Statistica Sinica **32** (2022), 391-415
doi:https://doi.org/10.5705/ss.202019.0116

# GRAPH-BASED TWO-SAMPLE TESTS FOR DATA WITH REPEATED OBSERVATIONS

Jingru Zhang and Hao Chen

*University of Pennsylvania and University of California, Davis*

*Abstract:* For two-sample comparisons, tests based on graphs constructed using the similarity information between observations are gaining attention, owing to their flexibility and good performance for high-dimensional/non-Euclidean data. However, when there are repeated observations, these graph-based tests can be problematic, because they are influenced by the choice of the similarity graph. We propose extended graph-based test statistics to resolve this problem. We also study the asymptotic properties of these extended statistics, and provide analytic formulae to approximate the $p$-values of the tests under finite samples, facilitating the application of the new tests in practice. The proposed tests are applied to analyze a phone-call network data set. All tests are implemented in the `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 comparisons present a fundamental problem in statistics, and have been studied extensively for univariate and low-dimensional data. However, research on the testing problem for high-dimensional and non-Euclidean data, such as network data, is gaining attention with the advent of big data. In the parametric domain for multivariate data, many studies have tested whether the means are the same (e.g., Srivastava and Du (2008)) and whether the covariance matrices are the same (e.g., Schott (2007); Srivastava and Yanagihara (2010); Xia, Cai and Cai (2015)). To improve their applicability, many semiparametric methods have been proposed to test means and covariance matrices (e.g., Bai and Saranadasa (1996); Chen and Qin (2010); Cai, Liu and Xia (2014); Xu et al. (2016); Li and Chen (2012); Cai, Liu and Xia (2013)) by adding conditions on the moment and/or covariance, rather than making assumptions about the underlying distributions. These parametric and semiparametric methods provide

Corresponding author: Hao Chen, Department of Statistics, University of California, Davis, CA 95616, USA. E-mail: hxchen@ucdavis.edu.

useful tools when the data follow their assumptions, but are often restrictive and not sufficiently robust if the model assumptions are violated.

In the nonparametric domain, researchers have extended the Kolmogorov–Smirnov test, Wilcoxon rank test, and Wald–Wolfowitz runs test to include high-dimensional data (see Chen and Friedman (2017) for a review). Of these, the first practical test was proposed by Friedman and Rafsky (1979) as an extension of the Wald–Wolfowitz runs test for multivariate data. They pool the observations from the two samples and construct a minimum spanning tree (MST) that connects all observations, minimizing the sum of the distances of the edges in the tree. They then count the number of edges in the MST that connect observations from different samples, and reject the null hypothesis of equal distributions if this count is significantly *smaller* than its expectation under the null hypothesis. This test was later extended to other similarity graphs in which observations that are closer together are more likely to be connected than those that are further apart. These extensions include the minimum distance pairing (MDP) of Rosenbaum (2005) and the nearest neighbor graph (NNG) of Schilling (1986) and Henze (1988). We refer to this type of tests as an *edge-count test*. Recently, a generalized edge-count test and a weighted edge-count test were proposed to address the problems of the original edge-count test under scale alternatives and unequal sample sizes (Chen and Friedman (2017); Chen, Chen and Su (2018)). Because these tests and the edge-count test are all based on a similarity graph, we call them *graph-based tests*. These tests have many advantages. They can be applied to data with an arbitrary dimension and to non-Euclidean data, and exhibit high power when detecting differences in distribution. They also have higher power than that of the likelihood-based tests when the dimension of the data is moderate to high for practical sample sizes (i.e., from hundreds to millions).

However, graph-based tests can be problematic for data with repeated observations. These tests all 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 can be equally "optimal" in terms of the criterion. Furthermore, the results of the graph-based tests can vary 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 are provided in the Supplementary Material, Section S2.1).

In this work, we seek ways to effectively summarize the tests over these equally "optimal" similarity graphs. As we show in Section 2.2, it is not uncommon to have more than a million equally optimal similarity graphs when there

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 using 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) |

are repeated 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, owing to the quadratic terms in the generalized edge-count test statistic, doing so is not feasible (see Section 3). However, we can 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 similarly to how they were designed for continuous data. Our results are as follows:

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

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

Based on (2), we study an extended max-type edge-count test that builds upon the two asymptotically independent normal random variables. The tests are implemented in the `R` package `gTests`.

The rest of the paper is organized as follows. Section 2 provides the notation used in the paper and preliminary setups. Section 3 discusses 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 by using them to analyze a phone-call network data set. Section 7 concludes the paper.

## 2. Notation and Preliminary Setup

### 2.1. Notation

Among the $N$ observations, we assume there are $K$ distinct values, indexed by $1, 2, \ldots, K$. The basic notation is summarized in Table 2. Here, $m_i = n_{1i}+n_{2i}$,

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 |

for $i = 1, \ldots, K$, $n_i = \sum_{k=1}^{K} n_{ik}$, for $i = 1, 2$, and $N = n_1 + n_2$.

Let $d(i, j)$ be the distance between the values indexed by $i$ and $j$. 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 an $n_1!n_2!/N!$ 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 $\mathsf{E}$, $\mathsf{Var}$, $\mathsf{Cov}$, and $\mathsf{Cor}$ to denote the expectation, variance, covariance, and correlation, respectively, under the permutation null distribution.

### 2.2. Similarity graphs on observations

Let $C_0$ be a similarity graph constructed on the distinct values. This can be the MST, MDP, or NNG on the distinct values, if it can be uniquely defined. If the common optimization rules do not result in a unique solution, we follow Chen and Zhang (2013) and use the union of all MSTs. Figure 1 is a simple example. The union of all MSTs on the distinct values can be obtained using Algorithm 1. For example, for the data in Figure 1, the distinct values **a** and **b**, **a** and **c**, **b** and **c**, and **d** and **e** are connected in the first step, and **b** and **d** and **c** and **e** are connected in the second step. We call this graph the nearest neighbor link (NNL). If one wants denser graphs, $k$-NNL can be considered, which is the union of the 1st,$\ldots$, $k$th NNLs, where the $j$th ($j > 1$) NNL is a graph generated by Algorithm 1, subject to the constraint that this graph does not contain any edge in the 1st,$\ldots$, $(j-1)$th NNLs.

**Algorithm 1** Generate a NNL

For each distinct value indexed by $i$ ($i = 1, \ldots, K$), 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**

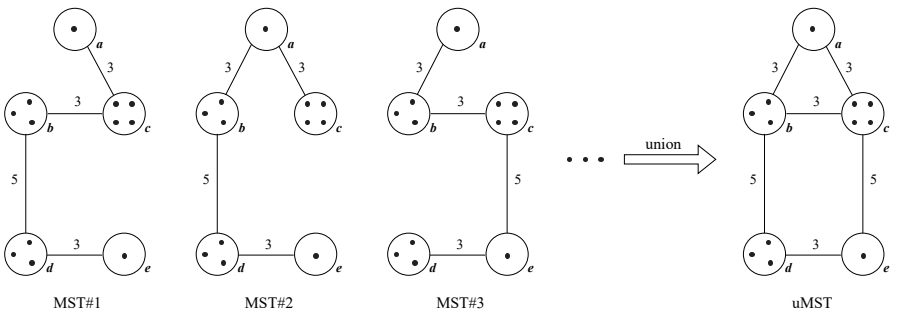


Figure 1. There are five distinct values (**a**, **b**, **c**, **d**, **e**), denoted by circles. Some distinct values have more than one observation, denoted by having more than one point in the circle. The distances between the distinct values are denoted on the edges. It is clear that there are six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

Then, a graph on the 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, indexed by $i$, and another observation, indexed by $j$, and connect the two. Then, for each $i$, if there is more than one observation indexed by $i$, connect these observations using a spanning tree (any edge in a spanning tree has distance zero). Let $\mathcal{G}_{C_0}$ be the set of all graphs initiated from $C_0$.

For the example in Figure 1, because the MST on the distinct values is not uniquely defined, let $C_0$ be the NNL. There are 15,552($= 1^2 \cdot 3^3 \cdot 4^3 \cdot 3^2 \cdot 1^2$) ways of assigning the six edges in $C_0$ to corresponding observations in each circle. In addition, by Cayley's lemma, for the observations equal to the same 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.

### 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 note 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, owing 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:

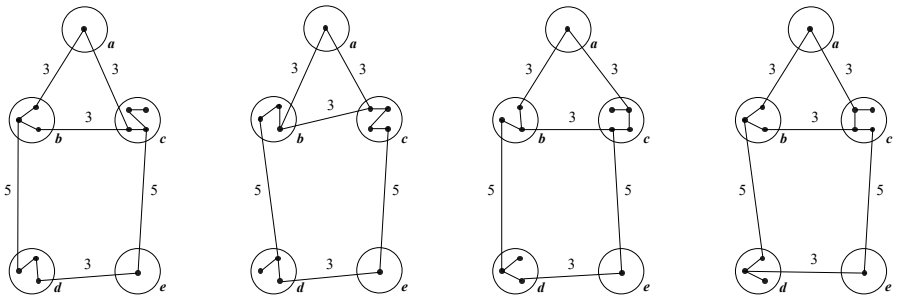


Figure 2. Four graphs, out of 2,239,488, on observations initiated from the NNL on distinct values.

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

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

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 a 5-MST, to boost their power (Friedman and Rafsky (1979); Chen and Friedman (2017)). Here, a $k$-MST is the union of the 1st,..., $k$th MSTs, where the first MST is the MST, and the $j$th ($j > 1$) 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,..., $(j-1)$th MSTs. However, Chen, Chen and Su (2018) found that, for a $k$-MST ($k > 1$), the edge-count test ($R_{0,G}$) behaves strangely when the two sample sizes are different. For example, consider the testing problem in which the two underlying distributions are $\mathcal{N}_d(0, \mathbf{I}_d)$ and $\mathcal{N}_d(\mu, \mathbf{I}_d)$ ($\|\mu\|_2 = 1.3,\ d = 50$), and we have 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 (see Chen, Chen and Su (2018)). To solve this issue, Chen, Chen and Su (2018) proposed a weighted edge-count test that inversely weights the within-sample edges using the sample sizes

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

The authors reason 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 should be down-weighted. This weighted edge-count test statistic addresses 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 consider two ways of summarizing the test statistics for $R_{0,G}$:

(1) averaging: $R_{0,(a)} = (1/|\mathcal{G}_{C_0}|)\sum_{G\in\mathcal{G}_{C_0}} R_{0,G}$, 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 $\bar{G}_{C_0} = \cup\{G \in \mathcal{G}_{C_0}\}$; that is, 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.

When there are many graphs in $\mathcal{G}_{C_0}$, it is often not feasible to compute these two quantities directly. Chen and Zhang (2013) derived analytic expressions to compute 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 can define $R_{1,(a)}$, $R_{1,(u)}$, $R_{2,(a)}$, and $R_{2,(u)}$ and their analytic expressions in terms of the summary quantities in Table 2 and $C_0$, as shown in Lemma 1.

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

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

$$R_{1,(u)} \equiv R_{1,\bar{G}_{C_0}} = \sum_{u=1}^{K} \frac{n_{1u}(n_{1u}-1)}{2} + \sum_{(u,v)\in C_0} n_{1u}n_{1v},$$

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

$$R_{2,(u)} \equiv R_{2,\bar{G}_{C_0}} = \sum_{u=1}^{K} \frac{n_{2u}(n_{2u}-1)}{2} + \sum_{(u,v)\in C_0} n_{2u}n_{2v}.$$

The notation $\{n_{ik}\}_{i=1,2;\ k=1,\ldots,K}$, $\{m_k\}_{k=1,\ldots,K}$ is defined in Table 2. These analytic expressions are obtained using similar arguments to those in Chen and Zhang (2013), and thus are omitted here.

## 3. Extended Graph-Based Tests

Because the generalized edge-count test can cover a wider range of alternatives than the original edge-count test can (Chen and Friedman (2017)), we would like the generalized edge-count test statistic to be well defined when there are repeated observations. For the generalized edge-count test statistic, $S_G = \begin{pmatrix} R_{1,G} - \mathsf{E}(R_{1,G}) \\ R_{2,G} - \mathsf{E}(R_{2,G}) \end{pmatrix}^T \mathbf{\Sigma}_G^{-1} \begin{pmatrix} R_{1,G} - \mathsf{E}(R_{1,G}) \\ R_{2,G} - \mathsf{E}(R_{2,G}) \end{pmatrix}$, one straightforward way of defining the average statistic is $(1/|\mathcal{G}_{C_0}|)\sum_{G\in\mathcal{G}_{C_0}} S_G$. However, $\mathbf{\Sigma}_G$ varies with $G$ in $\mathcal{G}_{C_0}$, making the averaging over $S_G$ difficult. Even in the simplified version in which $\Sigma_G$ is fixed over $G$ in $\mathcal{G}_{C_0}$, the quadratic terms in $S_G$ make the averaging analytically intractable. To view the problem more straightforwardly, note that $S_G$ can be written as $S_G = ((R_{w,G} - \mathsf{E}(R_{w,G}))/\sqrt{\mathsf{Var}(R_{w,G})})^2 + ((R_{d,G} - \mathsf{E}(R_{d,G}))/\sqrt{\mathsf{Var}(R_{d,G})})^2$, where $R_{w,G} = ((n_2-1)/(N-2))R_{1,G} + ((n_1-1)/(N-2))R_{2,G}$ and $R_{d,G} = R_{1,G} - R_{2,G}$. Let $\mathsf{E}_{\mathcal{G}_{C_0}}$ and $\mathsf{Var}_{\mathcal{G}_{C_0}}$ be the expectation and variance, respectively, defined on the sample space $\mathcal{G}_{C_0}$ that places the 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 $\mathsf{E}_{\mathcal{G}_{C_0}}(((R_{w,G} - \mathsf{E}(R_{w,G}))/\sqrt{\mathsf{Var}(R_{w,G})})^2) = (\mathsf{E}_{\mathcal{G}_{C_0}}((R_{w,G} - \mathsf{E}(R_{w,G}))/\sqrt{\mathsf{Var}(R_{w,G})}))^2 + \mathsf{Var}_{\mathcal{G}_{C_0}}((R_{w,G} - \mathsf{E}(R_{w,G}))/\sqrt{\mathsf{Var}(R_{w,G})})$. 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 with $G$ in $\mathcal{G}_{C_0}$. Thus, it is already difficult to derive an analytically tractable expression, even for $\mathsf{E}_{\mathcal{G}_{C_0}}((R_{w,G} - \mathsf{E}(R_{w,G}))/\sqrt{\mathsf{Var}(R_{w,G})})$. To get around these issues, we extend the generalized and weighted edge-count tests based on how they were introduced in Chen and Friedman (2017) and Chen, Chen and Su (2018), respectively, using 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. Furthermore, the key components in the latter form 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 this issue, Chen, Chen and Su (2018) proposed a weighted edge-count test $R_{w,G}$ (see (2.2)). When there are repeated observations, the above problem also exists for the extended edge-count test (see the Supplementary Material, Section S2.2). Following a similar idea, we can 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 $R_2$, respectively, defined on $\bar{G}$, so the weights should be the same; that is,

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

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

**Theorem 1.** *For all test statistics of the form $aR_{1,(a)} + bR_{2,(a)}$, with $a + b = 1$, for $a, b > 0$, 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 $\hat{p} = (n_1 - 1)/(N - 2)$.*

*Proof.* The minimum is achieved at

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

Substituting $\mathsf{Var}(R_{1,(a)})$, $\mathsf{Var}(R_{2,(a)})$, and $\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})$ from the Supplementary Material S1.4 into (3.2), we have $\hat{p} = (n_1 - 1)/(N - 2)$.

In the following lemma, we provide exact analytic formulae for the expectation and variance of $R_{w,(u)}$ and $R_{w,(a)}$, so that both extended weighted edge-count tests can be standardized easily. This lemma can be proved straightforwardly by substituting in the analytic expressions for $\mathsf{E}(R_{1,(a)})$, $\mathsf{E}(R_{2,(a)})$, $\mathsf{Var}(R_{1,(a)})$, $\mathsf{Var}(R_{2,(a)})$, $\mathsf{Cov}(R_{1,(a)}, R_{2,(a)})$, $\mathsf{E}(R_{1,(u)})$, $\mathsf{E}(R_{2,(u)})$, $\mathsf{Var}(R_{1,(u)})$, $\mathsf{Var}(R_{2,(u)})$, and $\mathsf{Cov}(R_{1,(u)}, R_{2,(u)})$, as provided in the Supplementary Material S1.4.

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

$$\begin{aligned} \mathsf{E}(R_{w,(u)}) &= |\bar{G}|\frac{(n_1-1)(n_2-1)}{(N-1)(N-2)}, \\ \mathsf{Var}(R_{w,(u)}) &= \frac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)} \end{aligned}$$

$$\left\{|\bar{G}| - \frac{1}{N-2}\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2 + \frac{2}{(N-1)(N-2)}|\bar{G}|^2\right\},$$

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

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

*where* $|\mathcal{E}_i^{\bar{G}}| = m_u - 1 + \sum_{\mathcal{V}_u^{C_0}} m_v$ *if observation* $i$ *has a distinct value index* $u$*, 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* $C_0$*.*

### 3.2. Extended generalized edge-count tests

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

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

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

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

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

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

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

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

$$\begin{aligned}
\mathsf{E}(R_{d,(a)}) &= (N - K + |C_0|)\frac{n_1 - n_2}{N}, \\
\mathsf{Var}(R_{d,(a)}) &= \frac{4n_1 n_2}{N(N-1)} \left\{ \sum_u \frac{(|\mathcal{E}_u^{C_0}| - 2)^2}{4m_u} - \frac{(|C_0| - K)^2}{N} \right\}, \\
\mathsf{E}(R_{d,(u)}) &= |\bar{G}|\frac{n_1 - n_2}{N}, \\
\mathsf{Var}(R_{d,(u)}) &= \frac{n_1 n_2}{N(N-1)} \left\{ \sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|^2 - \frac{4}{N}|\bar{G}|^2 \right\}.
\end{aligned}$$

Theorem 2 is proved in the Supplementary Material S1.1.

### 3.3. Extended max-type edge-count test statistics

Let $Z_{w,(a)} = (R_{w,(a)} - \mathsf{E}(R_{w,(a)}))/\sqrt{\mathsf{Var}(R_{w,(a)})}$, $Z_{d,(a)} = (R_{d,(a)} - \mathsf{E}(R_{d,(a)}))/\sqrt{\mathsf{Var}(R_{d,(a)})}$, $Z_{w,(u)} = (R_{w,(u)} - \mathsf{E}(R_{w,(u)}))/\sqrt{\mathsf{Var}(R_{w,(u)})}$, and $Z_{d,(u)} = (R_{d,(u)} - \mathsf{E}(R_{d,(u)}))/\sqrt{\mathsf{Var}(R_{d,(u)})}$. Under some mild conditions, $Z_{w,(a)}$ and $Z_{d,(a)}$ are asymptotically independent and follow a joint bivariate normal distribution; the same is true for $Z_{w,(u)}$ and $Z_{d,(u)}$ (see Theorems 3 and 4). Here, we define the extended max-type edge-count statistics as follows:

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

Because the following arguments hold for the averaging and the union statistics, we omit the subscripts $(a)$ and $(u)$, for simplicity. From the definition of the extended max-type edge-count test statistic, we can see that it uses both $Z_w$ and $Z_d$, and is similar to $S_G$ and effective for both the location and the scale alternatives. In addition, the introduction of $\kappa$ in the definition makes it more flexible than $S_G$.

Table 3. Relationship between $\gamma$ and $\kappa$.

| $\gamma$ | 8 | 4 | 2 | 1 | 1/2 | 1/4 | 1/8 |
|---|---|---|---|---|---|---|---|
| $\kappa$ | 1.63 | 1.47 | 1.31 | 1.14 | 1 | 0.88 | 0.79 |

Table 4. Rejection regions for the extended statistics.

| Statistic | Reject region |
|---|---|
| Extended generalized edge-count tests | $S \geq r_s^2$ |
| Extended weighted edge-count tests | $\frac{R_w - \mathsf{E}(R_w)}{\sqrt{R_w}} \geq r_w$ |
| Extended max-type edge-count tests | $M(\kappa) \geq \beta(\kappa)$ |

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

To investigate how the choice of $\kappa$ 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 differ in terms of their mean and/or variance. Three scenarios are considered; detailed results are presented in the Supplementary Material S3.2. 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 $M(\kappa)$.

### 3.4. Testing rule

We summarize the rejection regions for the extended statistics in Table 4, which are similar to their continuous counterparts. Because the testing rule is the same for the averaging and the union statistics, we omit the subscripts $(a)$ and $(u)$, for simplicity. In the table, $r_s$, $r_w$, and $\beta(\kappa)$ are the critical values, which can be obtained by drawing random permutations or using the asymptotic distributions of the extended statistics (see Section 5).

Schematic plots of the rejection regions in terms of $Z_w$ and $Z_d$ are shown in Figure 3. We can see that these statistics are closely related. More detailed comparisons on these statistics are presented in following sections.

## 4. Performance of the Extended Test Statistics

In this section, we study the performance of various tests using simulation studies. In Section 4.1, we study the preference-ranking problem, where two

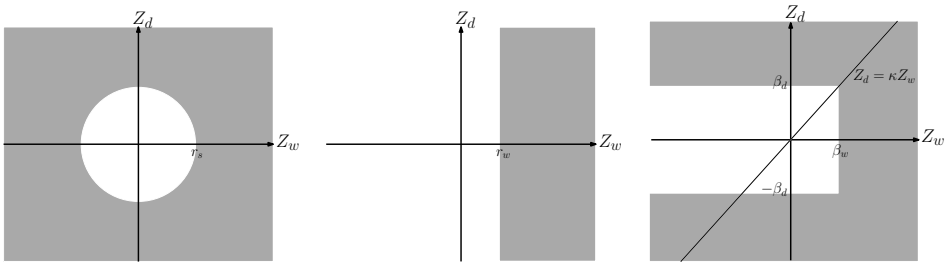


Figure 3. Rejection regions (in gray) of $S_G$, $R_{w,G}$, $M(\kappa)$. Left: $\{S_G \geq r_s^2\}$; middle: $\{Z_w \geq r_w\}$; right: $\{M(\kappa) \geq \beta(\kappa)\}$ ($\beta_d = \kappa\beta_w = \beta(\kappa)$).

groups of people are asked to rank six objects, and we test whether the two samples have the same preference. In Section 4.2, we compare the proposed tests on data generated directly from a multinomial distribution. Three existing tests are included in the comparison: Pearson's chi-squared test (denoted as "Pearson"), the deviance test (denoted as "LR"), and the kernel two-sample test of Gretton et al. (2012) (denoted as "Ker").

### 4.1. Preference-ranking problem

We consider the following two data-generating mechanisms:

(i) Data are generated 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 $\Xi$ is 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 samples 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 similar, so we present the results based on Spearman's distance in the following.

We compare the statistics $R_{0,(a)}$, $R_{0,(u)}$, $S_{(a)}$, $S_{(u)}$, $R_{w,(a)}$, $R_{w,(u)}$, $M_{(a)}(\kappa)$, and $M_{(u)}(\kappa)$ ($\kappa = 1.31, 1.14, 1$) using Pearson, LR, and Ker (Gretton et al. (2012)) in six scenarios (Scenarios 1–3 under (i), and Scenarios 4–6 under (ii)), with

balanced and unbalanced sample sizes. The settings with different $\theta$ and different $\eta$ under (i) are also considered, and the results are provided in the Supplementary Material S3.1. The parameters under each scenario are chosen such that the tests have moderate power, in order to be comparable.

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

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

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

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

- Scenario 5 (Different support): $\mathcal{D}_1 = \{\zeta \in \Xi : \zeta$ ranks No.1 before No.5$\}$ and $\mathcal{D}_2 = \{\zeta \in \Xi : \zeta$ ranks No.1 before No.6$\}$, with balanced ($n_1 = n_2 = 180$) and unbalanced ($n_1 = 180, n_2 = 220$) sample sizes.

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

The results are presented in Table 5, where the power is estimated using the fraction of trials (out of 1,000) that the test rejects the null hypothesis at the 0.05 significance level. Those above 95 percent of the best power under each setting are shown in bold.

We first check the results for the data generated by mechanism (i). We see that Pearson, LR, and Ker have low power under all three scenarios. For the extended statistics, $S_{(u)}$ and $M_{(u)}$ work well for all scenarios, whereas the others show obvious strengthes and weaknesses for different settings. For example, under 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 neither $R_{w,(a)}$ nor $R_{w,(u)}$ perform well when only $\theta$ differs (Scenarios 2 and 3). Overall, $M_{(u)}(\kappa)$ performs

Table 5. Estimated power of the tests under the six scenarios denoted by A1–A6, with (a) denoting the balanced setting and (b) denoting the unbalanced setting.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A1(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.866** | 0.759 | **0.866** | 0.837 | 0.815 | 0.780 | 0.194 | 0.197 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | **0.890** | 0.799 | **0.890** | **0.862** | **0.847** | 0.816 | 0.198 | |
| A1(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.654 | 0.880 | **0.955** | **0.942** | 0.930 | 0.910 | 0.469 | 0.469 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.885 | **0.965** | **0.984** | **0.977** | **0.970** | **0.962** | 0.312 | |
| A2(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.291 | 0.200 | 0.291 | 0.265 | 0.243 | 0.211 | 0.109 | 0.107 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.442 | **0.775** | 0.442 | 0.749 | **0.784** | **0.809** | 0.098 | |
| A2(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.526 | 0.332 | 0.352 | 0.361 | 0.349 | 0.335 | 0.017 | 0.014 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0 | **0.900** | 0.568 | 0.885 | **0.921** | **0.933** | 0.158 | |
| A3(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.297 | 0.217 | 0.297 | 0.278 | 0.269 | 0.240 | 0.107 | 0.116 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.464 | 0.780 | 0.464 | 0.754 | **0.791** | **0.820** | 0.092 | |
| A3(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.062 | 0.401 | 0.387 | 0.420 | 0.421 | 0.409 | 0.397 | 0.430 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | **0.962** | 0.884 | 0.582 | 0.867 | 0.903 | **0.920** | 0.113 | |
| A4(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.776** | 0.626 | **0.776** | **0.741** | 0.705 | 0.657 | 0.205 | 0.206 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.700 | 0.530 | 0.700 | 0.647 | 0.623 | 0.584 | 0.187 | |
| A4(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.865 | 0.791 | **0.914** | **0.876** | 0.850 | 0.825 | 0.300 | 0.306 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.812 | 0.688 | 0.818 | 0.779 | 0.761 | 0.732 | 0.216 | |
| A5(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.820** | 0.669 | **0.820** | 0.770 | 0.727 | 0.690 | **0.823** | **0.825** |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.656 | 0.525 | 0.656 | 0.620 | 0.573 | 0.537 | 0.742 | |
| A5(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.909** | 0.768 | **0.892** | 0.861 | 0.842 | 0.800 | **0.895** | **0.899** |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.769 | 0.640 | 0.730 | 0.708 | 0.683 | 0.659 | 0.794 | |
| A6(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.892** | 0.755 | **0.892** | **0.857** | 0.827 | 0.790 | 0.256 | 0.260 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.823 | 0.691 | 0.823 | 0.782 | 0.752 | 0.712 | 0.233 | |
| A6(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.940** | 0.902 | **0.970** | **0.958** | **0.943** | **0.925** | 0.352 | 0.350 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.891 | 0.822 | **0.930** | 0.903 | 0.881 | 0.859 | 0.291 | |

best among the tests. When $\theta$ 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 $\eta$ differs (Scenario 1). In general, the tests based on the "union" are slightly better than their "averaging" counterparts (except for some cases for $R_0$).

For the data generated by mechanism (ii), we see that all tests perform relatively well under Scenario 5. For the other two scenarios, 4 and 6, Pearson, LR, and Ker have low power. The proposed tests perform similarly well under both scenarios, with those based on "averaging" being slightly better than their "union" counterparts.

### 4.2. Multinomial distribution

Here, we generate data from a $d$-dimensional multinomial distribution. We consider $d = 100, d = 1{,}000$, and $d = 10{,}000$. Sample 1 consists of $n_1$ observations drawn randomly from $F_1 = \text{Mult}(M_1, p_1)$, for $i = 1, \ldots, n_1$, and sample 2 consists of $n_2$ observations drawn from $F_2 = \text{Mult}(M_2, p_2)$, for $i = 1, \ldots, n_2$. Here, $M_1$ and $M_2$ are the total counts of each observation in sample 1 and sample 2, respectively, and $p_1$ and $p_2$ are the respective compositions. We set $M_1 = M_2 = 3$, and consider the following choices of $p_i$. Let $p_1 = (a_1, a_2 \ldots, a_d)^T$ and $p_2 = (b_1, b_2, \ldots, b_d)^T$.

1) $d = 100$:

$$\text{Scenario 1 (B1): } a_i = 0.01, i = 1, \ldots, d; \;\; b_i = \begin{cases} 0.1 & i = 1 \\ \dfrac{0.9}{99} & i \geq 2 \end{cases}.$$

$$\text{Scenario 2 (B2): } a_i = \begin{cases} 0.002 & i \leq 70 \\ \dfrac{0.86}{30} & i \geq 71 \end{cases}; \;\; b_i = \begin{cases} 0.018 & i \leq 30 \\ \dfrac{0.46}{70} & i \geq 31 \end{cases}.$$

2) $d = 1{,}000$:

$$\text{Scenario 1 (C1): } a_i = 0.001, i = 1, \ldots, d; \;\; b_i = \begin{cases} 0.085 & i = 1 \\ \dfrac{0.915}{999} & i \geq 2 \end{cases}.$$

$$\text{Scenario 2 (C2): } a_i = \begin{cases} \dfrac{0.5}{970} & i \leq 970 \\ \dfrac{0.5}{30} & i \geq 971 \end{cases}; \;\; b_i = \begin{cases} \dfrac{0.6}{30} & i \leq 30 \\ \dfrac{0.4}{970} & i \geq 31 \end{cases}.$$

3) $d = 10{,}000$:

$$\text{Scenario 1 (D1): } a_i = 0.0001, i = 1, \ldots, d; \;\; b_i = \begin{cases} 0.18 & i = 1 \\ \dfrac{0.82}{9999} & i \geq 2 \end{cases}.$$

$$\text{Scenario 2 (D2): } a_i = \begin{cases} \dfrac{0.4}{9970} & i \leq 9970 \\ \dfrac{0.6}{30} & i \geq 9971 \end{cases}; \;\; b_i = \begin{cases} \dfrac{0.4}{30} & i \leq 30 \\ \dfrac{0.6}{9970} & i \geq 31 \end{cases}.$$

For each scenario, we examine both a balanced setting $n_1 = n_2 = 120$ and an unbalanced setting $n_1 = 120, n_2 = 200$. Under each setting, we randomly generate 1,000 data sets and estimate the power under the 0.05 significance level, are shown in Table 6. Values above 95 percent of the best power under each setting are shown in bold.

We see that Pearson and LR have no power under these scenarios. In Scenario 1 (B1, C1, D1), the graph-based statistics all perform reasonably well, except for $R_{0,(a)}$ and $R_{0,(u)}$ under the unbalanced setting. In Scenario 2 (B2, C2, D2), the extended generalized edge-count tests and extended max-type edge-count tests outperform all other tests, indicating that the alternative in this type of scenario is more in the scale domain than in the location domain.

## 5. Asymptotics

In this section, we provide the asymptotic distributions of the new test statistics described in Section 3. This provides us with a theoretical basis for obtaining approximate $p$-values in an analytic way. We examine how well these approximations work for finite samples by checking the empirical sizes of the new test statistics at the end of this section, and by comparing the $p$-values obtained from the asymptotic results and those using random permutations in the Supplementary Material S3.3. 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.** $|C_0|, \sum_{(u,v)\in C_0}(1/m_u m_v) = O(N)$; $K, \sum_u (1/m_u) = O(N^\alpha), \alpha \leq 1$.

Table 6. Estimated power of the tests under scenarios B1, B2, C1, C2, D1, and D2, with (a) denoting the balanced setting and (b) denoting the unbalanced setting.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| B1(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.637** | 0.560 | **0.637** | 0.600 | 0.600 | 0.570 | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | **0.633** | 0.507 | **0.633** | 0.590 | 0.557 | 0.547 | 0.313 | |
| B1(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.023 | **0.754** | **0.780** | **0.777** | **0.770** | 0.746 | 0.002 | 0.002 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.030 | 0.734 | **0.788** | **0.773** | **0.761** | 0.743 | 0.063 | |
| B2(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.050 | **0.822** | 0.050 | 0.550 | 0.620 | 0.674 | 0.004 | 0.004 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.044 | 0.774 | 0.044 | 0.366 | 0.436 | 0.486 | 0.364 | |
| B2(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.660 | **0.878** | 0.168 | 0.726 | 0.754 | 0.768 | 0.012 | 0.012 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.404 | **0.866** | 0.164 | 0.650 | 0.722 | 0.762 | 0.646 | |
| C1(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.773** | **0.766** | **0.773** | **0.768** | **0.766** | **0.758** | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | **0.773** | **0.766** | **0.773** | **0.768** | **0.766** | **0.758** | 0.675 | |
| C1(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.002 | **0.942** | **0.948** | **0.944** | **0.944** | **0.942** | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.002 | **0.942** | **0.948** | **0.944** | **0.944** | **0.942** | 0.550 | |
| C2(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.604 | **0.823** | 0.604 | 0.705 | 0.726 | 0.734 | 0.001 | 0.001 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.603 | **0.826** | 0.603 | 0.705 | 0.722 | 0.730 | 0.660 | |
| C2(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.006 | **0.921** | 0.245 | 0.763 | 0.807 | 0.824 | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.006 | **0.921** | 0.242 | 0.758 | 0.801 | 0.821 | 0.656 | |
| D1(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | **0.699** | **0.715** | **0.699** | **0.716** | **0.712** | **0.713** | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | **0.700** | **0.715** | **0.700** | **0.716** | **0.712** | **0.713** | 0.664 | |
| D1(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.227 | **0.936** | **0.923** | **0.930** | **0.930** | **0.933** | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.304 | **0.936** | **0.923** | **0.930** | **0.930** | **0.933** | 0.528 | |
| D2(a) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.075 | **0.877** | 0.075 | 0.608 | 0.649 | 0.677 | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.076 | **0.876** | 0.076 | 0.597 | 0.646 | 0.673 | 0.607 | |
| D2(b) | $R_{0,(a)}$ | $S_{(a)}$ | $R_{w,(a)}$ | $M_{(a)}(1.31)$ | $M_{(a)}(1.14)$ | $M_{(a)}(1)$ | LR | Pearson |
| | 0.588 | **0.897** | 0.301 | 0.767 | 0.788 | 0.810 | 0 | 0 |
| | $R_{0,(u)}$ | $S_{(u)}$ | $R_{w,(u)}$ | $M_{(u)}(1.31)$ | $M_{(u)}(1.14)$ | $M_{(u)}(1)$ | Ker | |
| | 0.571 | **0.895** | 0.300 | 0.765 | 0.785 | 0.806 | 0.756 | |

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

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

**Condition 3.** $\sum_u((|\mathcal{E}_u^{C_0}| - 2)^2/4m_u) - ((|C_0| - K)^2/N) = O(N)$.

**Remark 1.** A special case for Condition 1 is $|C_0|, \sum_{(u,v)\in C_0}(1/m_u m_v), K, \sum_u(1/m_u) = O(N)$. This and Condition 2 are the same as those stated in Chen and Zhang (2013) to obtain the asymptotic properties of $R_{0,(a)}$ and $R_{0,(u)}$. Condition 1 is easily satisfied, and Condition 2 sets constraints on the number of repeated observations and the degrees of the nodes in the graph $C_0$, such that they cannot be too large. When $m_u \equiv m$, for all $u$, Condition 2 simplifies 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) ensures that $(R_1, R_2)^T$ does not degenerate asymptotically. When $m_u \equiv m$, for all $u$, Condition 3 becomes $(1/4m)\sum_u |\mathcal{E}_u^{C_0}|^2 - (|C_0|^2/mK) = (1/4m)\sum_u(|\mathcal{E}_u^{C_0}| - (2|C_0|/K))^2 = O(N)$, which is the variance of the degrees of the nodes in $C_0$. When there is not enough variety in the degrees of the nodes in $C_0$, the correlation between $R_1$ and $R_2$ tends to one. (A similar condition is needed for the continuous counterpart (Chen and Friedman (2017)).)

**Theorem 3.** *Under Conditions* 1, 2, *and* 3, *as* $N \to \infty$, $\left(Z_{w,(a)}, Z_{d,(a)}\right)^T \xrightarrow{D} \mathcal{N}_2(0, \mathbf{I}_2)$ *under the permutation null distribution.*

The proof of this theorem is given in the Supplementary Material S1.2. 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* $\Phi(x/\kappa)(2\Phi(x) - 1)$ *under the permutation null distribution, where* $\Phi(x)$ *denotes the cumulative distribution function of the standard normal distribution.*

### 5.2. Statistics based on the union

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

**Condition 4.** $|\bar{G}| = O(N)$.

**Condition 5.** $\sum_{i=1}^{N} |\mathcal{E}_i^{\bar{G}}|^2 - (4/N)|\bar{G}|^2 = O(N)$.

**Condition 6.**

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

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

**Remark 2.** Condition 4 is easily satisfied. Condition 5 is mentioned in Chen and Friedman (2017) for the continuous version. When $m_u \equiv m$, for all $u$, Condition 5 can be rewritten as $\sum_{u=1}^{K} |\mathcal{E}_u^{C_0}|^2 - (4/K)|C_0|^2 = O(K)$. If $C_0$ is the $k$-MST, $k = O(1)$, constructed under the Euclidean distance, the above condition always holds, based on the results of 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 1. These conditions restrict the degrees of the nodes in graph $C_0$.

**Theorem 4.** *Under Conditions* 4, 5, *and* 6, *as* $N \to \infty$, $\left(Z_{w,(u)}, Z_{d,(u)}\right)^T \xrightarrow{D} \mathcal{N}_2(0, \mathbf{I}_2)$ *under the permutation null distribution.*

The proof of this theorem is given in the Supplementary Material S1.3. 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* $\Phi(x/\kappa)(2\Phi(x) - 1)$ *under the permutation null distribution, where* $\Phi(x)$ *denotes the cumulative distribution function of the standard normal distribution.*

To determine whether these theoretical results are useful in practice, we check the empirical sizes of these tests, with the $p$-value determined using the

Table 7. Empirical size at 0.05 nominal level.

| Statistic | $n_1 = 50$ $n_2 = 50$ | $n_1 = 50$ $n_2 = 100$ | $n_1 = 50$ $n_2 = 150$ | $n_1 = 100$ $n_2 = 100$ | $n_1 = 100$ $n_2 = 200$ | $n_1 = 100$ $n_2 = 300$ |
|---|---|---|---|---|---|---|
| $S_{(a)}$ | 0.032 | 0.043 | 0.043 | 0.038 | 0.030 | 0.033 |
| $S_{(u)}$ | 0.036 | 0.027 | 0.034 | 0.033 | 0.037 | 0.036 |
| $R_{w,(a)}$ | 0.038 | 0.039 | 0.039 | 0.041 | 0.037 | 0.037 |
| $R_{w,(u)}$ | 0.046 | 0.043 | 0.033 | 0.038 | 0.035 | 0.033 |
| $M_{(a)}(1.31)$ | 0.039 | 0.044 | 0.042 | 0.039 | 0.034 | 0.030 |
| $M_{(u)}(1.31)$ | 0.041 | 0.035 | 0.036 | 0.036 | 0.042 | 0.038 |
| $M_{(a)}(1.14)$ | 0.039 | 0.047 | 0.043 | 0.036 | 0.033 | 0.028 |
| $M_{(u)}(1.14)$ | 0.039 | 0.031 | 0.033 | 0.035 | 0.040 | 0.038 |
| $M_{(a)}(1)$ | 0.042 | 0.044 | 0.040 | 0.036 | 0.032 | 0.025 |
| $M_{(u)}(1)$ | 0.039 | 0.029 | 0.029 | 0.035 | 0.042 | 0.044 |

asymptotic results directly. We generate the data using mechanism (i) in Section 4.1, with $\theta_1 = \theta_2 = 5$ and $\eta_1 = \eta_2 = \{1, 2, 3, 4, 5, 6\}$. Table 7 shows the empirical sizes of the tests under difference choices of sample sizes. The empirical size is computed as the fraction of trials (out of 1,000) in which the asymptotic $p$-value ($p$-value computed based on the asymptotic distribution directly) is less than 0.05. We see that the empirical sizes are well controlled for all proposed tests, even when the sample sizes are in the 50s. We provide additional examinations of the asymptotic $p$-values by comparing them with the permutation $p$-values in the Supplementary Material S3.3.

## 6. Phone-Call Network Data Analysis

Here, we analyze phone-call network data. The MIT Media Laboratory conducted a study of 106 subjects, including students and staff of an institute, who use mobile phones with pre-installed software that can record call logs. The study lasted from July 2004 to June 2005 (Eagle, Pentland and Lazer (2009)). Given the richness of this data set, many problems can be studied. One question of interest is whether phone-call patterns on weekdays differ from those on weekends. The phone calls on weekdays and weekends can be viewed as representations of professional and personal relationships, 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 at least one call to person $j$ on that day. We encode the directed network of each day using an adjacency matrix, with element $[i, j]$ taking the value one if there is a directed edge pointing from subject $i$ to

Table 8. Breakdown statistics of the phone-call network data.

| | Value | Mean | Value-Mean | SD | | | |
|---|---|---|---|---|---|---|---|
| $R_{1,(a)}$ | 2,800.26 | 2,669.56 | 130.70 | 143.33 | | | |
| $R_{2,(a)}$ | 409.18 | 420.80 | -11.62 | 57.75 | | | |
| $(R_{1,(a)} + R_{2,(a)})/2$ | 1,604.72 | 1,545.18 | 59.54 | 44.74 | | | |
| $R_{w,(a)}$ | 1,087.14 | 1,058.40 | 28.73 | 11.79 | | | |
| $R_{d,(a)}$ | 2,391.08 | 2,248.76 | 142.32 | 199.37 | | | |
| | Value | Mean | Value-Mean | SD | | | |
| $R_{1,(u)}$ | 7,163.00 | 6,860.35 | 302.65 | 381.50 | | | |
| $R_{2,(u)}$ | 1,008.00 | 1,081.38 | -73.38 | 151.66 | | | |
| $(R_{1,(u)} + R_{2,(u)})/2$ | 4,085.50 | 3,970.86 | 114.64 | 116.22 | | | |
| $R_{w,(u)}$ | 2,753.17 | 2,719.93 | 33.24 | 15.65 | | | |
| $R_{d,(u)}$ | 6,155.00 | 5,778.97 | 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 |
| | $\kappa = 1.31$ | 3.19 | 0.009 | | $\kappa = 1.31$ | 2.78 | 0.022 |
| $M_{(a)}(\kappa)$ | $\kappa = 1.14$ | 2.78 | 0.013 | $M_{(u)}(\kappa)$ | $\kappa = 1.14$ | 2.42 | 0.032 |
| | $\kappa = 1$ | 2.44 | 0.022 | | $\kappa = 1$ | 2.12 | 0.050 |

subject $j$, and zero otherwise.

In the data set, there are 236 weekdays and 94 weekends. Among the 330 $(236 + 94)$ networks, there are 285 distinct values, 11 of which have more than one observation. We denote the distinct values as matrices $B_1, \ldots, B_{285}$. We adopt the distance measure used in Chen and Friedman (2017) and Chen, Chen and Su (2018), which is defined as the number of different entries, this is, $d(B_i, B_j) = \|B_i - B_j\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm of a matrix. In addition to the repeated observations, there are many equal distances among the distinct values. We set $C_0$ to be the 3-NNL, which has similar density to that of the 9-MST recommended in Chen, Chen and Su (2018).

Table 8 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 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. 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 the results based on "averaging." We can see that $R_{1,(a)}$ is much higher than its expectation, whereas $R_{2,(a)}$ is smaller than its expectation. 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, owing to the variance boosting issue, it fails to reject the null hypothesis at the 0.05 significance level. On the other hand, the weighted edge-count test chooses the proper weight to minimize the variance and performs well. Because $S_{(a)}$ and $M_{(a)}(\kappa)$ consider the weighted edge-count statistic and the difference between two within-sample edge-counts simultaneously, these tests all reject the null at the 0.05 significance level. Here, a larger value of $\kappa$ indicates greater similarity between the max-type test ($M_{(a)}(\kappa)$) and the weighted test ($R_{w,(a)}$). Thus, the $p$-values of $M_{(a)}(\kappa)$ are very close to those of $R_{w,(a)}$ when $\kappa$ is large. The results for the "union" counterparts are similar, except that $S_{(u)}$ cannot reject the null at the 0.05 significance level. Based on this table, there is clearly a mean difference between the two samples, but no significant scale difference.

We also compare the asymptotic $p$-values with the permutation $p$-values, and the results show they are quite close (see the Supplementary Material S3.3).

## 7. Conclusion

The generalized edge-count test and the weighted edge-count test are useful tools in two-sample testing frameworks. 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 these statistics, as well as the max-type statistic, to accommodate scenarios in which 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 a 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 for sample sizes in the hundreds or more, 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 such as that based on the phone-call network data

in Section 6 is recommended.

## Supplementary Material

The online Supplementary Material contains proofs of the lemmas and theorems, as well as some additional results.

## Acknowledgments

Jingru Zhang's research was supported, in part, by the CSC scholarship. Hao Chen's research was supported, in part, by NSF award DMS-1513653.

## References

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

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

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

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

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

Chen, H. and Zhang, N. R. (2013). Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica* **23**, 1479–1503.

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

Eagle, N., Pentland, A. S. and Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* **106**, 15274–15278.

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research* **13**, 723–773.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics* **16**, 772–783.

Li, J. and Chen, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics* **40**, 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**, 515–530.

Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association* **81**, 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**, 6535–6542.

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

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

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

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

Jingru Zhang

Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, USA

E-mail: jingru.zhang@pennmedicine.upenn.edu

Hao Chen

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

E-mail: hxchen@ucdavis.edu

(Received March 2019; accepted July 2020)