Statistica Sinica Preprint No: SS-2019-0116	
Title	Graph-based two-sample tests for data with repeated observations
Manuscript ID	SS-2019-0116
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0116
Complete List of Authors	Jingru Zhang and Hao Chen
Corresponding Author	Hao Chen
E-mail	hxchen@ucdavis.edu
Notice: Accepted version subject to English editing.	

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 (for example Srivastava and Du (2008)) and whether the covariance matrices are the same (see for examples Schott (2007); Srivastava and Yanagihara (2010); Xia et al. (2015)). To make applicability more extensive, many semi-parametric methods were proposed to test the means or covariance matrices (see for examples Bai and Saranadasa (1996); Chen et al. (2010); Cai et al. (2014); Xu et al. (2016); Li and Chen (2012); Cai et al. (2013)) by adding some conditions on moment and/or covariance instead of the assumption of underlying distributions. These parametric or semi-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 re-

view). 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 pooled the observations from the two samples together and constructed 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 counted the number of edges in the MST that connect 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 **??**).

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 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, 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 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 Notations

Among the N observations, we assume there are K distinct values and index them by $1, 2, \cdots, K$. Basic notations are 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 $d(i, j)$ be the distance between 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 $n_1!n_2!/N!$ 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 the nearest neighbor link (NNL) for easy reference. If one wants denser

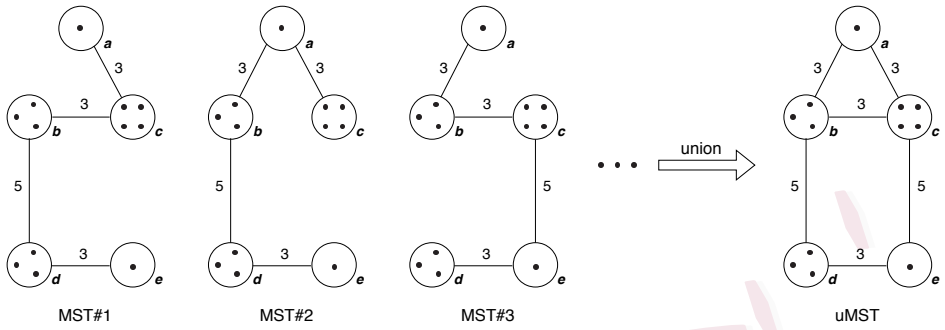

Figure 1: There are five distinct values (**a**, **b**, **c**, **d**, **e**) denoted by the circles. For some distinct values, there are more than one observations, denoted by the more than one point in the circle. The distance between the distinct values are denoted on the edges. It is clear that there 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.

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 observations 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.

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

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 $15,552 (= 1^2 \cdot 3^3 \cdot 4^3 \cdot 3^2 \cdot 1^2)$ 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 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

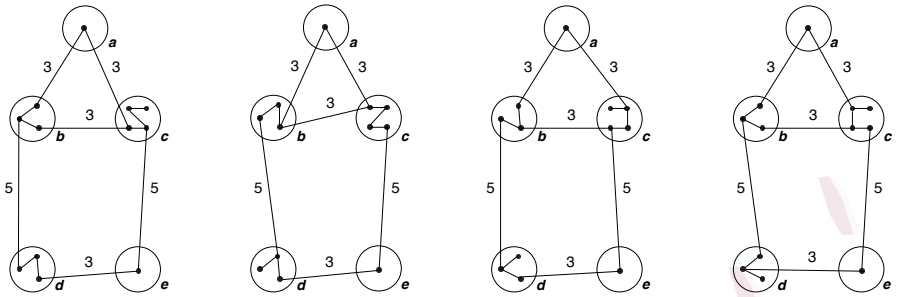

Figure 2: Four graphs, out of 2,239,488, on observations initiated from the NNL on distinct values.

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

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

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 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 $\mathcal{N}_d(\mu, \mathbf{I}_d)$ ($\|\mu\|_2 = 1.3,\ d = 50$), 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 see 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 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:*

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

The notations $\{n_{ik}\}_{i=1,2;\ k=1,\ldots,K}$, $\{m_k\}_{k=1,\ldots,K}$ are declared in Table 2. 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, $S_G = \begin{pmatrix} R_{1,G}-\mathsf{E}(R_{1,G}) \\ R_{2,G}-\mathsf{E}(R_{2,G}) \end{pmatrix}^T \mathbf{\Sigma}_G^{-1} \begin{pmatrix} R_{1,G}-\mathsf{E}(R_{1,G}) \\ R_{2,G}-\mathsf{E}(R_{2,G}) \end{pmatrix}$, one straightforward way of defining the average statistic would be $\frac{1}{|\mathcal{G}_{C_0}|}\sum_{G\in\mathcal{G}_{C_0}} S_G$. However, $\mathbf{\Sigma}_G$ 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 $R_{w,G} = \frac{n_2-1}{N-2}R_{1,G}+\frac{n_1-1}{N-2}R_{2,G}$, $R_{d,G} = R_{1,G}-R_{2,G}$. 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 tests based on how they were introduced in Chen and Friedman (2017) and Chen et al. (2018), 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 **??**). 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 **??** 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. 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 **??**.

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

$$
\begin{aligned}
\mathsf{E}(R_{w,(u)}) &= |\bar{G}|\tfrac{(n_1-1)(n_2-1)}{(N-1)(N-2)},\\
\mathsf{Var}(R_{w,(u)}) &= \tfrac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{|\bar{G}| - \tfrac{1}{N-2}\sum_{i=1}^{N}|\mathcal{E}_i^{\bar{G}}|^2 + \tfrac{2}{(N-1)(N-2)}|\bar{G}|^2\right\},\\
\mathsf{E}(R_{w,(a)}) &= (N-K+|C_0|)\tfrac{(n_1-1)(n_2-1)}{(N-1)(N-2)},\\
\mathsf{Var}(R_{w,(a)}) &= \tfrac{n_1(n_1-1)n_2(n_2-1)}{N(N-1)(N-2)(N-3)}\left\{-\tfrac{4}{N-2}\left(\sum_u \tfrac{(|\mathcal{E}_u^{C_0}|-2)^2}{4m_u} - \tfrac{(|C_0|-K)^2}{N}\right)\right.\\
&\qquad \left. + 2(K - \sum_u \tfrac{1}{m_u}) + \sum_{(u,v)\in C_0}\tfrac{1}{m_u m_v} - \tfrac{2}{N(N-1)}(|C_0|+N-K)^2\right\},
\end{aligned}
$$

where $|\mathcal{E}_i^{\bar{G}}| = m_u - 1 + \sum_{\mathcal{V}_u^{C_0}} m_v$ 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.

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 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 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:*

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

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

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

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

Theorem 2 is proved in Supplement **??**.

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 **??**. 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 reject regions for the extended statistics in Table 4, which are similar to their continuous counterparts. Since the testing rule is same for the averaging and the union statistics, we omit 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 through the asymptotic distributions of the extended statistics (see Section 5).

Table 4: Reject 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	$M(\kappa) \geq \beta(\kappa)$

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

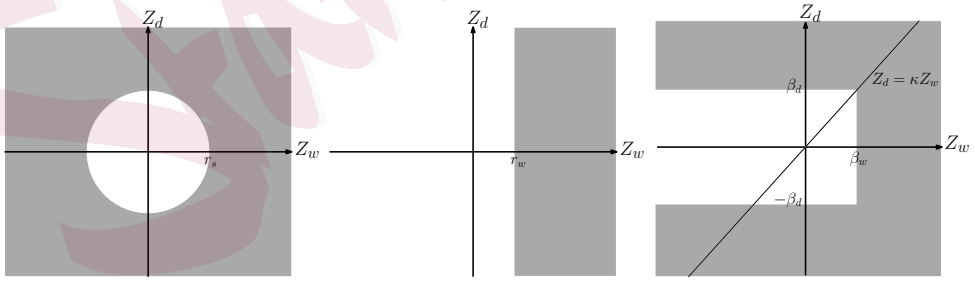

Figure 3: Rejection regions (in gray) of S_G, $R_{w,G}$, $M(\kappa)$. Left: $\{S_G \geq r_s^2\}$; middle: $\{Z_w \geq r_w\}$; right: $\{M(\kappa) \geq \beta(\kappa)\}$ $(\beta_d = \kappa\beta_w = \beta(\kappa))$.

4. Performance of the extended test statistics

In this section, we study the performance of various tests through simulation studies. In Section 4.1, we study the preference ranking problem, where two 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 directly generated from the multinomial distribution. Three existing tests are included in the comparison: the Pearson's Chi-square test (denoted as "Pearson"), the deviance test (denoted as "LR"), and the kernel two-sample test in Gretton et al. (2012) (denoted as "Ker").

4.1 Preference ranking problem

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 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 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 $M_{(u)}(\kappa)$ ($\kappa = 1.31, 1.14, 1$) with Pearson's Chi-square test, the deviance test and the kernel test (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 both different θ and different η under (i) are also considered and the results can be found in Supplement **??**. 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 ($n_1 = n_2 = 180$) and unbalance ($n_1 = 180, n_2 = 220$) 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 5 where the power is estimated by the fraction of trials (out of 1000) that the test rejects the null hypothesis at 0.05 significance level. Those above 95 percentage of the best power under each setting are in bold.

We first check results for the data generated by mechanism (i). We see that Pearson's Chi-square test, the deviance test, and the kernel two-sample test have low power under all three scenarios. For the extended statistics, $S_{(u)}$ and $M_{(u)}$ work well for all scenarios, while the others show obvious

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

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 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).

For the data generated by mechanism (ii), we see that all tests are doing pretty well under Scenario 5. For the other two scenarios, 4 and 6, Pearson's Chi-square test, the deviance test, and the kernel two-sample test have low power. The proposed tests perform similarly well under both scenarios with those based on "averaging" slightly better than their "union" counterparts.

4.2 Multinomial distribution

Here, we generate data from d-dimensional multinomial distribution. We consider $d = 100, d = 1,000$ and $d = 10,000$. Sample 1 consists of n_1 observations randomly drawn from $F_1 = \text{Mult}(M_1, p_1),\ i = 1, \cdots, n_1$ and sample 2 consists of n_2 observations from $F_2 = \text{Mult}(M_2, p_2),\ i = 1, \cdots, n_2$.

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 compositions. We set $M_1 = M_2 = 3$, and consider the following choices of p_i's. Let $p_1 = (a_1, a_2 \ldots, a_d)^T, p_2 = (b_1, b_2, \ldots, b_d)^T$. Different choices of p_i's are considered.

1) $d = 100$

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

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

2) $d = 1,000$

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

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

3) $d = 10,000$

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

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

For each scenario, we examine both balanced setting $n_1 = n_2 = 120$ and unbalanced setting $n_1 = 120, n_2 = 200$. Under each setting, we randomly generated 1,000 data sets and estimated power under 0.05 significance level

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

are presented in Table 6. Those above 95 percentage of the best power under each setting are in bold.

We see that the Pearson's Chi-square test and the deviance test have no power under these scenarios. In Scenario 1's (B1, C1, D1), all the graph-based statistics perform reasonably well except for $R_{0,(a)}$ and $R_{0,(u)}$ under the unbalanced setting. In Scenario 2's (B2, C2, D2), the extended generalized edge-count tests and extended max-type edge-count tests work much better than all other tests, indicating 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 new test statistics described in Section 3. This provides us theoretical bases for obtaining approximate p-values in an analytic way. We examine how well these approximations work for finite samples by checking the empirical size of the new test statistics at the end of this section and further by comparing the p-value obtained through asymptotic results and that through random permutations in Supplement **??**. 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. $|C_0|, \sum_{(u,v)\in C_0} \frac{1}{m_u m_v} = O(N);\ K, \sum_u \frac{1}{m_u} = O(N^\alpha), \alpha \le 1.$

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 1. One special case for Condition 1 is $|C_0|, \sum_{(u,v)\in C_0} \frac{1}{m_u m_v}, K, \sum_u \frac{1}{m_u} = O(N)$. This and Condition 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* $N \to \infty$, $\left(Z_{w,(a)}, Z_{d,(a)}\right)^T \xrightarrow{D} \mathcal{N}_2(0, \mathbf{I}_2)$ *under the permutation null distribution.*

The proof of this theorem is in Supplement **??**. 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 2. 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 1. These conditions restrict the degrees of 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 in Supplement **??**. 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.*

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

To see whether these theoretical results are useful in practice, we check the empirical size of these tests with the p-value determined by the asymptotic results directly. We generate data through mechanism (i) in Section 4 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 1000) that the asymptotic p-value (p-value computed based on the asymptotic distribution

directly) less than 0.05. We see that the empirical sizes are well controlled for all proposed tests even when the sample sizes are in 50s. We provide more examinations on the asymptotic p-values by comparing them with permutation p-values in Supplement **??**.

6. Phone-call network data analysis

We analyze the phone-call network data mentioned in Section 1 in details. 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.

In the dataset, there are 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 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 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 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, due

Table 8: 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

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.

We also compare the asymptotic p-values with the permutation p-values and the result shows they are quite close (details in Supplement **??**).

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.

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

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

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)