Statistica Sinica Preprint No: SS-2023-0408	
Title	A Robust Framework for Graph-Based Two-Sample Tests Using Weights
Manuscript ID	SS-2023-0408
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0408
Complete List of Authors	Yichuan Bai and Lynna Chu
Corresponding Authors	Yichuan Bai
E-mails	ycbai@iastate.edu
Notice: Accepted author version.	

A ROBUST FRAMEWORK FOR GRAPH-BASED TWO-SAMPLE TESTS USING WEIGHTS

Yichuan Bai and Lynna Chu

Iowa State University

Abstract: Graph-based tests are a class of non-parametric two-sample tests useful for analyzing high-dimensional data. The test statistics are constructed from similarity graphs (such as K-minimum spanning tree), and consequently, their performance is sensitive to the structure of the graph. When the graph has problematic structures (for example, hubs), as is common for high-dimensional data, this can result in low power and unstable performance among existing graph-based tests. We address this challenge by proposing new test statistics that are robust to problematic structures of the graph and can provide reliable inferences. We employ an edge-weighting strategy using intrinsic characteristics of the graph that are computationally simple and efficient to obtain. The limiting null distribution of the robust test statistics is derived and shown to work well for finite sample sizes. Simulation studies and data analysis of Chicago taxi-trip travel patterns demonstrate the new tests' improved performance across a range of settings.

Key words and phrases: curse of dimensionality, graph-based tests, high-dimensional data, non-parametric tests, robustness, similarity graphs.

1. Introduction

We focus on testing the equality of distributions for observations in the high dimensional setting, where the dimension of the observation d may be much larger than the sample size N. Suppose we have two samples $\{\mathbf{X_1}, \ldots, \mathbf{X_{n_1}}\}$ and $\{\mathbf{Y_1}, \ldots, \mathbf{Y_{n_2}}\}$ of d-dimensional observations that are independently and identically distributed from unknown distributions F_X and F_Y, respectively. The two-sample problem aims to test $H_0 : F_X = F_Y$ against an omnibus alternative $H_1 : F_X \neq F_Y$. This is a classic statistical problem but made more challenging by the increasing complexity of modern data, where observations can be high-dimensional data objects ($d >> N$). In this setting, it is often intractable to express or estimate F_X and F_Y directly due to the curse of dimensionality. Substantial developments have been made by the contemporary statistics community to address such challenges. For example, non-parametric two-sample tests for multivariate and high-dimensional data have been proposed using distances (Baringhaus and Franz (2004); Székely et al. (2004); Biswas and Ghosh (2014); Li (2018)), generalized ranking (Liu and Singh (1993); Hall and Tajvidi (2002)), and kernels (Gretton et al. (2012); Song and Chen (2023); Zhu and Shao (2021)).

While all of the mentioned methods can be applied to the high-dimensional setting, many do not explicitly address how to resolve various aspects of the curse of dimensionality. For example, distance-based test statistics are commonly used in the high-dimensional setting, but it has been observed that distances may not be meaningful in high-dimensional space since they have a tendency to concentrate when d is large. As such, distance-based

test statistics may have trouble effectively distinguishing similarities between observations, leading to reduced power. Moreover, the distribution of distances becomes considerably skewed as dimensionality increases, resulting in a phenomenon known as *hubness*. To be precise, let $N_k(x)$ be the number of times an observation x is among the k nearest neighbors of all other points in the data set. When the dimensionality is high, the distribution of N_k becomes right-skewed, resulting in the emergence of hubs. This hubness phenomenon affects methods that directly (or indirectly) make use of distances between observations; this includes pairwise distance-based tests such as the energy statistic (Székely et al. (2004)) and graph-based tests based on interpoint distances (described below). As a result, many existing two-sample tests are often vulnerable to the hubness aspect of the dimensionality curse, which can incur poor or unstable performance under various scenarios.

In this paper, we explore the hubness phenomenon and its effect on a class of tests based on geometric graphs constructed using interpoint distances. We refer to these as graph-based two-sample tests; the first test was proposed by Friedman and Rafsky (1979), and since then, numerous extensions and theoretical developments have been made. For example, Schilling (1986) and Henze (1988), Rosenbaum (2005), and Biswas et al. (2014) proposed test statistics specifically for k-NN graphs, minimum distance pairing, and Hamiltonian graphs, respectively. Chen and Friedman (2017), Chen et al. (2018), and Chu and Chen (2019) proposed new graph-based test statistics that target a wider range of alternatives. Banerjee et al. (2020) and Banerjee et al. (2024) proposed modifications of graph-based tests targeting the setting when heterogeneity is present in the

two samples due to latent subpopulations. Zhou and Chen (2023) proposed incorporating ranks in a similarity graph to boost the power of existing tests. Zhu and Chen (2024) studied asymptotic results for dense graphs. Details on constructing graph-based tests are provided in Section 2.1.

Despite their utility, these graph-based tests are sensitive to problematic data structures that can arise in the graphs. If the similarity graph is relatively flat, the existing tests work quite well. However in the presence of hubs and other problematic structures, the current tests suffer from reduced power and unreliable inference. We illustrate and explain why the hubness phenomenon can cause complications for the existing graph-based tests in Section 2.3. While some graph-based methods, such as the cross-match test based on non-bipartite matching (Rosenbaum, 2005) and the Shortest Hamiltonian Path (SHP)-based test (Biswas et al., 2014), can mitigate the hubness problem by placing constraints on the graph construction, these tests tend to suffer from low power under some common scenarios when the observations are high-dimensional (see Supplementary Section S1.1 for additional details). Recently, Zhu and Chen (2023) proposed a graph generation method that also places constraints on the graph; their approach involves optimizing an objective function with a penalty for a large node degree. However, if the similarity graph is constructed from domain knowledge or directly observed, as is often the case in real applications, their approach is no longer directly applicable. Moreover, their graph generating process could be computationally expensive and may also destroy vital internal connections between observations in the similarity graph. If the hubness is

too extreme, their approach deletes the hub from the graph. As demonstrated in Section 6, identifying the problematic hub may be nuanced, and straightforward deletion may not be ideal.

To address the hubness problem while preserving power and similarity information, we take a different approach and propose a robust framework for graph-based tests that employs an edge weighting strategy on the graph-based test statistics. We do not place constraints on graph construction nor generate a new graph, but instead use weights that are derived from the intrinsic characteristics of the similarity graph. These weights can mitigate the influence of hubs while effectively retaining power in the presence of hubs. We demonstrate through theoretical analysis, simulation studies, and real data applications the improved performance of this robust framework.

The paper is organized as follows. In Section 2, we review the graph-based testing framework and discuss the hubness phenomenon. We then propose a robust solution in Section 3, which involves choosing weights to dampen the effect of hubs. In Section 4, the asymptotic null distributions of the proposed test statistics are derived. Section 5 examines the power of the robust test statistics under different simulation settings. In Section 6, the robust test statistics are illustrated in the analysis of Chicago taxi data and some concluding remarks are given in Section 7.

2. Graph-based Testing Framework

2.1 Background

Graph-based tests provide a general framework to conduct two-sample tests for multivariate and non-Euclidean data. A similarity graph is constructed from the pooled observations of both samples according to a similarity measure (such as Euclidean distance). The similarity graph can be constructed based on a certain criterion. For example, a minimum spanning tree (MST) is a similarity graph that connects all observations in such a way that the total distance across edges is minimized. A k-MST is the union of MST and $k-1$ spanning trees, where the ith ($i > 1$) spanning tree does not contain any edges from the first $i-1$ spanning trees. Other examples include the k-nearest neighbor graph (k-NNG), where each observation is connected to its k nearest neighbors. Alternatively, the graph could be constructed according to domain knowledge and expertise.

Three quantities of the graph are computed: the number of edges connecting between the two samples (R_0), the number of edges connecting within sample $\mathbf{X}$ (R_1), and the number of edges connecting within sample $\mathbf{Y}$ (R_2). A combination of these edge counts is used to construct different graph-based test statistics. Friedman and Rafsky (1979) proposed using (a standardized) R_0 as the test statistic such that a small R_0 is evidence against the null hypothesis that the two distributions are equal. Their rationale was that if the two samples really do come from different distributions, then the number of edges connecting between different samples should be relatively small. While a small R_0 as evidence against the null holds well when the two distributions differ in means, this

rationale can be invalid for more general alternatives - for example, when the change in distribution also involves scale change or the two samples are unbalanced. To resolve this, graph-based test statistics were proposed in Chen and Friedman (2017), Chen et al. (2018), and Chu and Chen (2019) that use a combination of R_1 and R_2 and can target a wider range of alternatives.

2.2 Hubness phenomenon in high-dimensional data

Hubs, defined to be nodes in the graph with a large degree, are a product of the curse of dimensionality. The hubness phenomenon was carefully studied in Radovanovic et al. (2010), which showed that hubs are an inherent property of data distributions in high-dimensional and not an artifact of finite samples or specific data distributions. Their theoretical analysis showed that the probability a hub emerges increases as the data dimension increases. The high-dimensional setting amplifies the tendency of central observations (observations close to the mean) to become hubs, effectively making it easier for such an observation to become a 'popular' or 'central' node. As a result, k-MSTs and k-NNGs constructed on high-dimensional data tend to have large hubs under standard distance measures, such as L_p. To see that the presence of hubs is a common phenomenon for high-dimensional data, we construct 5-MST graphs using Euclidean distance and report the maximum and 95th percentile of node degrees. As shown in Figure 1, we see that the maximum node degrees are more than three times as much as the 95th percentiles. Similar results using 5-NN constructed from Euclidean distance are shown in Supplementary Section S2.1. Clearly

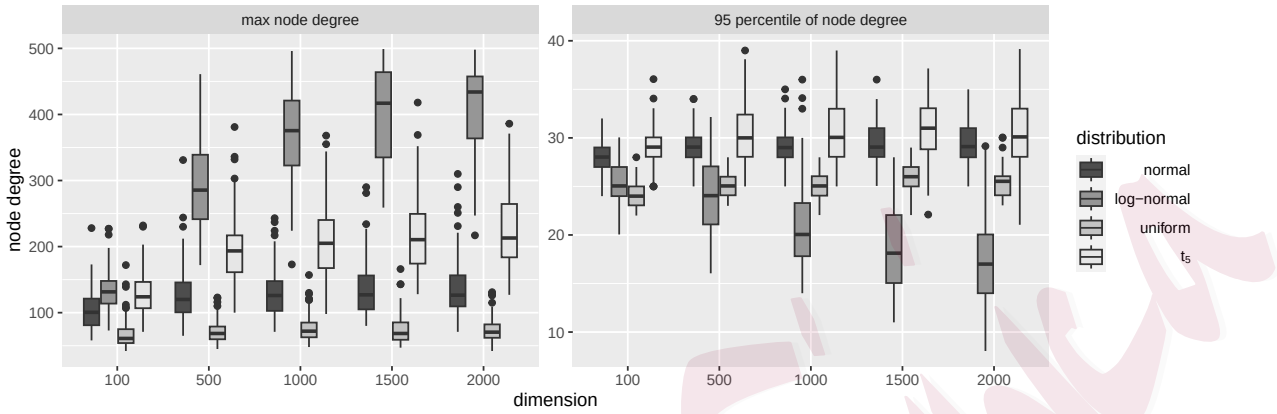


Figure 1: Boxplot of maximum and 95th percentiles of node degrees for different dimensions. Results are from 100 simulations with $n = 500$, where observations are drawn from d-dimensional normal, log-normal, uniform, and t distributions.

it is not uncommon to have a node with a degree much larger than the majority of other nodes' in the similarity graph. These hubs can be highly influential nodes and can distort final inference results depending on whether these observations are included or excluded in graph construction.

2.3 Limitations of current graph-based tests

Consider the following example that illustrates why hubs may cause problems in existing graph-based tests. In Scenario 1, we generate two samples ($n_1 = n_2 = 200$) with moderate dimension that differ in mean and variance, $\mathcal{F}_1 : \mathcal{N}(\mathbf{0}_s, \mathbf{I}_s)$ and $\mathcal{F}_2 : \mathcal{N}(\sqrt{\boldsymbol{(0.2log(s)/s)}}\mathbf{1_s}, (1 + 3log(s)/s)^2\mathbf{I}_s)$; in Scenario 2, the samples ($n_1 = n_2 = 200$) are generated with the same change in mean and variance but the observations are high-dimensional: $\mathcal{F}_1 : \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $\mathcal{F}_2 : \mathcal{N}((\sqrt{\boldsymbol{(0.2log(s)/s)}}\mathbf{1_s}, \mathbf{0_{d-s}}), \begin{pmatrix} (1+3log(s)/s)^2\mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-s} \end{pmatrix})$,

where $d = 1000$ and $s = \lfloor \sqrt{d} \rfloor$. A 5-MST is constructed from Euclidean distance on the pooled observations ($n_1 + n_2 = 400$). We observe that the maximum node degree of the graph is 64 in Scenario 1 and 80 in Scenario 2. The generalized edge-count test S (Chen and Friedman (2017)) and the max-type edge-count test M (Chu and Chen (2019)), which consist of different combinations of R_1 and R_2, are applied to both scenarios since they can detect general distributional differences (i.e., both mean and variance change). A large value of S or M serves as evidence against the null hypothesis. Both tests are capable of detecting the difference between two samples under Scenario 1. However, in Scenario 2, in the presence of a larger hub, both tests cannot reject the null hypothesis at the 10% significance level, with p-values 0.3899 and 0.2815, respectively.

Table 1: Graph-based quantities for 5-MST under Scenario 1 and 2.

	R_1	E(R_1)	Var(R_1)	R_2	E(R_2)	Var(R_2)
Scenario 1	956	497.5	1978.756	144	497.5	1978.756
Scenario 2	571	497.5	2872.114	431	497.5	2872.114

Table 1 sheds insight into why this happens. Two graph-based within-sample edge counts R_1 and R_2 and their expectations under the null $E(R_1)$ and $E(R_2)$ are reported. Under the alternative, we would expect the absolute value of differences between the within-sample edge counts and their null expectations to be relatively large ($|R_1 - E(R_1)|$ and/or $|R_2 - E(R_2)|$). Figure 2 illustrates how the edge counts behave in the two scenarios. We plot only those edges that are connected to hubs - which we define in this setting to

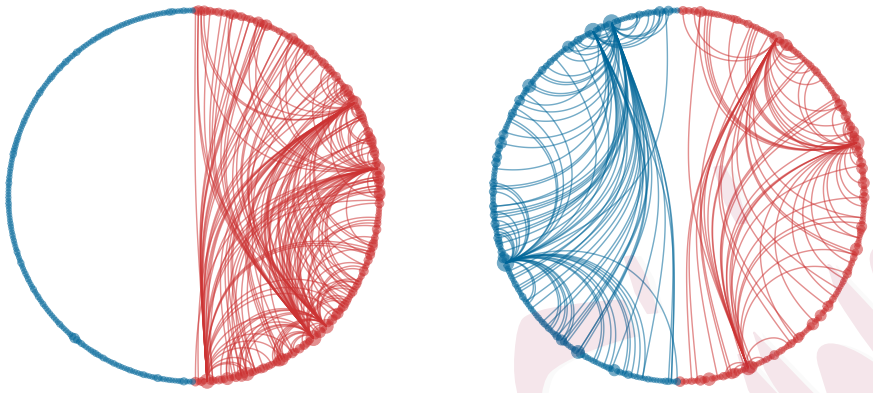


Figure 2: Illustration of edges connected to hubs (defined in this setting to be nodes with node degree larger than 50) in the similarity graph for Scenario 1 (left) and Scenario 2 (right). Hubs from Samples 1 and 2 are represented by red and blue points, respectively, along the circle perimeter. The size of the point corresponds to the node degree. Edges connecting observations from Sample 1 are in red and from Sample 2 are in blue.

be any node with a degree larger than 50.

We can see in Scenario 1, R_1 and R_2 behave as we expect. Observe that in Figure 2, hubs are generated in Sample 1 with many within-sample connections (so we can see many red edges), making R_1 large. On the other hand, most of the observations in Sample 2 (with a larger variance) connect to those in Sample 1, making R_2 small (we do not see any blue edges). Then, the differences between R_1 and R_2 and their respective expectations are relatively large as shown in the first row of Table 1, and it follows that existing tests have the power to reject the null.

In Scenario 2, two problems arise in the presence of a hub. First, hubs tend to form in

both samples. A hub in the sample with a larger variance will form edges with observations from the same sample, which increases R_2 (we see more blue edges in Figure 2), and form between-sample edges (decreasing R_1). The relative differences between the edge counts and their respective expectations in Scenario 2 become smaller as shown in the second row of Table 1, causing both tests to lose power. Second, the variances of both R_1 and R_2 increase, further inhibiting the power of the test statistics. This second problem can be clearly seen by studying the analytical expression of the variance of the edge counts, R_j $j = 1, 2$, under the permutation null distribution. Let G denote the similarity graph and its set of edges, $|G|$ denote the number of edges in G, G_i be the subgraph including all edge(s) that connect to node i, and $|G_i|$ be the degree of node i in G. The variance expression of R_j is:

$$\begin{aligned}\mathrm{Var}(R_j) =&[2C(N-n_j) + |G|(|G|-1)(n_j-3)]\frac{n_j(n_j-1)(n_j-2)}{N(N-1)(N-2)(N-3)} + \\ &\mu_j(1-\mu_j),\end{aligned}$$

where $j = 1, 2$, $\mu_j = E(R_j) = |G|\frac{n_j(n_j-1)}{N(N-1)}$, $C = \frac{1}{2}\sum_{i=1}^{N}|G_i|^2 - |G|$, n_j is the number of observations in sample j, and $N = n_1 + n_2$. In the presence of a hub, both $\sum_{i=1}^{N}|G_i|^2$ and C increase, where C represents the number of edge pairs sharing a common node, which in turn results in an inflated variance for R_1 and R_2.

2.4 Our Contribution

When the size or density of hubs is large, existing graph-based tests can suffer from limited power and unstable performance. We propose new test statistics that are useful even in the

presence of hubs. Specifically, we propose to apply appropriate weights to the test statistics that will dampen the effect of hubs while still retaining crucial similarity information. We show that these weights can improve power and resolve the variance boosting problem in the presence of problematic graph structures. We provide recommendations for weights as a function of node degrees and demonstrate that these work well in a range of scenarios. The limiting null distribution of these new robust test statistics is derived under mild conditions on the weights, and we show that the limiting distribution is quite accurate for finite sample sizes. Unless stated otherwise, we use the 5-MST constructed from L_2 distances of the pooled observations as the similarity graph in simulations. The robust edge-count tests can be implemented using the R package 'rgTest'. Code for simulations and our application is available at `https://github.com/stat-yb/robustEtest.git`.

3. Robust edge-count test statistics

Our approach is to flatten the similarity graph in order to limit the influence of hubs without incurring too much of a loss of similarity information so that the testing procedure can still retain power. To do so, we propose to apply weights that are functions of the graph's node degrees to the edge-count test statistics. The weights should be designed such that edges connected to a hub are down-weighted, while other edges are left mostly undisturbed. Let d_i denote the node degree of node i in a graph G. Let (i, j) represent the edge connecting observations i and j in graph G. Let w_{ij} denote the weight on edge (i, j) where w_{ij} is the value of the weight function $W(d_i, d_j)$, with $W(d_i, d_j)$ defined to be

a function of d_i and d_j. Discussions about the choice of weight functions are deferred to Section 3.1.

We apply weights w_{ij} to the edge counts $R_1(i,j)$, and $R_2(i,j)$, such that each edge $(i,j) \in G$ is weighted by a combination of d_i and d_j. Let $g_i = 0$ if the observation i is from Sample $\boldsymbol{X}$, and 1 otherwise. Let n_1 be the sample size of Sample $\boldsymbol{X}$, n_2 be the sample size of Sample $\boldsymbol{Y}$, and $N = n_1 + n_2$. We define

$$R_1^w = \sum_{(i,j)\in G} R_1(i,j),\ R_2^w = \sum_{(i,j)\in G} R_2(i,j).$$

where $R_1(i,j) = w_{ij} I(J_{(i,j)} = 1),\ R_2(i,j) = w_{ij} I(J_{(i,j)} = 2)$, and

$$J_{(i,j)} = \begin{cases} 1 & \text{if } g_i = g_j = 0, \\ 2 & \text{if } g_i = g_j = 1. \end{cases}$$

The robust generalized edge-count test statistic is defined to be:

$$S_R = (R_1^w - \mu_1^w, R_2^w - \mu_2^w)(\Sigma^w)^{-1} \begin{pmatrix} R_1^w - \mu_1^w \\ R_2^w - \mu_2^w \end{pmatrix},$$

where $\mu_1^w = E(R_1^w), \mu_2^w = E(R_2^w)$, and

$$\Sigma^w = \begin{pmatrix} \mathrm{Var}(R_1^w) & \mathrm{Cov}(R_1^w, R_2^w) \\ \mathrm{Cov}(R_1^w, R_2^w) & \mathrm{Var}(R_2^w) \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Theorem 1. *S_R can be expressed as*

$$S_R = (Z_{diff}^R)^2 + (Z_w^R)^2,$$

with $Cov(Z_{diff}^R, Z_w^R) = 0$, where

$$Z_{diff}^R = [(R_1^w - R_2^w) - E(R_1^w - R_2^w)]/[Var(R_1^w - R_2^w)]^{1/2},$$

$$Z_w^R = [(qR_1^w + pR_2^w) - E(qR_1^w + pR_2^w)]/[Var(qR_1^w + pR_2^w)]^{1/2},$$

$p = (n_1 - 1)/(N - 2)$, *and* $q = 1 - p$.

The proof of Theorem 1 can be found in the Supplementary Section S5.

Theorem 1 leads us to propose the robust max-type edge-count test statistic:

$$M_R = \max(Z_w^R, |Z_{\text{diff}}^R|).$$

If the graph is relatively flat and no hub is present, then d_i is similar for all $i \in [1, N]$, and the weights have little effect. However, in the presence of a problematic hub(s), the weights control the influence of edges connected to the hub, resulting in improved and reliable performance. This creates a test statistic that is increasingly robust to the underlying similarity graph and also resolves the variance boosting problem.

The analytic expressions of expectations and variances involved above can be obtained by combinatorial analysis under the permutation null distribution. We present them in the following lemma.

Lemma 1. *Under the permutation distribution, we have:*

$$\mu_1^w = \sum_{(i,j)\in G} w_{ij} \frac{n_1(n_1-1)}{N(N-1)},\ \mu_2^w = \sum_{(i,j)\in G} w_{ij} \frac{n_2(n_2-1)}{N(N-1)},$$

$$\Sigma_{11} = [-S_2 + \frac{2(2N-3)}{N(N-1)} S_3 + \frac{N-3}{n_2-1}(S_1 + S_2) - \frac{4(N-3)}{N(n_2-1)} S_3] D_N,$$

$$\Sigma_{22} = [-S_2 + \frac{2(2N-3)}{N(N-1)} S_3 + \frac{N-3}{n_1-1}(S_1 + S_2) - \frac{4(N-3)}{N(n_1-1)} S_3] D_N,$$

$$\Sigma_{12} = [-S_2 + \frac{2(2N-3)}{N(N-1)} S_3] D_N,$$

where $S_1 = \sum_{(i,j)\in G} w_{ij}^2$, $S_2 = \sum_{(i,j),(i,k)\in G} w_{ij} w_{ik}$, $S_3 = \sum_{(i,j),(k,l)\in G} w_{ij} w_{kl}$ *and* $D_N = [n_1 n_2 (n_1 - 1)(n_2 - 1)]/[N(N-1)(N-2)(N-3)]$.

The proof of this lemma can be found in the Supplementary Section S3. Using the results from Lemma 1, the expectations and variances involved in Z^R_{diff} and Z^R_w can be obtained as follows:

$$E(R^w_1 - R^w_2) = \sum_{(i,j)\in G} w_{ij}\frac{n_1 - n_2}{N},\ E(qR^w_1 + pR^w_2) = \sum_{(i,j)\in G} w_{ij}\frac{(n_1-1)(n_2-1)}{(N-1)(N-2)},$$

$$\text{Var}(R^w_1 - R^w_2) = [(S_1 + S_2) - \frac{4}{N}S_3]\frac{n_1 n_2}{N(N-1)},$$

$$\text{Var}(qR^w_1 + pR^w_2) = [\frac{N-3}{N-2}S_1 - \frac{S_2}{N-2} + \frac{2S_3}{(N-1)(N-2)}]D_N.$$

Large values of S_R and M_R are evidence against the null hypothesis of no distributional difference. The constructions of both S_R and M_R allow them to be powerful for general alternatives. When there is a change in the mean, both R^w_1 and R^w_2 tend to be larger than their null expectation - it follows that Z^R_w will be large, which leads to a large S_R and M_R. When a change in variance is present, without loss of generality, suppose the sample with the smaller variance is sample $\boldsymbol{X}$. Then R^w_1 is relatively large compared to its null expectation while R^w_2 is relatively small. In this case, $|Z^R_{\text{diff}}|$ tends to be large, which also leads to a large S_R and M_R. The robust test statistics S_R and M_R default to the tests proposed in S and M when $w_{ij} = 1$ for all (i, j). In S and M, each edge has an equal contribution to the test statistic so that even those edges connected to problematic hubs are treated with the same weight as those that are not. By placing weights on the edges, we dampen the influence of hubs and effectively flatten the graph.

Remark 1. The test statistics are well-defined under the following conditions:

(a.) $\sum_{\{j,\text{s.t.}(i,j)\in G\}} w_{ij}$ are not all equal for all $i \in [1, N]$;

(b.) $(N-3)S_1 - S_2 + \frac{2}{N-1}S_3 > 0$, where $S_1 = \sum_{(i,j)\in G} w_{ij}^2$, $S_2 = \sum_{(i,j),(i,k)\in G} w_{ij}w_{ik}$ and $S_3 = \sum_{(i,j),(k,l)\in G} w_{ij}w_{kl}$.

The proof of this remark can be found in the Supplementary Section S4. For example, for a completely flat graph (all nodes have the same degree), then $w_{ij} = w_{i'j'}$, $\forall i \neq j \neq i' \neq j' \in [1, N]$ and Z^R_{diff} is not well-defined. For a star-shaped graph, in which all observations connect to the same node, Z^R_w is not well-defined. Theorem 2 ensure $R^w_j, j = 1, 2$ does not vanish to zero when the sample size goes to infinity. The proof of this theorem can be found in the Supplementary Section S6.

Theorem 2. *Let W be a weight function such that $W(i, j) = w_{ij}$, $\forall (i, j) \in G$. If the weight function W is asymptotically bounded below by $1/|G|$ as $N \to \infty$, then $\lim_{N\to\infty} R^w_s > 0$, for $s = 1, 2$.*

3.1 Proposed Weights

The test statistics are defined for general weights that are functions of the node degrees and monotonically decreasing, as defined below.

Definition 1. A bivariate function is called monotonically decreasing if for all x_1, x_2 and y_0 such that $x_1 < x_2$, then $f(x_1, y_0) > f(x_2, y_0)$; and for all y_1, y_2 and x_0 such that $y_1 < y_2$, then $f(x_0, y_1) > f(x_0, y_2)$.

In practice, users have the flexibility to choose their weights, provided that the test statistics are well-defined given the conditions in Remark 1. Since we allow the graph to be general, and the weights are properties of the graph, obtaining optimal weights for

general similarity graphs is challenging, and we reserve this line of theoretical analysis for future work. Instead, we provide recommendations for data-driven weights based on empirical studies. We recommend a weight that (1) demonstrates reasonable power and (2) meets the conditions for our asymptotic theory.

For edge (i, j), we recommend the following weight function:

$$W(d_i, d_j) = \frac{1}{\max(d_i, d_j)}. \tag{3.1}$$

The weight function W is bounded below by $1/|G|$ asymptotically and monotonically decreasing.

We present the following examples to demonstrate how the weight function works in the robust test and its utility. First, we present an example to show how weights can temper the impact of hubness on the variance. A dataset with 100 observations is simulated from a 100-dimensional uniform distribution. According to Lemma 1, the change in the variance of the test statistics is contingent on the change in S_1, S_2, and S_3 for different similarity graph structures. When applying equal weights (which effectively treats all the edges as equal since w_{ij} equals a constant c for all $(i, j) \in G$), S_1 and S_3 are constant given a fixed number of edges, and any hubness in the similarity graph only affects S_2. In Figure 3, the boxplots of $\sum_{\{j,k:j\neq k,(i,j),(i,k)\in G\}} w_{ij}w_{ik}$ for $i \in [1, N]$, which is the dominant component in S_2, compares this quantity under equal weights and the weight function W (3.1). There are several observations that form hubs in this setting. When using equal weights, it is clear that these hubs are still present and the variance boosting problem continues to manifest itself with large values of $\sum_{\{j,k:j\neq k,(i,j),(i,k)\in G\}} w_{ij}w_{ik}$. On the

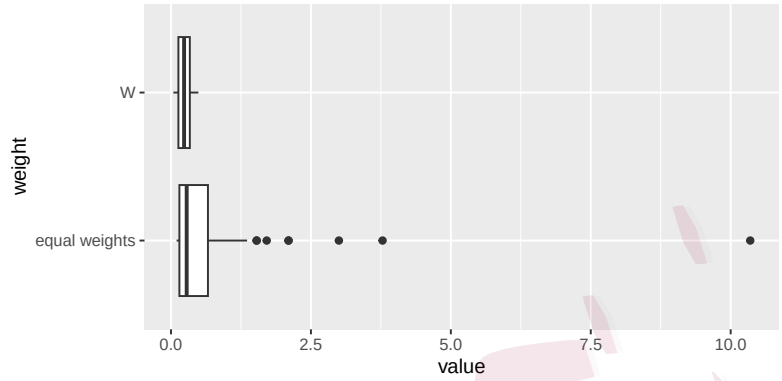


Figure 3: The boxplots of $\sum_{\{j,k:j\neq k,(i,j),(i,k)\in G\}} w_{ij}w_{ik}$ for $i \in [1, N]$ using equal weights and weighted function W.

other hand, when applying the weight function, the impact of the hubs is well-controlled.

To more comprehensively evaluate the performance of the weight function, we simulate 500 replications of two samples with $n_1 = 100$, $n_2 = 100$, and $d = 400$ from the d-dimensional log-normal distribution. The difference between the two samples is reflected by Δ_μ, where μ is the expected value of the variable's natural logarithm. The difference is equal across all dimensions such that $||\Delta_\mu||^2 = 2.5$. We record the maximum node degree of the similarity graph in each simulation, and group the simulations according to their maximum node degrees from low to high by each tenth percentile. The increase in the maximum node degree is indicative of a more severe hubness phenomenon.

Figure 4 presents boxplots of the difference between the robust within-sample edge counts and their expectations and the variances of robust within-sample edge counts. We compare these with their corresponding quantities using equal weights. Under the alternative, we anticipate a relatively large difference between the within-sample edge counts

and their expectations. Under equal weights, as the maximum node degree increases, the relative difference decreases in Sample 1. However, the boxplots using W do not exhibit a similar trend in Sample 1, which suggests higher power. Under equal weights, the variance also increases as the node degree of the hub increases. On the other hand, it is clear the weight function W controls the variance from increasing.

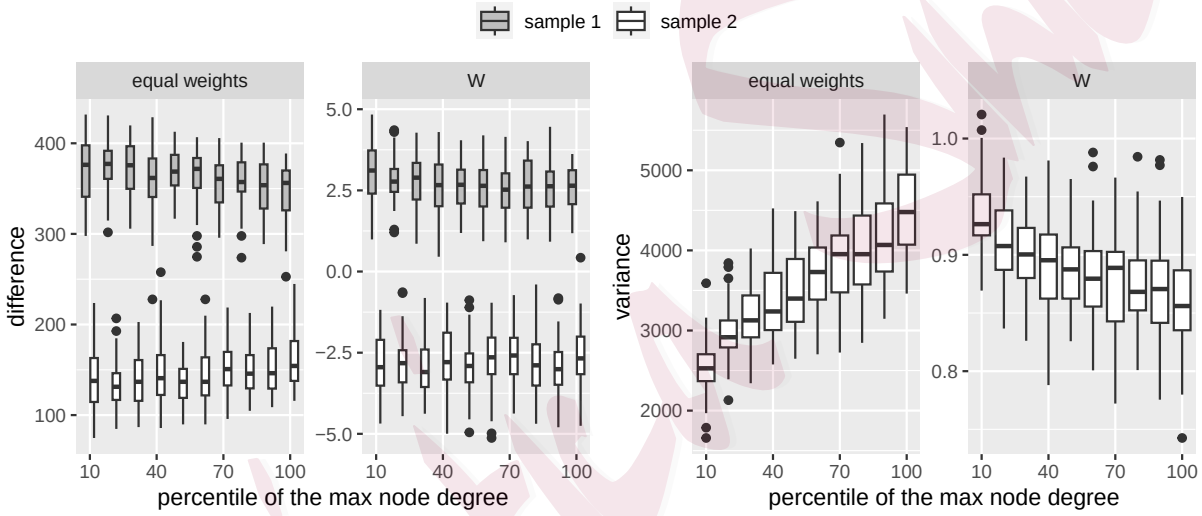


Figure 4: Boxplots of $R_j^w - E(R_j^w)$ (left) and $\mathrm{Var}(R_j^w)$ (right), $j = 1, 2$, using weights (W) compared to boxplots of corresponding quantities using equal weights. Simulations are grouped according to the percentile of the max node degrees. Only variances of sample 2 are presented since the sample sizes are equal, and the variances are roughly the same for both samples.

4. Asymptotic null distribution

The robust edge-count test statistics are computationally straightforward to calculate and their significance can be obtained via resampling from the permutation distribution. However, as the sample size increases, permutation becomes increasingly computationally prohibitive. To make the tests practical for modern data sets, we study the limiting distributions of the robust edge-count test statistics.

We define

- $A_{(i,j)} = \{(i,j)\} \cup \{(i',j') \in G, (i,j) \text{ and } (i',j') \text{ share a node}\}$,
- $B_{(i,j)} = A_{(i,j)} \cup \{(i'',j'') \in G, \exists (i',j') \in A_{(i,j)}, \text{ such that } (i',j') \text{ and } (i'',j'') \text{ share a node}\}$,
- $W(A_{(i,j)}) = \sum_{(i',j') \in A_{(i,j)}} w_{i'j'}$, and $W(B_{(i,j)}) = \sum_{(i'',j'') \in B_{(i,j)}} w_{i''j''}$.

Theorem 3. *Under conditions:*

(i) $G = \mathcal{O}(N^{\alpha}), 1 \leq \alpha < 1.25$,

(ii) $S_1 + S_2 - \frac{4}{N} S_3 = \mathcal{O}(S_1 + S_2)$,

(iii) $\sum_{(i,j) \in G} (w_{ij} |A_{(i,j)}|)^2 = o(S_1 \sqrt{N})$,

(iv) $\sum_{(i,j) \in G} w_{ij} W(A_{(i,j)}) W(B_{(i,j)}) = o(S_1)^{1.5}$,

as n_1, n_2, $N \to \infty$ *and* $n_1/N \to \lambda \in (0,1)$, $Z_w^R \xrightarrow{\mathcal{D}} N(0,1)$ *and* $Z_{diff}^R \xrightarrow{\mathcal{D}} N(0,1)$ *under the permutation null distribution.*

The proof utilizes Stein's theorem from Chen and Shao (2005), and details are provided in the Supplementary Section S7.

Corollary 1. *Under the conditions given in Theorem 3, as* $n_1, n_2, N \to \infty$ *and* $n_1/N \to \lambda \in (0,1)$, $S_R \xrightarrow{\mathcal{D}} \chi_2$ *under the permutation null distribution.*

Condition (ii) ensures Z^R_{diff} is asymptotically well-defined. The condition is automatically met when using the proposed weight function W (3.1). Utilizing the proposed weight ensures S_3 is bounded by a constant independent of N.

Conditions (iii) and (iv) prevent the sum of weights in the hub from growing too large. To see that the conditions hold easily in the presence of hubs in high dimensions, we generated data from the normal distribution, uniform distribution, log-normal distributions, and heavy-tailed t distributions. Ratios of the key quantities involved in the conditions are shown in Figure 5. Once we assign weights, the ratios $\sum(w_{ij}|A_{(i,j)}|)^2/(S_1\sqrt{N})$ and $(\sum w_{ij}W(A_{(i,j)})W(B_{(i,j)})/(S_1)^{1.5}$ are bounded by $o(1)$ as N increases under all scenarios.

To evaluate the accuracy of our asymptotic theory for finite sample sizes, we compare the critical values generated from 10,000 permutations with those obtained using our asymptotic theory under the null hypothesis. The boxplots of the differences between asymptotic and permutation critical values are shown in Figure 6. We observe that the p-value approximations are reasonable based on the small differences shown in the boxplots.

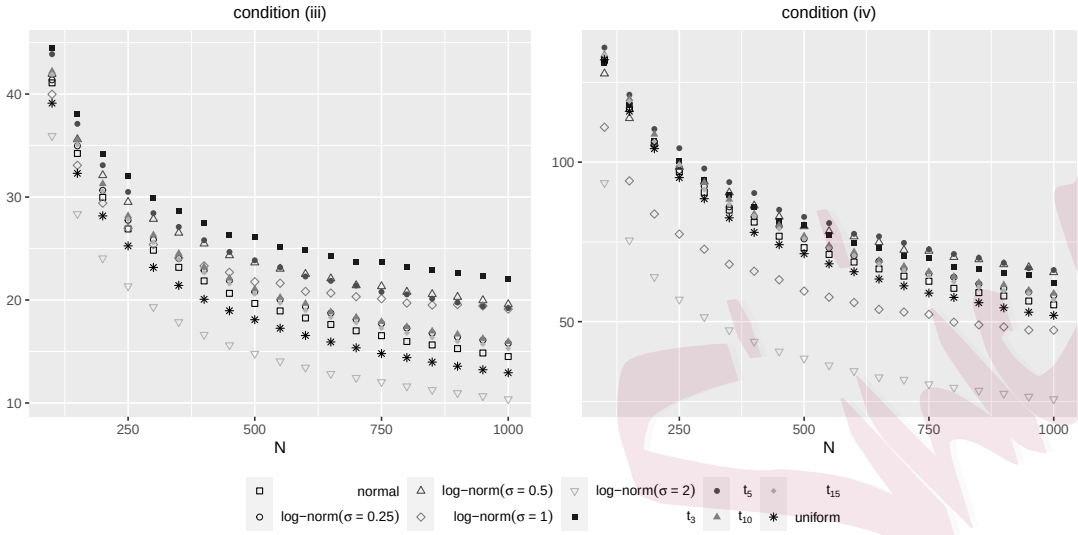


Figure 5: Ratios of key quantities for the proposed robust test statistics generated using data from normal distribution, uniform distribution, log-normal distributions with different skewness levels (controlled by σ), and heavy-tailed t distributions with varying degrees of freedom. The dimension of each observation is $d = N$. Left: the ratio of $\sum(w_{ij}|A_{(i,j)}|)^2$ to $S_1 N^{0.5}$. Right: the ratio of $\sum w_{ij} W(A_{(i,j)}) W(B_{(i,j)})$ to $(S_1)^{1.5}$.

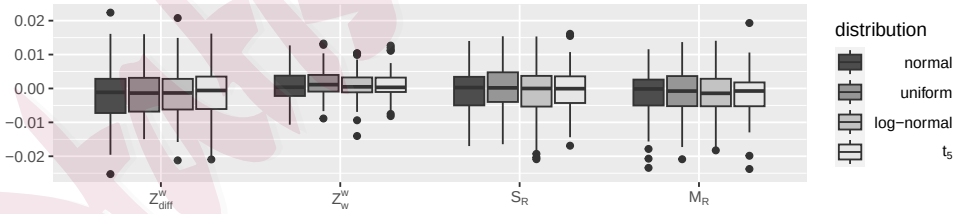


Figure 6: Boxplots of differences between asymptotic critical values and permutation critical values. Data are generated from different distributions with $n_1 = n_2 = 100$ and $d = 100$.

5. Performance Analysis

5.1 Hubs in high-dimensional data

We examine the performance of the robust edge-count test statistics on high-dimensional data. We present the power of the tests, which is estimated to be the number of trials (out of 100) with significance less than 5%. We also report the median of the maximum node degrees of 5-MST in each trial (over 100 trials), denoted as $\tilde{d}_{\max}$. We compare the robust tests S_R and M_R with the following tests: MMD (Gretton et al. (2012)), energy (Székely et al. (2004)), the generalized edge-count test S (Chen and Friedman (2017)), the max-type edge-count test M (Chu and Chen (2019)), and rank-based tests R_g-NN and R_o-MST (Zhou and Chen (2023)). The test statistics R_g-NN and R_o-MST also apply weights to a similarity graph (NN or MST) in the form of ranks. However, their ranking weights are not designed to mitigate problematic structures in the graph and, as we'll demonstrate, can still suffer from reduced power in some scenarios. For R_g-NN, we follow the authors' recommendation and use the 10-NN graph. The detailed settings of the simulations are as follows:

- Simulation I and III: Observations are generated from multivariate log-normal distributions. $\mathbf{X} \sim \exp(\mathcal{N}(\mathbf{1}_d, 0.6\mathbf{I}_d))$, $\mathbf{Y} \sim \exp(\mathcal{N}((1+\sqrt{0.01log(d)/d})\mathbf{1}_d, (0.6+1.8log(d)/d)\mathbf{I}_d))$, where d denotes the dimension. $n_1 = n_2 = 100$.

- Simulation II and IV: Observations are generated from multivariate mixture Gaussian distributions. $\mathbf{X} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, $\mathbf{Y} \sim 0.1\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) + 0.9\mathcal{N}(\sqrt{0.1log(d)/d}\mathbf{1}_d, (1+$

$2.5log(d)/d)\mathbf{I}_d)$, where d denotes the dimension. $n_1 = n_2 = 100$.

- Simulation V: Observations are generated from multivariate Gaussian distributions: $\mathbf{X} : \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, $\mathbf{Y} : \mathcal{N}((\sqrt{\boldsymbol{(0.2log(s)/s)}}\mathbf{1_s}, \mathbf{0_{d-s}}), \begin{pmatrix} (1+3log(s)/s)^2\mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-s} \end{pmatrix})$, where $s = \lfloor\sqrt{d}\rfloor$. $n_1 = n_2 = 200$.

The results for Simulations I and II are presented in Table 2 and 3. In both settings, there is a mean and variance change. For each of the 100 trials, the maximum node degree appears in the sample with a larger variance. Both simulations exhibit a pattern of hubness in the high-dimensional setting, which creates difficulties for the existing two-sample tests. For log-normal data, the hubness is more pronounced as the dimension increases ($\tilde{d}_{\max}$ is quite large). The MMD test struggles in this setting. When the dimension is moderate ($d = 500$), the remaining non-parametric tests perform reasonably. However as d increases, the power for some of the tests begins to suffer. We observe that as d increases, the robust edge-count tests S_R and M_R outperform all other methods and the gap becomes more pronounced for larger d. We observe a similar pattern in Table 3: as d increases the robust edge-count tests have considerable power gains compared to other methods.

The results for Simulations III and IV are presented in Table 4 and 5. These settings are similar to Simulations I and II but allow the maximum node degree to appear in either sample. Under these scenarios, the robust edge-count tests show comparable results to R_o-MST while outperforming other tests for the log-normal data (see Table 4). In Table 5, we observe that the robust edge-count tests excel when compared to other tests for the mixture Gaussian data as the dimension of the observation increases.

Table 2: Simulation I: number of trials that reject the null with $\alpha = 0.05$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	111.5	6	75	55	66	67	89	97	**98**
800	128	4	51	40	48	46	81	89	**93**
1100	124	6	33	36	39	36	77	**90**	**90**
1400	131	4	19	20	28	18	65	87	**90**
1700	127.5	5	17	10	15	13	53	77	**80**
2000	134.5	7	14	15	24	18	54	70	**76**

Table 3: Simulation II: number of trials that reject the null with $\alpha = 0.05$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	69	26	30	84	84	84	91	**100**	**100**
800	71	14	19	56	68	55	76	92	**93**
1100	74	8	20	53	49	52	62	84	**91**
1400	71	6	14	35	43	35	56	80	**86**
1700	71	8	15	27	31	31	50	66	**73**
2000	71	5	17	23	26	23	38	64	**65**

Table 6 presents the performance under Simulation V. We simulate observations where the change does not occur in all dimensions; this setting can easily induce a large hub in

Table 4: Simulation III: number of trials that reject the null with $\alpha = 0.05$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	123	50	83	91	91	93	**100**	98	99
800	130	44	61	77	83	81	96	95	**97**
1100	137	37	45	81	82	81	94	94	**98**
1400	134	35	29	71	78	70	**94**	89	**94**
1700	137.5	32	19	62	72	62	**90**	85	**90**
2000	143.5	26	19	50	63	54	85	80	**86**

Table 5: Simulation IV: number of trials that reject the null with $\alpha = 0.05$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	74	49	25	96	**99**	96	**99**	98	98
800	74.5	38	20	90	92	89	96	99	**100**
1100	72.5	24	13	76	89	77	86	91	**94**
1400	75	25	14	71	75	70	83	86	**90**
1700	72	29	15	78	82	78	87	83	**88**
2000	72	17	11	53	55	47	65	73	**77**

the similarity graph when the dimension is high. Similar to before, when the dimension is not too high ($d = 500$) all the tests have comparable power. But as d increases, we

see that the robust tests start to out-compete most of the other graph-based tests. When $d = 2000$, it is evident that the robust tests have the superior power.

Table 6: Simulation V: number of trials that reject the null with $\alpha = 0.05$.

d	$\tilde{d}_{\max}$	MMD	Energy	S	M	R_g-NN	R_o-MST	S_R	M_R
500	107	47	65	**100**	**100**	**100**	**100**	**100**	**100**
800	109	41	61	91	92	93	**98**	96	96
1100	112	34	46	86	89	87	92	**95**	**95**
1400	109	34	45	75	79	78	87	87	**92**
1700	107	21	42	72	79	70	84	84	**87**
2000	108	20	33	67	70	68	77	84	**85**

The robust tests are also well-designed to deal with the hubness phenomenon under imbalanced sample sizes as explored in previous studies Chen et al. (2018); Banerjee et al. (2020, 2024). In particular, the test statistic Z_w^R is constructed to mitigate any power loss from imbalanced samples. Since S_R and M_R are functions of Z_w^R, both test statistics are equipped to handle the imbalanced setting and hubness phenomenon for general changes. Under the imbalanced setting, the larger sample is more likely to develop a hub. Additional simulations demonstrating the performance of the robust graph-based tests under imbalanced sample sizes are provided in Supplementary Section S1.2.

5.2 Calibration Under the Null Hypothesis

To assess the calibration of robust edge-count tests under the null hypothesis, we simulate two samples with $n_1 = n_2 = 100$ from the standard normal distribution. In Table 7, the number of trials (out of 1000) to reject the null (at $\alpha = 0.05$) are reported for both asymptotic and permutation critical values. Rejection rates are around 5% for dimensions ranging from 600 to 2000, indicating that the type I error rate is well-controlled.

Table 7: Number of trials (out of 1000) that reject the null with $\alpha = 0.05$ under the null hypothesis.

	d	600	800	1000	1200	1400	1600	1800	2000
Permutation	S_R	45	48	53	42	42	38	55	49
	M_R	44	43	62	53	46	45	51	46
Asymptotic	S_R	39	45	51	41	41	38	57	48
	M_R	45	42	59	50	46	44	50	46

5.3 Consistency of the proposed tests

The robust edge-count tests show increasing power as the number of observations grows. We simulate samples with various sample sizes to exhibit the consistency of the test. The simulated data are generated from log-normal distributions with $\mathbf{X} \sim \exp(\mathcal{N}(\mathbf{1}_d, 0.6\mathbf{I}_d))$, $\mathbf{Y} \sim \exp(\mathcal{N}((1+\sqrt{0.01log(d)/d})\mathbf{1}_d, (0.6+1.8log(d)/d)\mathbf{I}_d))$, and mixture Gaussian distri-

butions with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$, $\mathbf{Y} \sim \lfloor 0.1n_2 \rfloor \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) + \lfloor 0.9n_2 \rfloor \mathcal{N}(\sqrt{0.1log(d)/d}\mathbf{1}_d, (1 + 2.5log(d)/d)\mathbf{I}_d)$, where $d = 2000$. The powers of the robust edge-count tests at 5% significance level for 100 simulations are presented in Table 8. With more observations, the number of rejections increases and quickly approaches 100, even for moderate sample sizes, demonstrating the consistency of the proposed tests.

Following Theorem 5.2.1 from Chen and Friedman (2017), it is straightforward to show that the robust edge-count tests are consistent against all alternatives on k-MST with $k = O(1)$.

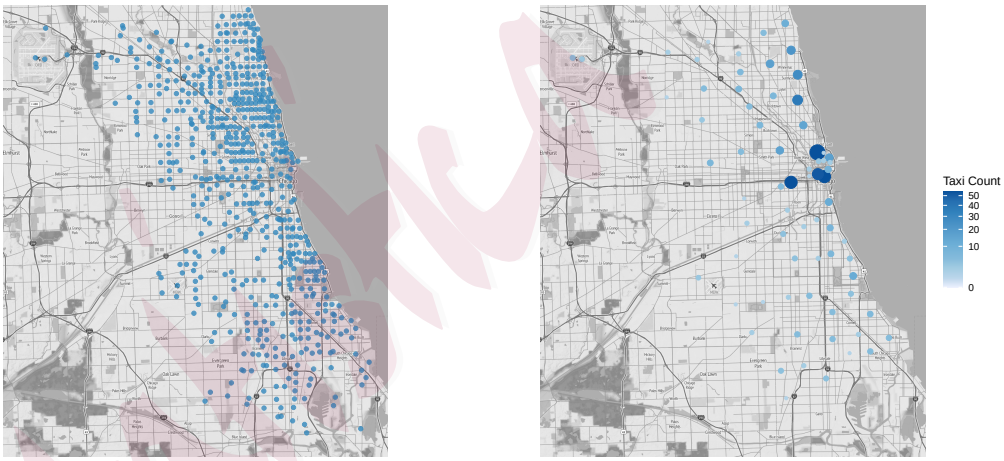
Table 8: Number of trials (out of 100) that reject the null for $\alpha = 0.05$ as $N = n_1 + n_2$ increases.

$n_1 = n_2$		100	125	150	175	200	225	250	275	300
log-normal	S_R	76	76	93	93	94	97	100	100	100
	M_R	85	84	95	93	98	97	100	100	100
mixture Gaussian	S_R	83	92	91	99	99	99	99	100	100
	M_R	85	94	90	99	100	99	99	100	100

6. Real Data Application

We illustrate the robust graph-based tests on the Chicago taxi trip dataset in 2020. This data is publicly available on the Chicago Data Portal website (`https://data.cityofc`

hicago.org/Transportation/Taxi-Trips/wrvz-psew) and includes drop-off dates, times, and locations for each taxi trip. There are 635 unique drop-off locations (as shown in Figure 7a). We count the frequency of taxi drop-offs in each location for a specified time interval. Each observation is a 635×1 vector of taxi trip counts that occur within a time interval for each day; each element of the vector represents the number of drop-offs at a specific drop-off location. In Figure 7a, the position of the dot indicates the location of taxi drop-offs. Figure 7b shows an example when the time interval is set to be 7 am - 10 am on September 1. The size and color of the dot indicate the number of trips at that location.



(a) Pick-up locations in Chicago. (b) Frequency of taxi dropoffs.

Figure 7: Left: plot of all drop-off locations. Right: plot of the number of taxi trips that occurred from 7 am to 10 am on September 1st at all possible drop-off locations. The larger dot size indicates more trips took place at the location. A color key is also provided.

Since taxi trips may not happen in some locations for a specified time interval, it is

often the case that many entries in our vector of observations are 0 or very low counts. Given that the dimension of the observation is much larger than the number of observations in each sample, a large node degree is likely to arise when constructing the similarity graph. Since the underlying data distribution is unknown, it is difficult to identify problematic hubs just by examining the constructed similarity graph. We will demonstrate that the robust edge-count test can circumvent any hub-related issues and lead to reasonable inference.

To illustrate the new tests, we consider two different scenarios and compare the performance of the new tests with existing graph-based tests, as well as the energy test and MMD test. For similarity graphs, we use 5-MST constructed from the L_1 distance between observations. For all tests, p-values are obtained via 10,000 permutations.

6.1 Scenario I

We compare the taxi drop-offs in morning rush hours from 7 am to 10 am between September and November. Sample 1 consists of the number of taxi drop-offs that occurred during morning rush hour in September. Each day is an observation, resulting in 30 observations ($n_1 = 30$). Sample 2 consists of the number of taxi drop-offs that occurred in the morning hours in November, with each day being an observation ($n_2 = 30$). The dimension of each observation is 635, which is clearly far more than the number of observations. The heatmaps of the taxi counts in each district are shown in Figure 8. The changes are subtle but taxi trips in September appear busier and more dispersed than in November. While

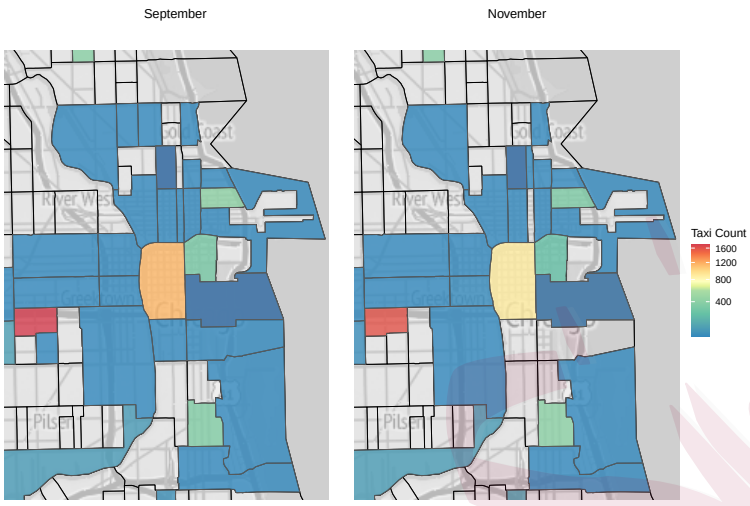


Figure 8: Heatmap illustrating the number of taxi trips in each district for the month of September (left) and November (right).

we might be able to visualize this difference between months, what we want to effectively discern is whether this change in distribution is meaningful or just by random chance.

To address this question, Table 9 presents the two-sample test results. With a max node degree of 19 in the similarity graph, all graph-based tests provide significant evidence in favor of a difference between September and November at a 10% significance level. However, the energy and MMD tests cannot reject the null hypothesis. This demonstrates, at least in this setting, that the graph-based methods show superior performance as omnibus tests when comparing high-dimensional distributions.

As a sanity check, to demonstrate that the tests are well-calibrated, we randomly split the morning rush hour taxi dropoffs in September and November into two samples. As shown in table 10, all tests fail to reject the null hypothesis of no difference at a 10%

Table 9: P-values for tests comparing taxi dropoffs in morning rush hours between September and November.

MMD	Energy	S	M	S_R	M_R
0.3871	0.1696	0.0032	0.0024	0.0024	0.0025

significance level.

Table 10: P-values for tests comparing randomly split taxi drop-offs in morning rush hours in September and November.

MMD	Energy	S	M	S_R	M_R
0.9293	0.5720	0.9933	0.9217	0.9912	0.9630

6.2 Scenario II

The statewide stay-at-home order signed by the Illinois Governor took effect on March 21 in response to the spread of COVID-19, leading to a sharp decline in Chicago taxi trips. This is a setting where large hubs cause issues for the existing graph-based tests. In the early morning hours (1 am - 5 am), the number of taxi rides is relatively low, especially after the lockdown in March; this sparse taxi activity induces the formation of hubs with large node degrees in the high-dimensional setting. We compare the number of taxi dropoffs during the early morning hours between weekdays and weekends over two

months (April and May). Sample 1 comprises the taxi dropoffs on 43 weekdays during early morning hours ($n_1 = 43$), while Sample 2 comprises the taxi dropoffs on 18 weekends during early morning hours ($n_2 = 18$).

Figure 9 displays the heatmap of the pairwise distance within two samples, indicating that the weekday observations tend to be closer (more similar) compared to weekends. We conjecture that there is a difference between the weekdays and weekends in the early morning, even post-lockdown. The similarity graph generated has several hubs with node degrees exceeding 20. Table 11 presents the two-sample test results. In this scenario, only the robust tests S_R and M_R provide evidence of a difference in travel patterns between the two samples at $\alpha = 0.1$. The MMD and energy test, while not significant, have p-values that are seemingly in the right direction compared to the other graph-based tests S and M.

Table 11: P-values for tests comparing taxi dropoffs on weekdays and weekends in April and May.

MMD	Energy	S	M	S_R	M_R
0.1289	0.1098	0.2617	0.2150	0.0223	0.0804

To better understand the behavior of the graph-based tests, we conduct a small sensitivity analysis to see how observations with large node degrees influence the tests' conclusions. One influential observation generating a hub with node degree of 31 is from April 26. As shown in Table 12, after removing this observation, the MMD test still

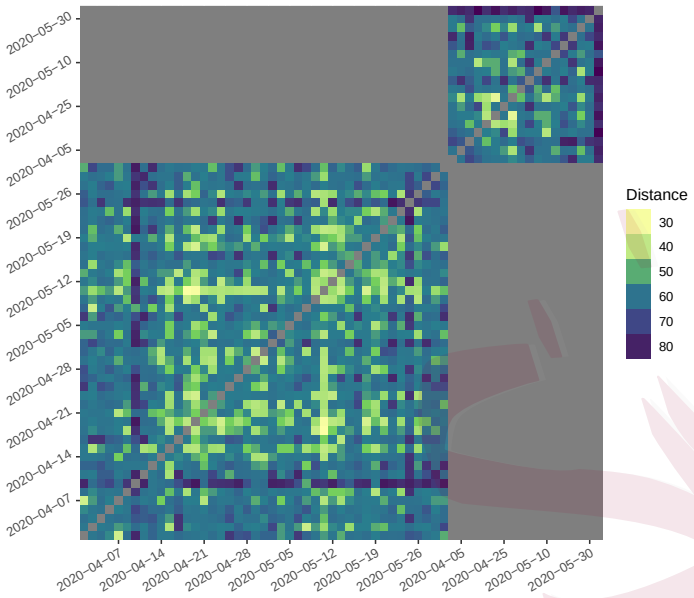


Figure 9: Heatmap of the pairwise distance within the two samples (weekdays versus weekends). The bottom left square shows the pairwise distance for the weekday taxi dropoffs, and the top right square for weekends. Lighter colors indicate closer distances. For ease of visualization, the between-sample distances are not shown.

cannot reject the null, while the other tests have significant test results rejecting the null at $\alpha = 0.1$.

Table 13 shows that for this setting, the tests are well-calibrated under the null. By randomly assigning the early morning hour taxi drop-offs in April and May into Sample 1 ($n_1 = 43$) and Sample 2 ($n_2 = 18$), all tests fail to reject the null at a 10% significance level.

In practice, identifying the influential observations, (such as the taxi drop-offs on April 26th), is challenging. The problematic observation may not necessarily be largest

Table 12: P-values for tests comparing taxi dropoffs on weekdays and weekends in April and May after removing activity on April 26th.

MMD	Energy	S	M	S_R	M_R
0.1217	0.0977	0.0432	0.0774	0.0149	0.0600

Table 13: P-values for tests comparing randomly split taxi dropoffs in the early morning hours on weekdays and on weekends in April and May.

MMD	Energy	S	M	S_R	M_R
0.9144	0.9467	0.3846	0.3587	0.5849	0.4393

hub (node with max degree) but a collection of hubs. The influence of an observation depends heavily on the connectivity of the graph and which edges are connected to this hub. While the inclusion and exclusion of potentially problematic observations may lead to conflicting results across existing tests, in contrast, the proposed robust test statistics (S_R and M_R) are shown to provide consistent and stable results. This is crucial for drawing statistical conclusions in real data applications where the ground truth is unknown.

7. Discussion

In this article, we propose robust edge-count two-sample tests that provide reliable inference even in the presence of problematic graph structures that arise as a product of

the curse of dimensionality. Our proposed robust tests can outperform the existing non-parametric tests in the presence of the hubs while providing comparable power even when the graph is relatively flat. The robust edge-count tests are constructed by applying weights that are functions of node degrees to the edge counts. A specific weight function with desirable properties is recommended.

These robust test statistics are computationally straightforward to calculate. While finite-sample p-values can be obtained via permutations, to make the test more computationally tractable, the limiting null distributions of the robust test statistics are derived under some mild conditions on the data-driven weights. Through empirical studies, these conditions are shown to be easily satisfied even in the presence of hubs. The p-value approximations based on asymptotic results are reasonably close to the permutation p-value for finite sample sizes, making the approach easy to apply to large data sets when permutation may be computationally prohibitive. Simulation studies show that the robust edge-count tests have power gains over existing edge-count tests when the dimension increases and hubs are more easily generated. An application of the tests on Chicago taxi data demonstrates the robust test statistics utility in high-dimensional settings.

Our results pave the way for future work in a few directions. It is of great interest to obtain optimal weights for robust graph-based test statistics. While this may be difficult to derive for generic similarity graphs, we may first focus on well-behaved graphs such as k-MSTs or k-NNs. Secondly, the study could be broadened to incorporate dense similarity graphs (where $k = O(n^{\alpha}), 0 \leq \alpha < 1$), which would require more careful theoretical

analysis. Lastly, the robust edge-count tests can be adapted to the scan statistic setting and applied to high-dimensional change-point problems, where the effect of the hubs over time may hamper our ability to effectively detect changes in distribution.

Supplementary Materials

The supplement contains additional simulations, figures, and technical proofs for Lemma 1, Remark 1, Theorem 1, Theorem 2 and Theorem 3.

References

Banerjee, T., B. B. Bhattacharya, and G. Mukherjee (2020). A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data. *The Annals of Applied Statistics 14*(4), 1777 – 1805.

Banerjee, T., B. B. Bhattacharya, and G. Mukherjee (2024). Bootstrapped edge count tests for nonparametric two-sample inference under heterogeneity. *Journal of Computational and Graphical Statistics*, 1–24.

Baringhaus, L. and C. Franz (2004). On a new multivariate two-sample test. *Journal of multivariate analysis 88*(1), 190–206.

Biswas, M. and A. K. Ghosh (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis 123*, 160–171.

Biswas, M., M. Mukhopadhyay, and A. K. Ghosh (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika 101*(4), 913–926.

Chen, H., X. Chen, and Y. Su (2018). A weighted edge-count two-sample test for multivariate and object data.

Journal of the American Statistical Association 113(523), 1146–1155.

Chen, H. and J. H. Friedman (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association 112*(517), 397–409.

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

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

Friedman, J. H. and L. C. Rafsky (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics 7*(4), 697–717.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *The Journal of Machine Learning Research 13*(1), 723–773.

Hall, P. and N. Tajvidi (2002). Permutation tests for equality of distributions in high-dimensional settings. *Biometrika 89*(2), 359–374.

Henze, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics 16*(2), 772–783.

Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika 105*(3), 529–546.

Liu, R. Y. and K. Singh (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association 88*(421), 252–260.

Radovanovic, M., A. Nanopoulos, and M. Ivanovic (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research 11*(sept), 2487–2531.

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.

Song, H. and H. Chen (2023, 11). Generalized kernel two-sample tests. *Biometrika 111*(3), 755–770.

Székely, G. J., M. L. Rizzo, et al. (2004). Testing for equal distributions in high dimension. *InterStat 5*(16.10), 1249–1272.

Zhou, D. and H. Chen (2023, 12–15 Jul). A new ranking scheme for modern data and its application to two-sample hypothesis testing. In G. Neu and L. Rosasco (Eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, Volume 195 of *Proceedings of Machine Learning Research*, pp. 3615–3668. PMLR.

Zhu, C. and X. Shao (2021). Interpoint distance based two sample tests in high dimension. *Bernoulli 27*(2), 1189 – 1211.

Zhu, Y. and H. Chen (2023). Robust graph-based methods for overcoming the curse of dimensionality. *arXiv*, 2307.15205.

Zhu, Y. and H. Chen (2024). Limiting distributions of graph-based test statistics on sparse and dense graphs. *Bernoulli 30*(1), 770 – 796.

Department of Statistics, Iowa State University

E-mail: ycbai@iastate.edu

Department of Statistics, Iowa State University

E-mail: lchu@iastate.edu