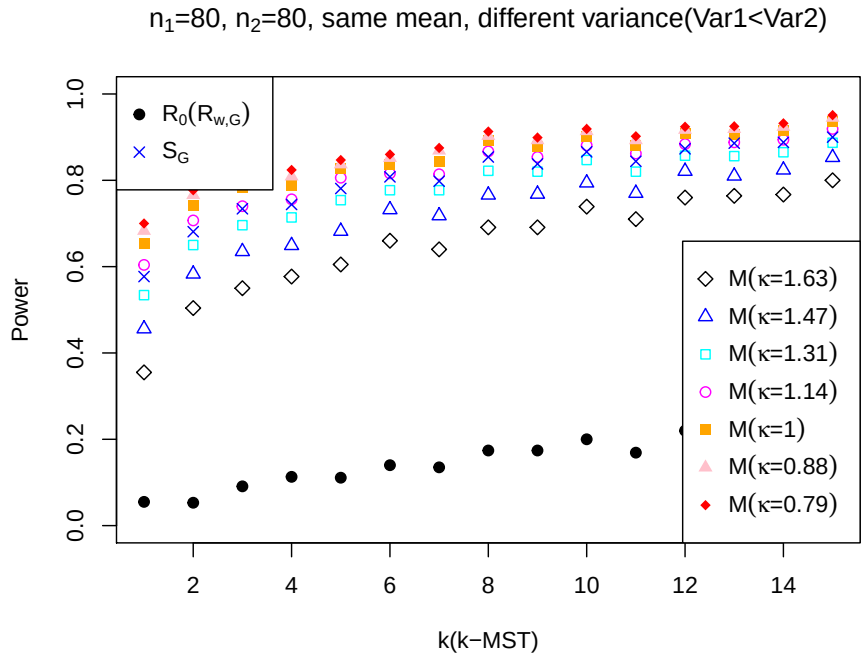


Figure 2: Under Scenario 2, the fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level.



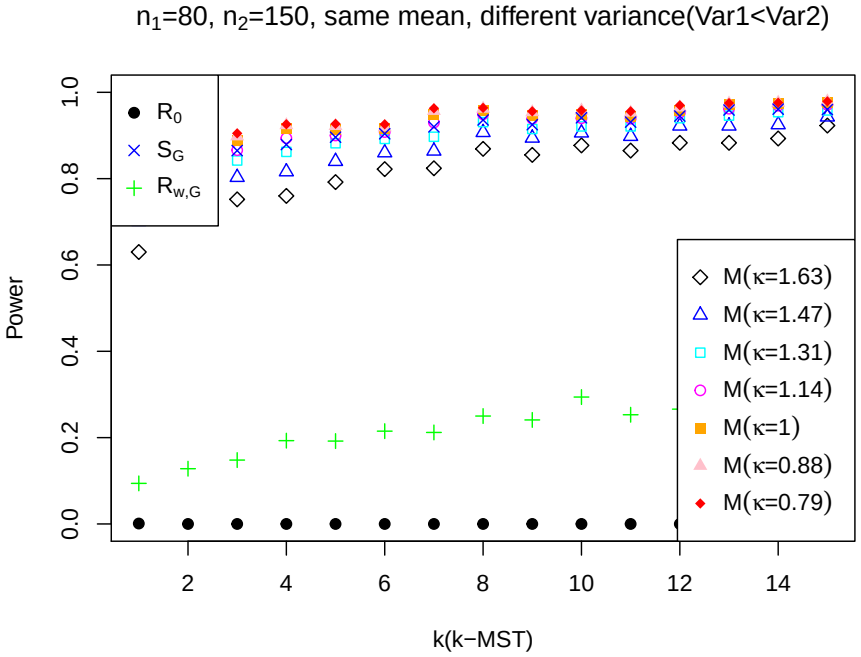


Figure 3: Under Scenario 3, the fraction of trials (out of 1000) that the test rejected the null hypothesis at 0.05 significance level.

3.3 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.

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.

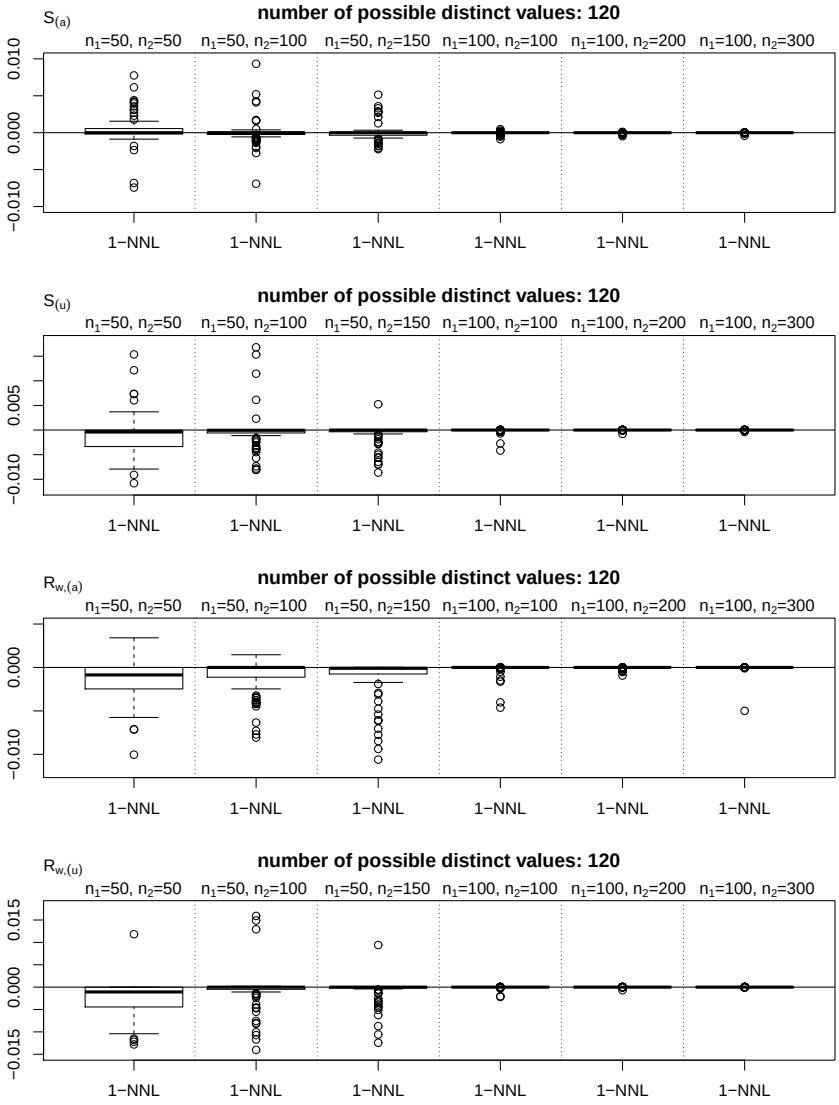


Figure 4: Boxplots for the differences between the asymptotic p-value and the permutation p-value based on 100 simulation runs under each setting for $S_{(a)}, S_{(u)}, R_{w,(a)}$ and $R_{w,(u)}$.

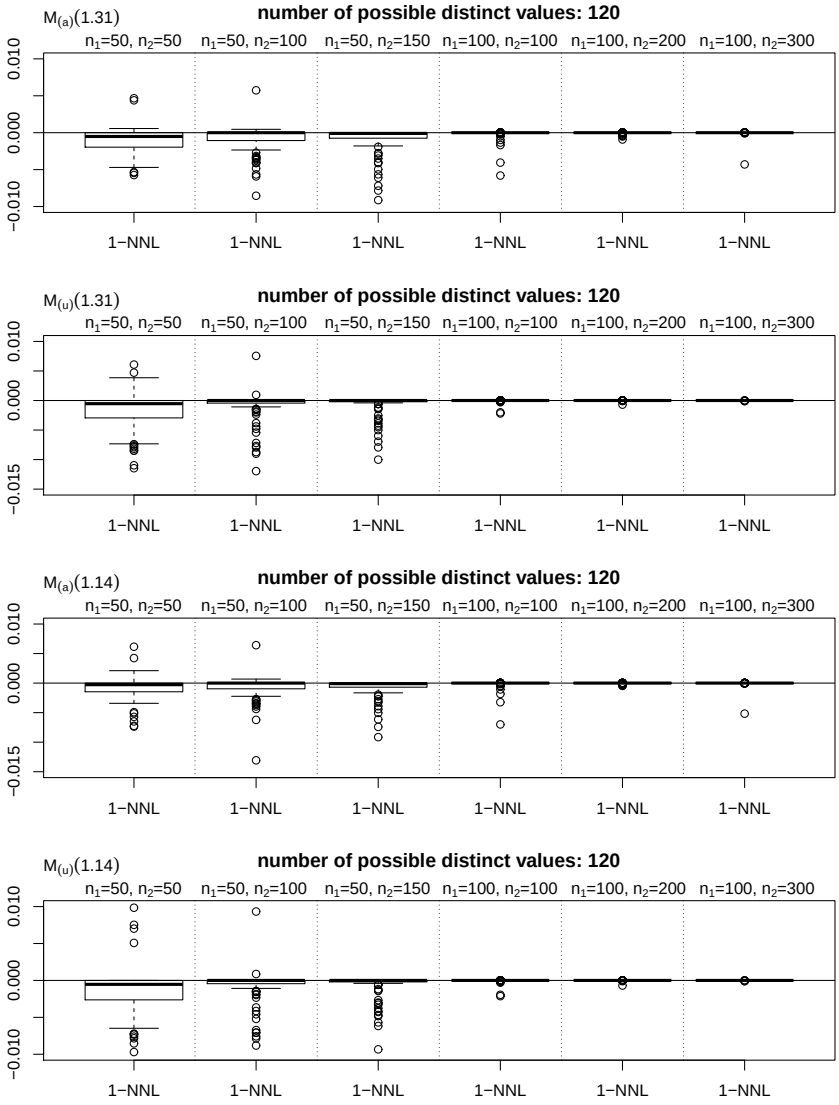


Figure 5: Boxplots for the differences between the asymptotic p-value and the permutation p-value based on 100 simulation runs under each setting for $M_{(a)}(\kappa)$ and $M_{(u)}(\kappa)$ with $\kappa = 1.31, 1.14$.

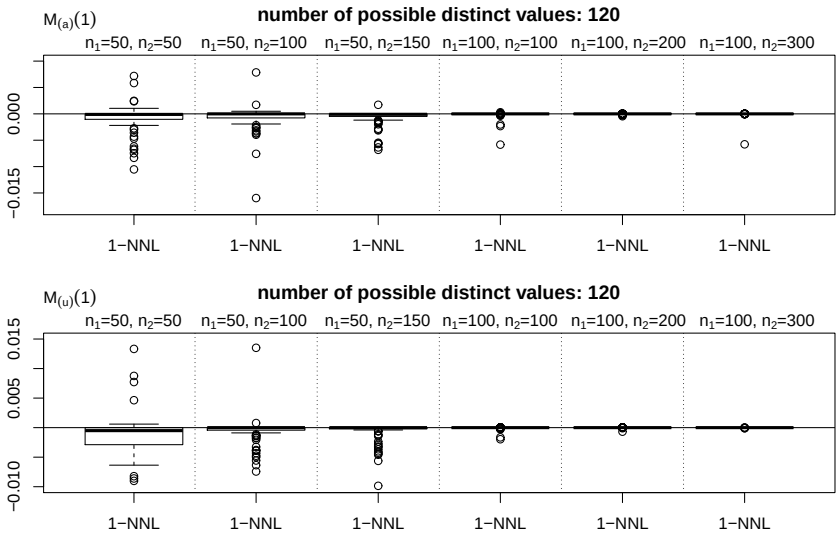


Figure 6: Boxplots for the differences between the asymptotic p-value and the permutation p-value based on 100 simulation runs under each setting for $M_{(a)}(1)$ and $M_{(u)}(1)$.

References

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.