A TRIDENT SCHOLAR PROJECT REPORT

NO. 475

Innovations to Increase the Power of State-of-the-Art Graph-Theoretic Two-Sample Statistical Tests

by

Midshipman 1/C Michael J. Wallace, USN

UNITED STATES NAVAL ACADEMY
ANNAPOLIS, MARYLAND

This document has been approved for public release and sale; its distribution is unlimited.

USNA-1531-2

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
05-21-18

2. REPORT TYPE

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Innovations to Increase the Power of State-of-the-Art Graph-Theoretic Two-Sample Statistical Tests

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Wallace, Michael J.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Naval Academy
Annapolis, MD 21402

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Trident Scholar Report no. 475 (2018)

12. DISTRIBUTION / AVAILABILITY STATEMENT

This document has been approved for public release; its distribution is UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
One of the classic problems in statistics is to determine whether a group of observations can be characterized as statistically different from some other group. In the case of the well-known two-sample *t*-test, observations are univariate (1-dimensional) and underlying probability distributions are normal (or approximately normal). However, in real-world problems, the number of covariates may be very large and there may be little known about underlying distributions. Finding powerful tests for group differences in this general multivariate case presents challenges, and this difficult case has attracted recent research interest.
In the setting of graph-theoretic approaches, the first consequential two-sample test was introduced by Friedman and Rafsky (FR1979) as a multivariate generalization of the Wald- Wolfowitz runs test. The rationale of this test and newer, similar tests is that, if two samples are from different distributions, observations would be preferentially closer to others from the same sample than those from the other sample. This project explores the tradeoffs between graph density, test power, and computational costs in a variety of scenarios and recommends guidelines for edge-counting criteria. The benefits and drawbacks of using denser subgraphs are analyzed to extend recent findings in statistical literature. A power simulation study is used to examine state-of-the-art tests in competition under the same conditions and compare performance. A novel exploratory approach is then introduced that enables finding group differences at lower computational costs. Next, the efficacy of a newly-proposed dissimilarity measure for mixed data, "treeClust", is investigated using real-world medium-sized and large-sized data sets. Finally, we introduce a new test that involves ranking all of the edges with respect to weight instead of selecting a subset of edges based on some other more time-consuming optimality criterion, as is done in other such tests. This Cumulative Cross-Count (CCC) test is a competitively powerful, user-friendly, nonparametric, multivariate, multi-group test. We derive moment information and employ permutation approaches to approximate p-values.

15. SUBJECT TERMS
Nonparametrics, Graph-theoretic procedure, Two-sample hypothesis testing, Minimum spanning tree, Minimum non-bipartite matching, Tree-based dissimilarity

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		79	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

U.S.N.A. --- Trident Scholar project report; no. 475 (2018)

INNOVATIONS TO INCREASE THE POWER OF STATE-OF-THE-ART GRAPH-THEORETIC TWO-SAMPLE STATISTICAL TESTS

by

Midshipman 1/C Michael J. Wallace
United States Naval Academy
Annapolis, Maryland

_____
(signature)

Certification of Adviser(s) Approval

CAPT David M. Ruth, USN
Mathematics Department

_____
(signature)

_____
(date)

Acceptance for the Trident Scholar Committee

Professor Maria J. Schroeder
Associate Director of Midshipman Research

_____
(signature)

_____
(date)

USNA-1531-2

ABSTRACT

One of the classic problems in statistics is to determine whether a group of observations can be characterized as statistically different from some other group. In the case of the well-known two-sample *t*-test, observations are univariate (1-dimensional) and underlying probability distributions are normal (or approximately normal). However, in real-world problems, the number of covariates may be very large and there may be little known about underlying distributions. Finding powerful tests for group differences in this general multivariate case presents challenges, and this difficult case has attracted recent research interest.

In the setting of graph-theoretic approaches, the first consequential two-sample test was introduced by Friedman and Rafsky (FR1979) as a multivariate generalization of the Wald-Wolfowitz runs test. The rationale of this test and newer, similar tests is that, if two samples are from different distributions, observations would be preferentially closer to others from the same sample than those from the other sample.

This project explores the tradeoffs between graph density, test power, and computational costs in a variety of scenarios and recommends guidelines for edge-counting criteria. The benefits and drawbacks of using denser subgraphs are analyzed to extend recent findings in statistical literature. A power simulation study is used to examine state-of-the-art tests in competition under the same conditions and compare performance. A novel exploratory approach is then introduced that enables finding group differences at lower computational costs.

Next, the efficacy of a newly-proposed dissimilarity measure for mixed data, “treeClust”, is investigated using real-world medium-sized and large-sized data sets.

Finally, we introduce a new test that involves ranking all of the edges with respect to weight instead of selecting a subset of edges based on some other more time-consuming optimality criterion, as is done in other such tests. This Cumulative Cross-Count (CCC) test is a competitively powerful, user-friendly, nonparametric, multivariate, multi-group test. We derive moment information and employ permutation approaches to approximate p-values.

KEYWORDS: Nonparametrics; Graph-theoretic procedure; Two-sample hypothesis testing; Minimum spanning tree; Minimum non-bipartite matching; Tree-based dissimilarity

ACKNOWLEDGEMENTS

I would first and foremost like to express my sincere gratitude to CAPT David Ruth for his incredible support and invaluable guidance as an advisor, professor, and mentor over the past four years. His patience, motivation, enthusiasm, and immense knowledge helped me in all the time of research and writing this Trident project. I could not have imagined having a better advisor.

Besides my advisor, I would like to thank the members of the Trident Committee, especially Prof. Anastasios Liakos, Prof. Ahmed Rahman, and Prof. Maria Schroeder, for their encouragement, insightful comments, and constructive feedback.

Last but not the least, I would like to thank my friends, roommates, company-mates, and parents for supporting me throughout the entire process.

TABLE OF CONTENTS

I.	Introduction……..……..…….	4
II.	Problem Background.……..……..…….……..…	13
III.	Motivation for Proposed Improvements……..	21
IV.	Simulation Results……..…….	27
V.	The Cumulative Cross Count (CCC) Test……...	46
VI.	Conclusion & Opportunities for Future Work……...….…….	64
VII.	Appendices…….……..	66
VIII.	List of References…..……..……..…....	79

I. INTRODUCTION

The ability to detect subtle changes in systems that are subject to random variability is a central problem in statistics that has great practical importance. Decision-makers in a wide variety of industries often want to determine whether or not two groups of observations can be characterized as being statistically different. For example, consider these meaningful real-world applications:

- *Healthcare* – Is the condition of patients who receive a new drug better than that of those who received a placebo?

- *Industry* – When multiple factories are used to manufacture the same product, are consistency and uniformity standards maintained across different plants?

- *Business* – Is a company more successful after implementing a new sales tool?

- *Government* – Does the enforcement of a new law lead to a better functioning society?

- *Equal Opportunity* – Is the work environment experience different based on gender or race?

- *Military* – Is the overall health of a helicopter after the completion of a mission different than before?

The classic approach to these two-sample problems, often first encountered in an undergraduate-level statistics course, is referred to as the two-sample location *t*-test. Based on the well-known Student's *t*-test, which was developed in the early 1900's by William Gossett (under the pen name "Student"), the two-sample location *t*-test seeks to determine if there is a *statistically significant* difference between two group means (averages). More precisely, the two-sample location *t*-test is used to test the null hypothesis that two populations have equal means.

In order to better understand the concept of *statistical significance*, we introduce some basic statistical terminology. In statistics, a *population* is a set of similar items or events which is of interest for some question or experiment (e.g. the set of all college students in Maryland, the set of all trucks manufactured in Detroit, the set of all women in the military, etc.). In

general, the goal of statistical analysis is to produce information about some specified population. Typically, however, the population is very large, making it impractical or impossible to systematically acquire and record information about every element of a given population. Thus, a reasonably-sized subset of the population, called a *data sample*, is collected to represent the population in a statistical analysis. The elements of a sample are commonly referred to as *observations*. Statistics are then calculated from the sample to estimate certain characteristics of the larger population, such as its mean or standard deviation.

Clearly, since the sample does not include all members of the population, statistics on the sample are not perfectly precise estimates of the population. For example, the average grade-point-average of 10 randomly-selected students at the U.S. Naval Academy is typically not the same as the average grade-point-average of all 4,500 students at the school. In order words, sample statistics (such as sample mean and sample standard deviation) depend on the specific observations in the sample and will vary from sample to sample. The difference between the sample and population values is called *sampling error*. With these basic ideas in mind, the discussion now returns to the concept of *statistical significance*.

When testing for a difference between two groups of data, we proceed in the following manner:

1.) Assume that there is no underlying difference between the two groups of data (i.e. both sets of data were drawn from the same population, or distribution).
2.) Based on the collected data, calculate the probability of obtaining a result at least as extreme as that which is present (given that there is no underlying difference between the two groups). This calculated probability is called a *p-value*.
3.) Compare this *p-value* to a predetermined *significance level* (0.05 is commonly used). If the *p-value* is less than the *significance level*, conclude that a *statistically significant* difference exists between the two groups of data. Otherwise, no conclusion may be drawn.

This process for determining statistical significance is known as *statistical hypothesis testing*. First of all, note that we never conclude that two groups of data are statistically “the same”. Secondly, note that the determination of statistical significance depends on the predetermined *significance level*. A lower significance level requires a lower p-value (i.e. a more “extreme” result from the collected data) in order to conclude that a difference exists between

two groups. In a sense, the use of statistical hypothesis testing to determine whether a group difference exists is similar to the "presumption of innocence" in the U.S. court of law. Compare the following description of the U.S. court of law to the three numbered points listed above and observe the similarities:

1.) Upon entering the courtroom, the defendant is assumed to be not guilty (i.e. the "presumption of innocence").
2.) The judge and/or jury then decide how much the presented evidence disinclines them to continue believing in the innocence of the defendant.
3.) This notion of "how much the judge and/or jury disbelieve in the innocence of the defendant" is then compared to a legal standard for burden of proof (e.g. some evidence, reasonable suspicion, preponderance of the evidence, clear and convincing evidence, beyond reasonable doubt, etc.). Based on the amount of evidence collected and the legal standard being used, the defendant is then found to be "guilty" or "not guilty" of the crime for which he was charged. Note that a defendant is never found by a judge or jury to be "innocent".

In the classic two-sample location *t*-test the sample data is used to estimate the mean and standard deviation of the two groups. The means and standard deviations are used to calculate the *t*-statistic, which is then compared to a percentile from the known *t*-distribution. The equation for calculating the *t*-statistic is shown below. Note that the calculation is simply the difference between the two sample averages ($\bar{x}_1$ and $\bar{x}_2$), scaled by estimates of variability (s_1 and s_2) and the respective sample sizes of the two groups (n_1 and n_2). Assumptions about normality allow us to know the exact distribution of this test statistic, t.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The two-sample location *t*-test is an example of a *parametric test*, a statistical test that makes assumptions about the underlying probability distribution. In the *t*-test, we estimate population parameters (mean and standard deviation) and we assume that the underlying populations are normally distributed (or, at least, approximately normal). Figures 1 and 2 provide a qualitative description of a parametric two-sample test. In Figure 1, there are two data samples: the blue circles are univariate (1-dimensional) observations that belong to Group 1 and

the red triangles are univariate observations that belong to Group 2. In a two-sample test, we want to determine if these two groups of observations are statistically different.

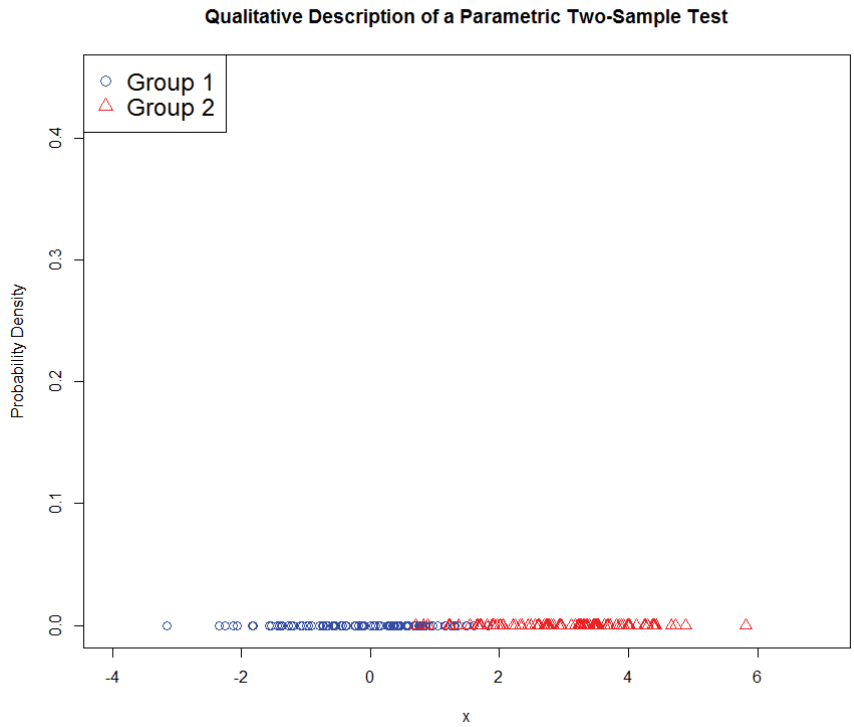


Figure 1: Two groups of univariate observations (blue circles and red triangles) plotted on a graph

In Figure 2, the sample means and variances of each group are used to estimate the underlying populations as normal distributions (“bell curves”). From a qualitative perspective, we then decide whether or not these two groups seem to come from different underlying population distributions, depending on the degree of overlap between the two bell curves.

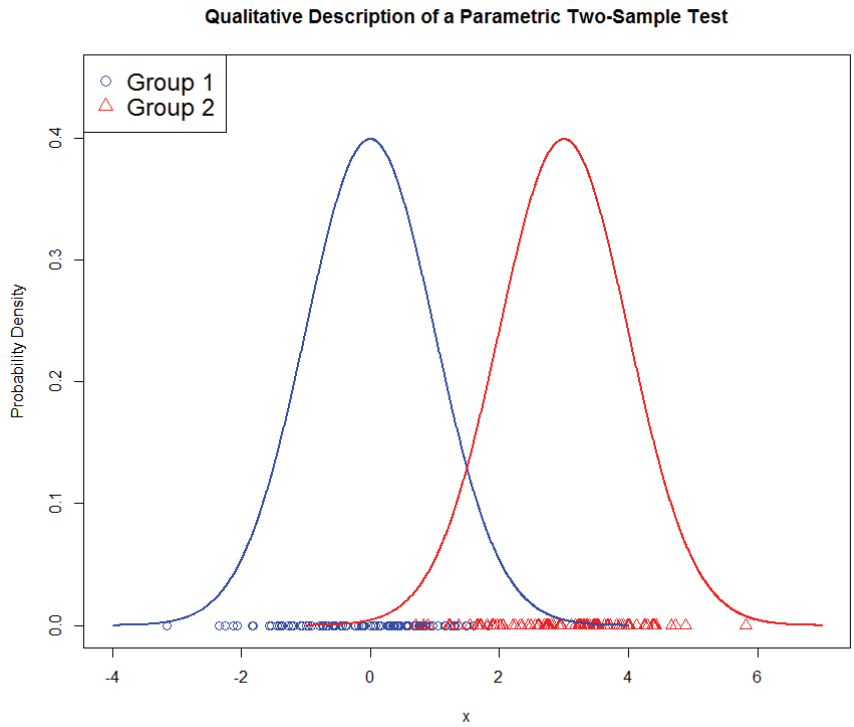


Figure 2: Two bell curves plotted as estimates of the underlying probability distributions

Unfortunately, these sort of parametric approaches to two-sample testing are subject to some notable limitations, namely *multi-variation* and *model assumptions*. Multi-variation refers to the idea that real-world processes are often best described by more than one attribute. For instance, overall human health might be appropriately modeled by blood pressure, whereby a very high or very low blood pressure reading may be an indicator of poor health. However, it is likely that a collection of different attributes, such as blood pressure, age, weight, smoking habits, exercise habits, family history, and respiratory rate, would provide a much better model of overall human health. The problem with parametric tests is that the ability to estimate parameters well is degraded in higher dimensions. In other words, as the number of attributes increases, a much greater amount of data is needed to create a useful model. In fact, if the number of attributes exceeds the number of collected observations, which is a realistic possibility in many practical applications, it is not even mathematically possible to use some parametric models for two-sample testing.

The second limitation of parametric testing, *model assumptions*, refers to the concept that the power of a parametric test depends on how well model assumptions are satisfied. The problem is that distributional assumptions are often difficult to justify in higher dimensions. In

the one-dimensional case, graphical tools such as a quantile-quantile plot may be used to assess if a set of data was plausibly drawn from some theoretical distribution, such as a normal (Gaussian) distribution. However, there are no such graphical tools for reliability determining if a multi-dimensional data set (e.g. 100-dimension data) was drawn from a known multivariate distribution (e.g. 100-variate normal distribution). One approach to overcome these major limitations is to use nonparametric testing.

A nonparametric test is a statistical test that does not rely on assumptions that the data are drawn from a given probability distribution. Compared to parametric tests, these nonparametric tests are much more robust and widely applicable. They can readily accommodate a large number of covariates, and they may be used when very little is known about the underlying probability distributions. In many cases, such as the graph-theoretic approaches discussed in this paper, two-sample nonparametric tests are intuitively straightforward. However, *test power* for multivariate cases invites room for improvement. Put simply, test power is the ability to detect a difference between two groups when a difference actually exists.

In this research, we are interested in exploring graph-based nonparametric approaches to two-sample statistical testing. Figure 3 displays a very simple example of a two-dimensional case where 20 observations are divided into two equal-sized groups labelled “1” and “2”, and the quantitative covariates x and y are plotted on the horizontal and vertical axes. The problem is to determine if Group 1 is statistically different from Group 2.

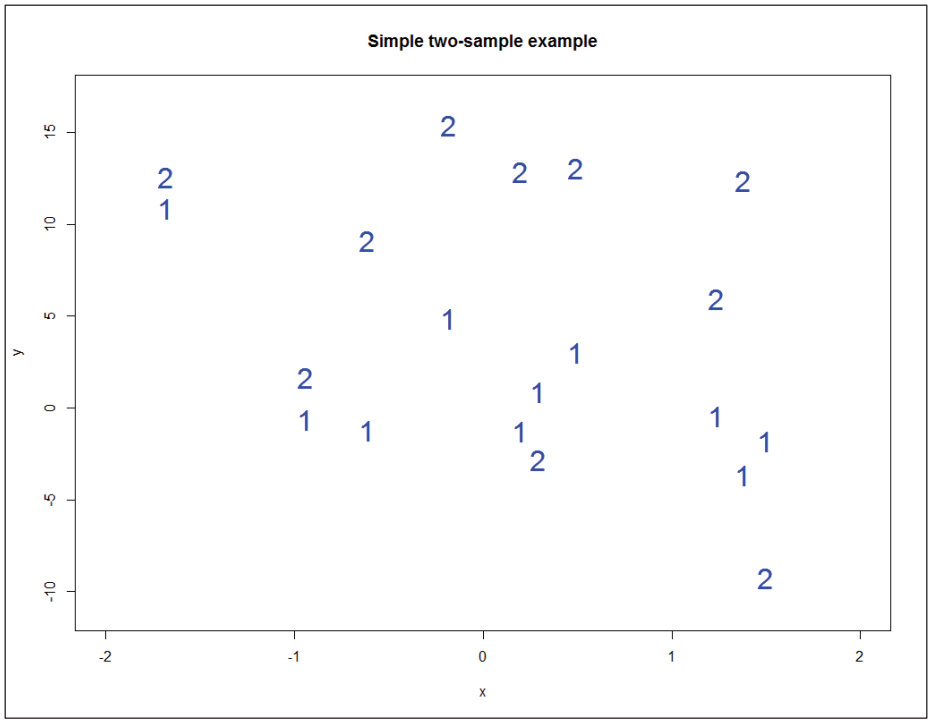


Figure 3: Example of bivariate quantitative data where the problem is to determine if Group 1 is statistically different from Group 2.

Most approaches to solve this problem involve some notion of specifying interpoint dissimilarities (i.e. how far is the upper-left-most observation from the lower-right-most observation?) In this particular case, any metric on $\mathbb{R}^2$ might be a reasonable dissimilarity measure. Many approaches also involve assumptions about the distributions which generate the covariates.

In contrast, Figure 4 shows a less simple (but still fairly simple) example of a four-dimensional case where groups are labelled “1” and “2”, and the four covariates are *x* (horizontal axis; quantitative), *y* (vertical axis; quantitative), color (blue or red; categorical), and shape (triangle, square, or circle; categorical). The problem is to determine if Group 1 is statistically different from Group 2. Again, most approaches to solve this problem involve some notion of specifying interpoint dissimilarities, but this case is complicated by the fact that it is not as clear how to incorporate color and shape into the dissimilarity measure.

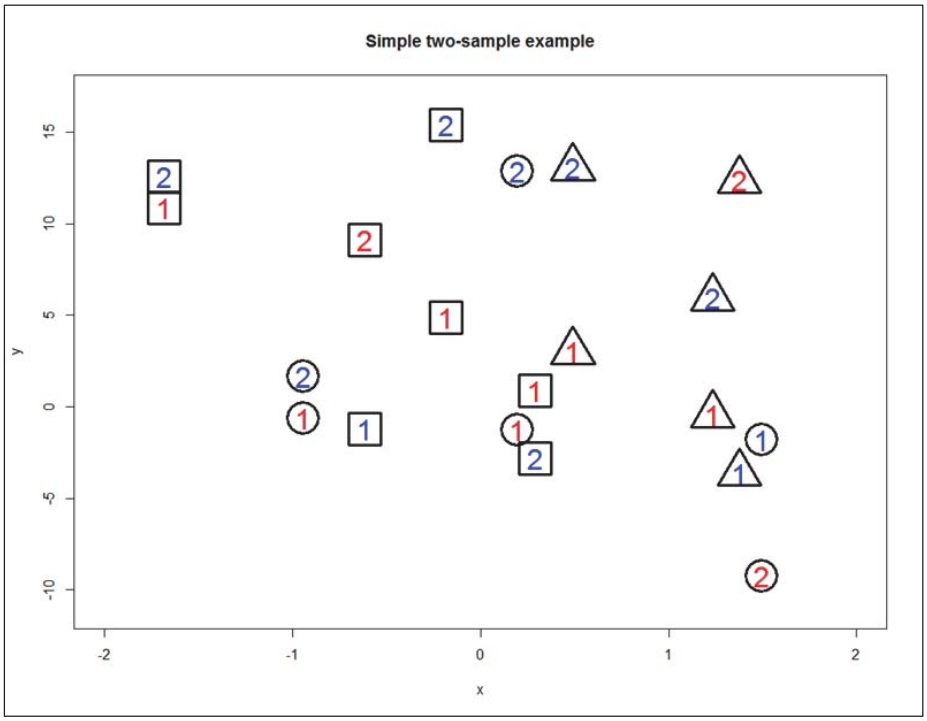


Figure 4: Example of four-variate mixed data where the problem is to determine if Group 1 is statistically different from Group 2.

Our research seeks to investigate and improve upon some of the latest solutions to problems like this one, but in the more challenging cases where

- the number of observations, *N*, may be very large,
- the number of covariates, *dim*, may be very large,
- little is known about underlying probability distributions of two groups *F* and *G*,
- there may be no natural dissimilarity measure available.

Specifically, we aim to validate, refine, and extend results on graph-based two-sample tests which are described in the following section, the first by Friedman and Rafsky (FR1979), the second by Rosenbaum (Ro2005), the third by Ruth (Ru2014) and the fourth by Chen and Friedman (CF2017). All of the approaches convert a set of observations into a particular weighted, undirected graph, and then count graph edges in a particular way to identify whether a group difference exists. One of the main goals of this research is to find the most efficient way to build graphs on a set of observations that contain informative edges which enable detecting a difference between two groups of data.

Our work towards these ends is organized as follows: In Section II, we formally introduce the two-sample problem in the context of a graph-theoretic setting. We then explore

the theoretical underpinnings of this problem through a review of recent approaches in relevant statistical literature. This section culminates with a brief discussion of a new dissimilarity measure based on tree-clustering. In Section III, we discuss the motivation behind our proposed improvements to current methods, namely the rationale to increase the density of optimal subgraphs used in these graph-based tests. We then propose the use of an alternative optimal subgraph based on shortest edges. In Section IV, we validate and refine the results of test power simulation studies from a key journal article published in 2017 in the *Journal of the American Statistical Association*. We then use those refined power estimates to investigate the effects of increasing subgraph density on test power and computational costs as well as to compare competing state-of-the-art methods under the same test conditions. We conclude this section by applying a newly-proposed dissimilarity measure for mixed data, “treeClust”, to a real-world medium-sized data set consisting of test results for patients, with or without a heart condition, undergoing angiography at the Cleveland Clinic in Ohio. In Section V, we introduce a new statistical test for the first time: the Cumulative Cross-Count (CCC) test. We show that the CCC test is a competitively powerful, user-friendly, nonparametric, multivariate test. We then derive moment information and employ permutation approaches to approximate p-values. Lastly, we compare the CCC test to the best existing parametric and nonparametric approaches and explore its practical efficacy using a real-world large-sized data set consisting of features extracted from electric current drive signals from a drive which has intact and defective components. In Section VI, we summarize all of our findings and highlight opportunities for future work within the field and on the CCC test in particular.

II. PROBLEM BACKGROUND

A. PROBLEM FORMULATION

Consider N independent observations in two distinct groups of sizes m and n, respectively. So, $N = m + n$. Each observation is drawn from the same probability distribution as all other observations in that same group. Formally, we call the groups $\mathcal{O}_1 = \{Y_1, \ldots, Y_m\}$ and $\mathcal{O}_2 = \{Y_{m+1}, \ldots, Y_N\}$, where each Y_i is drawn from distribution F for $1 \leq i \leq m$ and from distribution G for $m + 1 \leq i \leq N$. The $Y_i's$ may have any number of attributes; we will use dim to denote the number of attributes. For example, if the $Y_i's$ measure the height, weight, and hair color of two groups of Midshipmen where 20 are male and 10 are female, then $m = 20$, $n = 10$, and $dim = 3$. dim does not depend on N. The covariates may be quantitative or categorical. Assume there exists some function d that measures dissimilarity between observations; that is $d(Y_i, Y_j) = 0$ if $i = j$, $d(Y_i, Y_j)$ is small if Y_i and Y_j are "close", and $d(Y_i, Y_j)$ is large if Y_i and Y_j are "distant." We will develop a test statistic T whose distribution can be derived (or approximated) for the case $F = G$, and is unusually small or large when $F \neq G$. Furthermore, we would like T to have no dependence on F or G.

We will restrict our work to graph-theoretic approaches, which generally adhere to the following setting:

- Each observation is considered to be a vertex in a graph $\mathcal{G}$.
- Each pair of observations is an (undirected) edge of $\mathcal{G}$.
- Edge weights are assigned based on interpoint dissimilarities (as determined by d).

In our setting, $\mathcal{G}$ is a complete weighted graph, which means that every vertex is connected to every other, and every connecting edge has a weight (i.e., dissimilarity value) assigned to it. A variety of test statistics T can be found by considering $\mathcal{G}^*$, a subgraph of $\mathcal{G}$ that is optimal by some measure. In the ensuing discussion, we will use the following notation throughout:

R_0 = number of edges in $\mathcal{G}^*$ that connect vertices in $\mathcal{O}_1$ to vertices in $\mathcal{O}_2$;

R_1 = number of edges in $\mathcal{G}^*$ that connect vertices in $\mathcal{O}_1$ to vertices in $\mathcal{O}_1$;

R_2 = number of edges in $\mathcal{G}^*$ that connect vertices in $\mathcal{O}_2$ to vertices in $\mathcal{O}_2$.

In other words, R_0 is the number of across-group edges, R_1 is the number of within-group edges for the first group, and R_2 is the number of within-group edges for the second group. A

foundational concept in this setting is that *under the null hypothesis of no group difference (i.e. our default assumption that the distributions F and G are the same) each vertex is equally likely to be paired in $\mathcal{G}^*$ with any other vertex.*

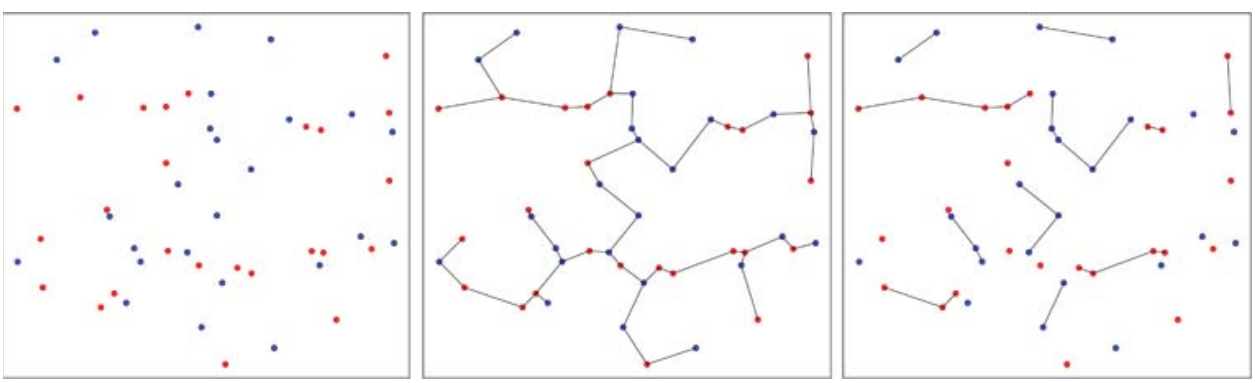
B. RECENT GRAPH-THEORETIC APPROACHES TO TWO-SAMPLE STATISTICAL TESTING

1. Friedman and Rafsky (FR1979)

The first consequential two-sample test in this setting was introduced by Friedman and Rafsky (FR1979). FR1979 proposes the use of a minimum spanning tree (MST), which is a subset of edges of a connected, edge-weighted undirected graph (i.e. a subgraph) that connects all vertices together ("spanning"), without any cycles ("tree") and with the minimum possible total edge weight ("minimum"). Refer to Appendix 1 for a review of basic graph theory definitions and a more detailed discussion of minimum spanning trees. FR1979 lets $\mathcal{G}^*$ be a minimum spanning tree (MST) of $\mathcal{G}$, and uses the test statistic

$$T_{\mathrm{FR}} = R_0 + 1.$$

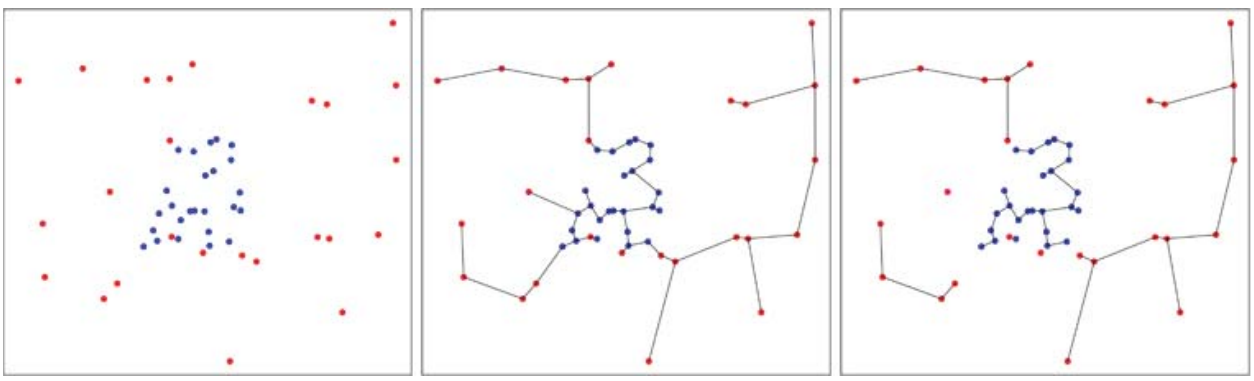
The quantity T_{FR} can be interpreted as the number of within-group clusters of $\mathcal{G}^*$. Figure 5 provides a visual representation of how an MST is constructed on a set of observations and then decomposed into within-group clusters, or subtrees. Red dots and blue dots may be thought of as observations from $\mathcal{O}_1$ (Group 1) and $\mathcal{O}_2$ (Group 2), respectively.



Plot of Observations **Construction of MST** **Decomposition into Subtrees**

Figure 5: Visual representation of how an MST is constructed on a set of observations and then decomposed into subtrees. In this case, there are 29 subtrees ($T_{\mathrm{FR}} = 29$) and the expected number of subtrees is 26 ($E[T_{\mathrm{FR}}] = 26$, based on $m = n = 25$). This is not strong evidence of a group difference.

The null hypothesis is rejected for small values of T_{FR}, which makes this test essentially a multivariate extension of the Wald–Wolfowitz runs test. In other words, if the two groups are actually different, we would expect observations from $\mathcal{O}_1$ to be preferentially closer to other observations from $\mathcal{O}_1$ than to those from $\mathcal{O}_2$. Figure 6 provides another visual representation of how an MST is constructed on a set of observations and then decomposed into subtrees; however, in this case, there is much stronger evidence of a group difference. As before, the red dots and blue dots may be thought of as observations from $\mathcal{O}_1$ (Group 1) and $\mathcal{O}_2$ (Group 2), respectively.



Plot of Observations **Construction of MST** **Decomposition into Subtrees**

Figure 6: Visual representation of how an MST is constructed on a set of observations and then decomposed into subtrees. In this case, there are 8 subtrees ($T_{\mathrm{FR}} = 8$) and the expected number of subtrees is 26 ($E[T_{\mathrm{FR}}] = 26$, based on $m = n = 25$). This is strong evidence of a group difference.

The null distribution of T_{FR} does not depend on the distribution of F or G, but is conditional on the structure of $\mathcal{G}^*$. Friedman and Rafsky showed that, under the null hypothesis, the test statistic T_{FR} has expected value

$$E[T_{\mathrm{FR}}] = \frac{2mn}{N} + 1$$

and variance

$$\mathrm{Var}[T_{\mathrm{FR}}|C] = \frac{2mn}{N(N-1)}\left\{\frac{2mn - N}{N} + \frac{C - N + 2}{(N-2)(N-3)}[N(N-1) - 4mn + 2]\right\},$$

where C is the number of adjacent edge pairs in the MST, which depends on the topology of the minimum spanning tree and is determined by the node degrees. Refer to Appendix 2 for a derivation of the expected value and variance of T_{FR}. Like the Wald–Wolfowitz runs test,

however, the power of T_{FR} can be somewhat weak, which means that this test is not very good at detecting group differences when they actually exist. To improve test power, FR1979 considers a more complicated $\mathcal{G}^*$ that consists of the union of two or three disjoint MSTs. The same test statistic on this more complicated graph improves test power. This idea of using denser graphs will be an important part of our work.

2. Rosenbaum (Ro2005)

Rosenbaum (Ro2005) takes a similar approach to this problem, but uses as $\mathcal{G}^*$ a minimum-weight non-bipartite matching on $\mathcal{G}$, which is the lowest-weight spanning subgraph of $\mathcal{G}$ for which the degree of each vertex in $\mathcal{G}^*$ is exactly 1. In other words, Rosembaum's minimum-weight non-bipartite matching is the cheapest way ("minimum-weight") to connect every observation to *exactly* one other observation ("matching"), without regard as to whether an observation is from $\mathcal{O}_1$ or $\mathcal{O}_2$ ("non-bipartite"). Strictly speaking, N must be even for $\mathcal{G}^*$ to be a perfect matching, but odd N may be easily accommodated. The test statistic is

$$T_{\mathrm{Ro}} = R_0,$$

Figure 7 shows an example of Rosenbaum's minimum-weight non-bipartite matching on a set of 20 bivariate observations in two groups (red circles and blue triangles). Ro2005 proves that the exact null distribution of T_{Ro} can be expressed in closed form as a relatively simple combinatorial expression that depends only on R_0, m, and n, but not on F or G. Specifically, Rosenbaum shows

$$\Pr(T_{\mathrm{Ro}} = k) = \frac{(2^k)\left(\frac{N}{2}\right)!}{\binom{N}{n} k! \left(\frac{n-k}{2}\right)! \left(\frac{N-n+k}{2}\right)!}.$$

The null hypothesis is rejected for small values of T_{Ro}. With the same logic as before, if the two groups are actually different, we would expect observations from $\mathcal{O}_1$ to be preferentially closer to other observations from $\mathcal{O}_1$ than to those from $\mathcal{O}_2$. The fact that Ro2005's is an exact, distribution-free test is attractive but, like the FR1979 test, Ro2005 test power is somewhat weak.

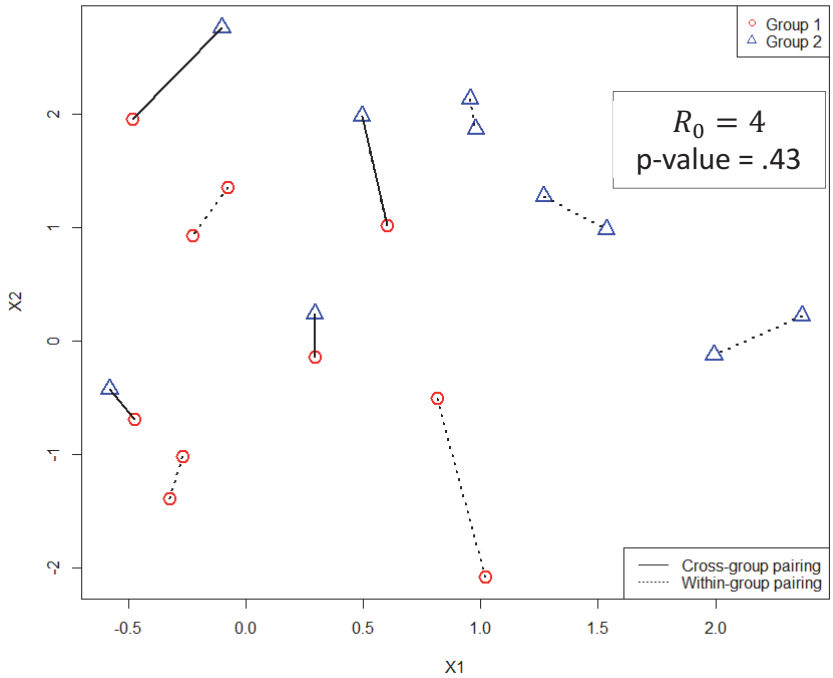


Figure 7: Example of a minimum-weight non-bipartite matching on a set of 20 bivariate observations in two groups (red circles and blue triangles), each containing 10 observations.

3. Ruth (Ru2014)

Ru2014 extends Ro2005's work by using a similar but denser (i.e., more edges) $\mathcal{G}^*$ than Ro2005: For even N, pick an integer $r \in \{1, \ldots, N/2\}$ and let $\mathcal{G}^*$ be a minimum-weight r-regular spanning subgraph, which is the lowest-weight spanning subgraph of $\mathcal{G}$ for which the degree of each vertex in $\mathcal{G}^*$ is r. (Odd N may be accommodated as in Ro2005.) Note that for $r = 1$ this is the same as the Ro2005 case. The associated test statistic is defined as

$$T_{\mathrm{Ru}} = \frac{R_0}{r}.$$

Figure 8 shows an example of a minimum-weight 3-regular spanning subgraph on 20 bivariate observations in two groups (red circles and blue triangles). The plot shows the graph $\mathcal{G}^*$ with respect to Euclidean distance (i.e., distance as measured by a ruler) for $r = 3$; solid edges are cross-group edges and dashed edges are within-group edges. For this particular example, $R_0 = 12 \Rightarrow T_{\mathrm{Ru}} = \frac{12}{3} = 4$. Like T_{FR} and T_{Ro}, the null hypothesis is rejected for small values of T_{Ru}.

Unlike T_{Ro}, the null distribution of T_{Ru} may not necessarily be expressed exactly for $r > 1$. However, Ru2014 shows that, under the null hypothesis, the test statistic T_{Ru} has expected value

$$E[T_{\mathrm{Ru}}] = \frac{mn}{N-1}$$

and variance

$$Var[T_{\mathrm{Ru}}] = \frac{2m(m-1)n(n-1)(N-1-r)}{r(N-3)(N-2)(N-1)^2}.$$

Refer to Appendix 3 for a derivation of the expected value and variance of T_{Ru}. Approximate p-values for T_{Ru} may be computed easily for fairly large N using a permutation test on the observation group labels (for example, the p-value for the case in Figure 8 is approximately 0.11). For larger N, T_{Ru} can be approximated using the normal distribution. The main advantage of this test is that it has been shown to have impressive power over a broad range of alternatives (Ruth, D. 2014).

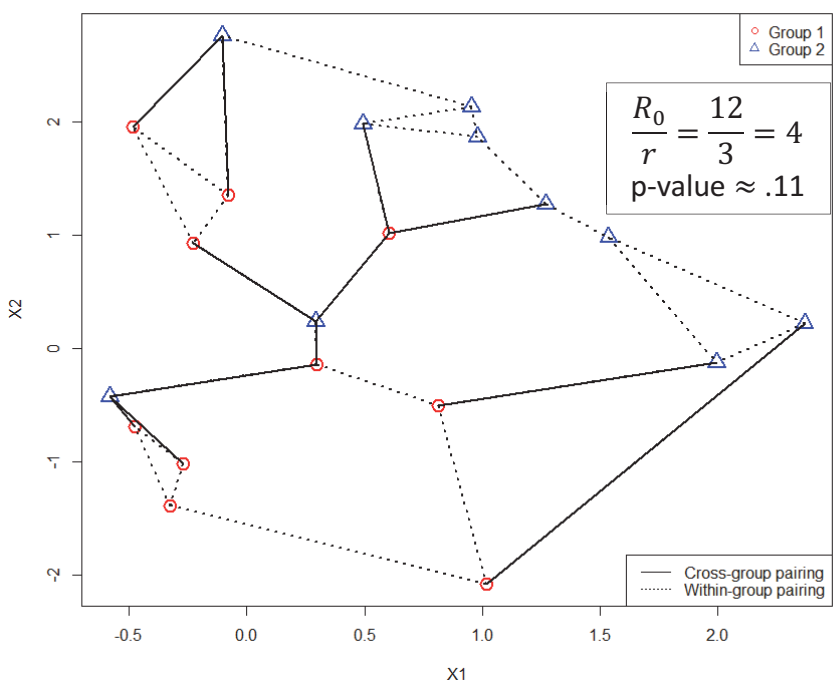


Figure 8: An example of a minimum-weight 3-regular spanning subgraph (3-MWSS) on 20 bivariate observations in two groups (red circles and blue triangles), each containing 10 observations.

4. Chen and Friedman (CF2017)

The three tests described so far tend to perform best against location alternatives (i.e., alternatives for which F and G differ in location; say, by a shift in mean). They share a common pitfall; namely that their power can be very low against scale alternatives (i.e., alternatives for which F and G differ in scale; say, by a shift in variance). Very recently, Chen and Friedman published a new graph-theoretic test (CF2017) in the *Journal of the American Statistical Association* that exhibits high power for location or scale alternatives (and both). They consider a variety of $\mathcal{G}^*$ possibilities, including unions of a limited number of disjoint spanning trees or unions of a limited number of disjoint non-bipartite matchings, and they use the test statistic

$$T_{\mathrm{CF}} = (X_1 - \mu_1, X_2 - \mu_2)\Sigma_{12}^{-1}\begin{pmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{pmatrix},$$

where $\mu_1 = E[X_1]$, $\mu_2 = E[X_2]$, and Σ_{12} is the covariance matrix for the vector $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. This statistic is simply a centered and scaled measure of the joint deviation of X_1 and X_2 from their expected values. In this case, large values of T_{CF} lead to a rejection of the null hypothesis. Again, approximate p-values for T_{CF} may be computed easily for fairly large N using a permutation test. For larger N, T_{CF} can be approximated using the chi-squared distribution. Interestingly, CF2017 speculates that $\mathcal{G}^*$ ought not to be overly dense. This is contrary to results in Ru2014 that indicate a dense $\mathcal{G}^*$ leads to better test power.

C. A NEW DISSIMILARITY MEASURE BASED ON TREE-CLUSTERING

Finally, it is the case that every one of the tests described above relies on a given dissimilarity measure, d. But for mixed data with many covariates, it is not necessarily clear what dissimilarity measure ought to be used. In effect, mixed data are datasets that include both quantitative and qualitative variables. While finding the distance between two numbers is generally straight-forward, finding the distance between two categories presents a much greater challenge. For example, for some individual attributes such as race, sex, hair color, and educational level, assigning an appropriate dissimilarity measure may be difficult. Without a good dissimilarity measure, mixed data may not contribute much to detecting whether or not a group difference exists in a two-sample test. A study of dissimilarity measures could constitute its own separate research

project; however, in this work we will explore the efficacy of a dissimilarity measure based on tree-clustering recently proposed by Buttrey and Whitaker (BW2015) in the context of graph-theoretic two-sample tests. The BW2015 approach is as follows: For each covariate $k \in \{1, \dots, p\}$, construct a classification or regression tree modeling covariate k as the response variable and including all other covariates as predictor variables in the tree. After applying some pruning and discarding rules, a collection of $k_0 \leq p$ trees remains, where each observation is assigned to one leaf in each tree. The key idea is that *observations are considered dissimilar with respect to a particular tree when they fall in different leaves of that tree*. Among many options for an associated dissimilarity measure d is to define

$$d(Y_i, Y_j) = \frac{1}{k_0} \sum_{k=1}^{k_0} I_k(i, j)\,;$$

where $I_k(i, j)$ is the indicator function that observations i and j fall in different leaves of tree k. That is, $d(Y_i, Y_j)$ is the proportion of trees in which Y_i and Y_j fall in different leaves. This proposed measure has the great advantage that it can be used for quantitative, categorical, or mixed data. BW2015 suggests that it is also very resilient to noise.

III. MOTIVATION FOR PROPOSED IMPROVEMENTS

We propose that increasing the density of the subgraphs used in these graph-based two-sample tests leads to impressive improvements in statistical test power. Although past literature in this area of research has acknowledged that denser subgraphs lead to power improvements, we believe that they might have severely underestimated the value of graph density relative to its cost. In the context of these problems, graph density refers to the fraction of edges of the complete graph that are considered in a statistical test. Our rationale is that considering a greater number of edges will provide more "closeness information". Of course, we expect that increasing graph density will come at the expense of more difficult optimization and more complicated test statistic null distributions, but we believe that these tradeoffs may be worthwhile in some cases. We will now discuss some of the recent work that has motivated this exploration into the tradeoffs between graph density, test power, and computational costs.

A. INCREASING SUBGRAPH DENSITY IN EXISTING LITERATURE

In his extension of Rosenbaum's Cross-Match Test, Ruth (Ru2014) explores the value of using optimal *r*-regular graphs with $r > 1$. In that paper, Ruth's power simulations suggest that using greater graph densities improves the ability of the test to detect location changes on multivariate data. Figure 9 is a reproduction of a power graph featured in Ru2014. The graph provides power estimates for the mean cross-count test on 5-variate normal data with $m = 20$ and $n = 40$. In addition to validating the results of Ru2014, this reproduction shows how increasing the graph density ($r = 1$ to $r = 4$ to $r = 10$ to $r = 30$) steadily increases the corresponding power estimates. These power estimates are compared to the exact power of the Hotelling T^2 test (the multivariate analog of the univariate *t*-test), which is the parametric test of choice for multivariate two-sample testing against location alternatives under assumptions of normality.

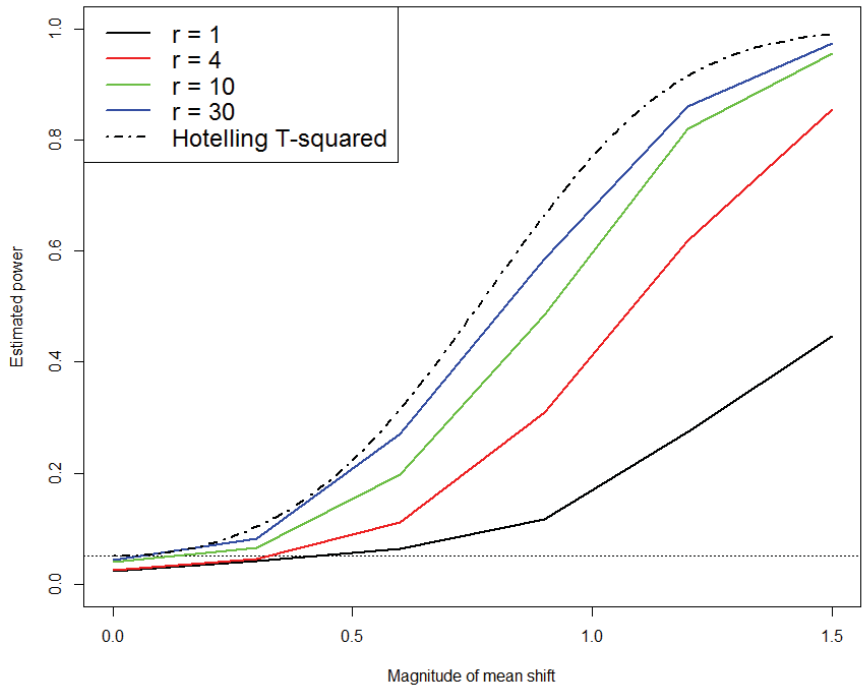


Figure 9: Power estimates for the mean cross-count (MCC) test at r = 1, 4, 10, and 30 and exact power for Hotelling's T^2 statistic for 5-variate normal mean alternatives with m = 20 and n = 40. These results validate results featured in Ru2014.

Figure 10 is a reproduction of another power graph featured in Ru2014. This graph provides power estimates for the mean cross-count test on 5-variate lognormal data with $m = 20$ and $n = 40$. In addition to validating the results of Ru2014, this reproduction also shows how increasing the graph density ($r = 1$ to $r = 4$ to $r = 10$ to $r = 30$) steadily increases test power. In addition, Figure 10 reveals the relative power of nonparametric tests (in this case, the mean cross-count test) over the Hotelling T^2 test when its assumptions are violated. When the underlying probability distributions are non-normal, as is generally encountered in real-world data, and the dimension starts to increase, nonparametric tests remain impressively robust. Collectively, these two figures seem to suggest that the use of greater graph densities ought to be explored in other areas of graph-based two-sample testing. We also note that Ru2014 does not provide computational time estimates for varying graph densities; however, these are explored and discussed in the next section ("Simulation Results").

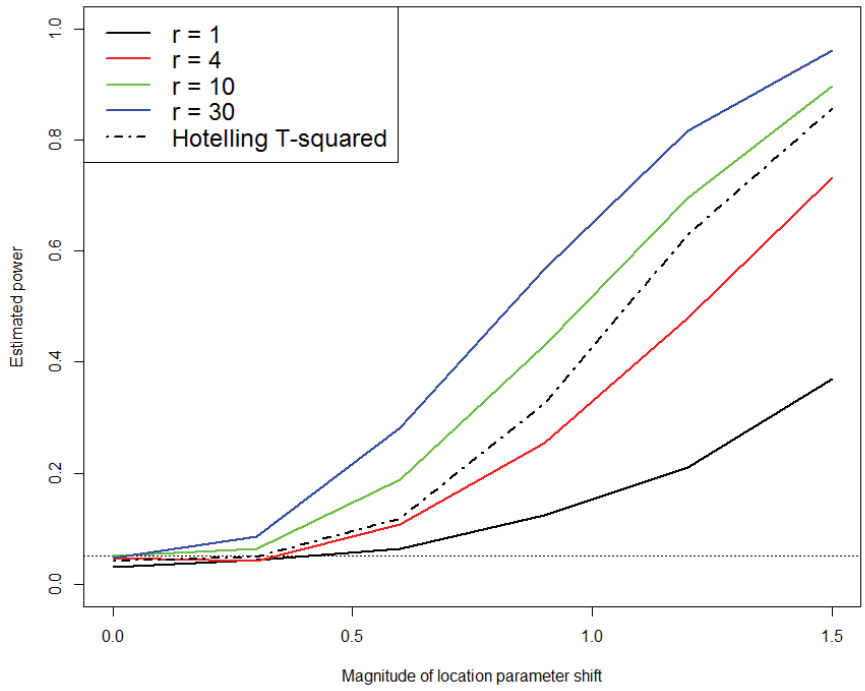


Figure 10: Power estimates for the mean cross-count (MCC) test at r = 1, 4, 10, and 30 and exact power for Hotelling's T^2 statistic for 5-variate lognormal location parameter alternatives with m = 20 and n = 40. These results validate results featured in Ru2014.

As mentioned previously, CF2017 exhibits high power for both location or scale alternatives (and both); this test is considered to be among the best of its kind. In their approach, Chen and Friedman suggest building successive MSTs and then counting the deviation of each within-group edge count from the expected edge count and scaling appropriately. The simulations in CF2017 show that increasing the graph density from a 1-MST to a 3-MST (3 successive MSTs) to a 5-MST (5 successive MSTs) leads to steady power improvements in all scenarios studied. We will save a more detailed discussion of these scenarios and results for the next section ("Simulation Results").

CF2017 acknowledges that, for the simulation settings studied, the 5-MST did not achieve the optimal point since the trend of increasing power from a 1-MST to a 5-MST had not been stabilized. However, rather than identify an "optimal" density, they note that computational costs increase with graph density and argue that the 5-MST is "good enough", even for sample sizes in the hundreds or thousands. Furthermore, they suggest that making the similarity graph too dense might even provide "counter information", which would reduce the power of the test.

However, based on the simulation studies in Ru2014 and some intuition, we believe that, in some cases, there may be some significant value in building subgraphs more dense than a 5-MST. Having already discussed the findings from Ru2014, we will now build some intuition on graph density.

The rationale of all of these graph-based tests is that, if two samples are from different distributions, observations would be preferentially closer to those from the same sample than those from the other sample. Thus, edges in these optimal subgraphs (MSTs, minimum-weight spanning subgraphs, etc.) would be more likely to connect observations from the same sample. These tests reject the null hypothesis if the number of between-sample edges is significantly *less* than what is expected (or, alternatively, if the number of within-sample edges is significantly *more* than what is expected). Therefore, the goal in building these graphs is to find an "optimal" graph density that captures as much of the "closeness information" as possible (i.e. the differences caused by the two samples being from different underlying distributions) without capturing too much random variability (which decreases test power).

B. COMBINATORIAL JUSTIFICATION FOR INCREASING SUBGRAPH DENSITY

Consider a two-sample test in which the number of observations in Group 1 is 500 ($m = 500$) and the number of observations in Group 2 is also 500 ($n = 500$), giving a total sample size of 1000 ($N = 1000$). This might be viewed a medium-sized data set. The number of edges in the complete graph of N observations is $\binom{N}{2}$; thus, the number of edges in the complete graph of 1000 observations is $\binom{1000}{2} = 499{,}500$ edges. Recall that CF2017 recommends using a 5-MST. Combinatorially, the number of edges in a 1-MST is $N - 1$, which means that the number of edges in a 5-MST is 5 x ($N - 1$). Thus, the number of edges in a 5-MST on 1000 observations is 5 x $(1000 - 1) = 4995$ edges.

By following the recommendation of CF2017 to build a 5-MST on a data set of 1000 observations, we are only capturing $\frac{4{,}995}{499{,}500} = 0.01 = 1\%$ of the total number of available edges in the complete graph. Worse yet, this fraction continues to decrease as the total sample size increases. Intuitively, we contend that limiting ourselves to a 5-MST causes us to "miss out" on a great deal of available "closeness information", especially as sample sizes increase.

Furthermore, we raise the conjecture that the "optimal" density for graph-based two-sample testing (to maximize test power, without considering computational costs) is roughly 50% of the total number of available edges in the complete graph. Theoretically, in graph-based statistical testing, any subgraph carries the same statistical information as its complement. In other words, using *just* the edges in the 1-MST on a set of observations is equivalent to using *all* of the edges *except* those in the 1-MST. This explains why any test on *no edges* has no power at all and any test on *the complete graph* also has no power at all (the complement of no graph is the complete graph). Similarly, for example, a subgraph containing 20% of the total number of available edges has the same information as a subgraph containing 80% of the total number of available edges. Following this line of thinking, our intuition leads us to believe that the "optimal" number of edges to use is roughly 50%, since the complement of a subgraph containing roughly 50% of the available edges is also roughly 50%. We contend that using optimal subgraphs to capture roughly 50% of the total number of available edges would, in a sense, maximize the amount of "closeness information" collected without gathering "redundant edges". For reference, whereas CF2017 would recommend building a 5-MST graph on 1000 observations, we might recommend building a $\frac{1}{2} \times \frac{499{,}500}{999}$ = 250-MST graph. In the next section, we will explore some of the test power and computational time tradeoffs involved in this decision-making process.

C. AN ALTERNATIVE OPTIMAL SUBGRAPH: SHORTEST EDGES

Another proposed improvement to existing methods involves finding alternative optimal subgraphs that enable detecting group differences at lower computational costs. All of the graph-building (i.e. "edge-gathering") approaches discussed so far (MSTs, MWSSs, etc.) require integer programming optimization, which can be very time-consuming, especially when sample sizes are large. Historically, these graphs have been used in two-sample testing because they have some desirable structural properties. Using the known structure of these graphs enables us to derive theoretical moments (e.g. mean and variance) for a given test statistic under the null hypothesis. Deriving additional moments (e.g. skewness, kurtosis, etc.) provides a better estimation of the test statistic null distribution, which can make the computation of approximate p-values feasible. However, a statistical tool called *permutation testing* gives us a simple way to estimate the sampling distribution for any test statistic under the null hypothesis. Thus,

permutation testing allows us to use optimal subgraphs that lack well-understood statistical structure and still compute approximate p-values for two-sample testing.

We propose a graph-based approach that considers only a subset of shortest edges from the undirected complete graph on all observations. Intuitively, this approach seems to make sense since, if two samples are from different distributions, observations would be preferentially closer to those from the same sample than those from the other sample. In terms of computational times, this approach should be much faster than existing methods since its only requires a sorting algorithm instead of integer programming optimization. Furthermore, although the optimal subgraphs from such an approach lack a well-understood structure, permutation testing makes it possible to find approximate p-values. Put simply, permutation testing involves randomly permuting or “shuffling” the observation vertex labels many times (i.e. thousands) in order to approximate the test statistic null distribution and, in turn, compute p-values.

IV. SIMULATION RESULTS

In practice, the process of taking a set of observations from two groups and determining whether or not a statistically significant difference exists between those two groups involves four distinct steps. First, a data array (in which the rows are observations and the columns are attributes) is converted to a pairwise distance matrix by assigning an inter-point dissimilarity measure to the original data set. This dissimilarity measure may be Euclidean distance (distance as measured by a ruler), Gower distance (a commonly-used distance measure for mixed data), treeClust (a new distance measure based on classification or regression trees), or any other specified measure. The measure need not be a metric in a formal sense, but in practice it generally is one. Next, the pairwise distance matrix is used to create some sort of optimal subgraph of the complete graph. These optimal subgraphs may be minimum spanning trees (MSTs), minimum-weight spanning subgraphs (MWSSs), a collection of shortest edges, or any other specified subgraph. Then, a test statistic (e.g. Cross-Count, Scaled Within-Group Count, etc.) is computed using the edges from the optimal subgraph. Finally, permutation testing or an approximation of the test statistic null distribution is used to compute a p-value, which is then compared to a pre-specified significance level to conclude whether or not a difference exists. Figure 11 provides a basic flowchart of this process, along with some possible choice of test statistic, which we will describe next.

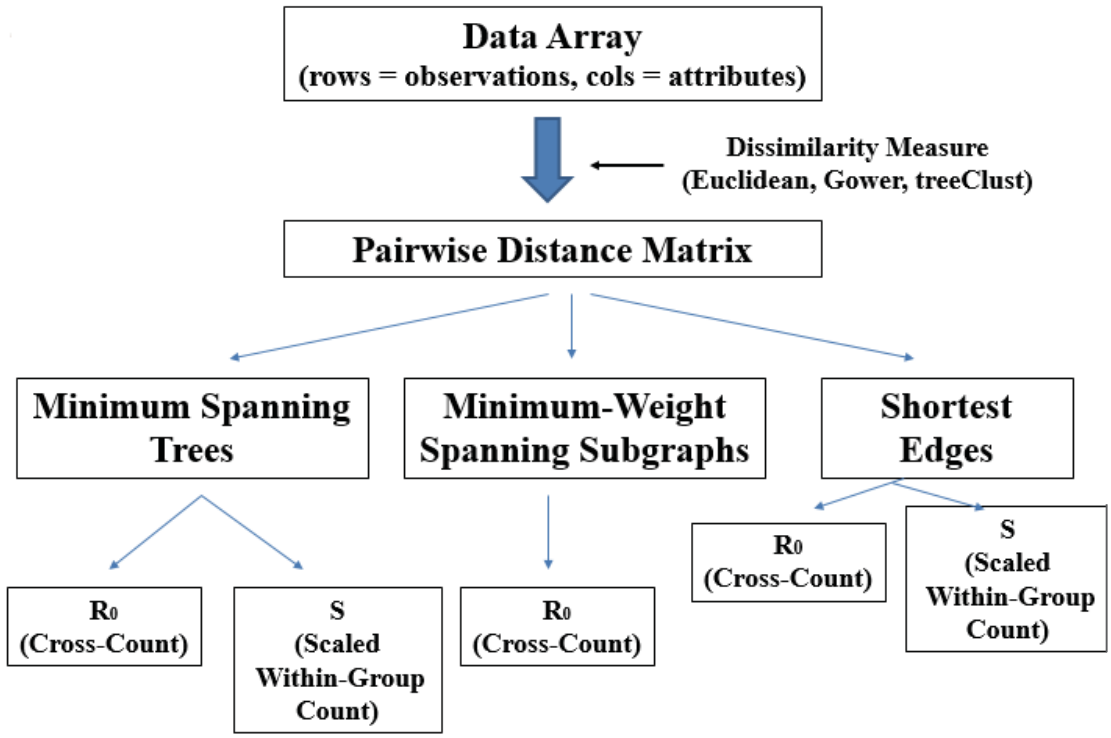


Figure 11: A basic flowchart showing the distinct steps from data array to test statistic.

A. VALIDATION AND REFINEMENT OF RESULTS FROM TEST POWER SIMULATION STUDIES IN CF2017

We began our simulations by validating and refining the power results published in Chen and Friedman (2017) for the classic Cross-Count (R_0) test and their new Scaled Within-Group Count (S) test on minimum spanning trees. Power in this context refers to the probability of correctly identifying a group difference when a group difference exists. In the following study, we simulate under conditions identical to those in CF2017, but with 1000 simulations per case rather than with 100 per case as is used in the original paper. The estimated margin of error, MOE, for these power estimates is

$$MOE = \sqrt{\frac{\hat{p}(1-\hat{p})}{nsims}},$$

where $\hat{p}$ is the estimated power ($0 \leq \hat{p} \leq 1$) and $nsims$ is the number of simulations. Thus, using 1000 simulations instead of 100 decreases the margin of error by a factor of $\frac{1}{\sqrt{10}}$.

To have a baseline for comparison in their simulation studies, Chen and Friedman chose the distribution to be multivariate normal in order to have the asymptotically most powerful tests based on normal theory – the Hotelling's two-sample T^2 test if assuming equal covariance matrices ("Hotelling's T^2), and the generalized likelihood ratio test if not assuming equal covariance matrices ("GLR"). In addition to the two tests based on the normal theory, they also included in their comparison the new Scaled Within-Group Count test on a 1-MST, 3-MST, and 5-MST ("S: 1-,3-,5-MST"), the classic Cross-Count test on MSTs ("R_0: 1-,3-,5-MST") and on minimum distance non-bipartite pairings ("R_0: 1-,3-,5-MDP"), as well as the degree test on a 1-MST ("deg 1"). All optimal subgraphs were constructed using the Euclidean distance. For a more detailed description of the above tests, see Chen and Friedman (2017).

Table 1, Table 2, and Table 3 show power results for three types of group difference. All values in the tables are listed as percentages between 0 and 100; that is, the percentage of trials in which the test correctly detects a group difference. Table 1 shows results for two multivariate normal distributions where their means are different (Δ corresponds to the magnitude of mean shift). The results range from low dimension ($dim = 2$) to high dimension ($dim = 100$). Table 2 shows results for two multivariate normal distributions where their standard deviations are different (σ corresponds to standard deviation difference). These results range from low dimension ($dim = 2$) to medium dimension ($dim = 20$). Table 3 shows results for two multivariate lognormal distributions differing in the location parameter (Δ corresponds to the difference of the two location parameters). Changing the location parameter affects both the mean and variance of a lognormal distribution, so this is both a location and scale alternative. For all cases, the specific location and/or scale alternative was chosen so that the tests have moderate power. The numbers in the upper rows of each table are the power results published in CF2017 using 100 simulation trials. The numbers in the lower rows of each table are our reproduction of a subset of those power results (specifically, from the two tests that use orthogonal minimum spanning trees). We chose to refine their power estimates using 1000 simulation trials so that they could be used with greater confidence in future comparisons. The tables below serve to verify the reproducibility of and to refine the CF2017 power results.

Table 1: Validation and refinement of power results from CF2017 Table 1. Numbers indicate the percentage of trials with significance less than 5%. Normal data. The means of the two distributions differ in Δ in L_2 distance. $n = m = 50$.

Location alternatives							
(100 trials, CF2017)							
dim	2	10	30	50	70	90	100
Δ	0.6	0.8	1.1	1.4	1.7	2	2
Hotelling's T^2	77	71	74	76	70	26	--
GLR	52	30	14	--	--	--	--
R_0: 1-,3-,5-MST	22 35 40	12 35 47	27 46 49	37 67 73	41 76 89	61 85 92	57 85 90
R_0: 1-,3-,5-MDP	9 25 32	10 26 38	18 36 43	21 47 64	27 63 86	41 74 89	50 75 87
deg 1	4	6	4	4	3	4	4
S: 1-,3-,5-MST	10 22 24	9 23 34	20 30 34	25 40 59	23 54 80	36 76 83	34 74 82
(1000 trials, Wallace)							
R_0: 1-,3-,5-MST	17 31 37	18 35 43	23 42 52	33 57 67	49 75 84	59 88 92	52 85 93
S: 1-,3-,5-MST	9 21 29	13 23 31	17 28 37	20 41 56	30 59 69	38 75 84	39 72 82

Note in the above table that the Hotelling's T^2 test performs very well in low-to-moderate dimensions since all assumptions for the Hotelling's T^2 test are satisfied. However, as the dimension increases, the power of the Hotelling's T^2 test diminishes. Furthermore, when the dimension becomes greater than or equal to the sample size (as in the above case where *dim* = 100), the Hotelling's T^2 test cannot even be used. Table 1 above also shows that the Cross-Count test and the Scaled Within-Group Count test are not severely limited by dimension size, and the classic Cross-Count test on MSTs slightly outperforms the new Scaled Within-Group Count test on MSTs for normal location differences.

Table 2: Validation and refinement of power results from CF2017 Table 2. Numbers indicate the percentage of trials with significance less than 5%. Normal data. The two distributions differ in σ. $n = m = 50$.

Scale alternatives				
(100 trials, CF2017)				
dim	2	5	10	20
σ	1.4	1.25	1.2	1.15
Hotelling's T^2	7	7	5	5
GLR	69	42	28	12
R_0: 1-,3-,5-MST	22 34 41	12 22 24	7 17 28	7 15 18
R_0: 1-,3-,5-MDP	16 28 36	12 14 17	7 9 18	5 5 10
deg 1	8	27	59	62
S: 1-,3-,5-MST	20 43 56	37 64 64	57 76 78	66 73 80
(1000 trials, Wallace)				
R_0: 1-,3-,5-MST	17 22 30	11 17 22	11 18 25	7 15 21
S: 1-,3-,5-MST	14 43 61	34 59 66	55 73 78	66 82 83

Note in the above table that, since the equal covariance matrices assumption for the Hotelling's T^2 test is not satisfied in this scenario, the Hotelling's T^2 test performs very poorly. The GLR test performs well in low dimensions (*dim* = 2) but its power decreases very rapidly as dimension increases due to issues with parameter estimation in high dimensions. The degree test, which has no power to detect normal location group differences (Table 1), shows impressive power in high dimensions but is still outperformed by the Scaled Within-Group Count test in all scenarios. Table 2 above also suggests that the Scaled Within-Group Count test on MSTs significantly outperforms the Cross-Count test on MSTs for normal scale alternatives. This highlights a well-known limitation of the Cross-Count test to detect scale differences among groups, especially in high dimensions.

Table 3: Validation and refinement of power results from CF2017 Table 3. Numbers indicate the percentage of trials with significance less than 5%. Product lognormal data, $n = m = 50$.

Log location alternatives						
(100 trials, CF2017)						
dim	2	10	30	50	70	90
Δ	0.8	1	1.3	1.3	1.5	1.7
Hotelling's T^2	82	81	79	52	39	20
GLR	27	18	16	--	--	--
R_0: 1-,3-,5-MST	38 58 62	26 49 58	22 45 51	14 44 52	16 48 60	21 42 53
R_0: 1-,3-,5-MDP	25 44 54	18 34 50	11 31 40	11 23 35	15 36 49	12 34 47
deg 1	4	10	29	41	50	47
S: 1-,3-,5-MST	19 39 53	25 46 57	43 52 61	40 57 62	46 65 69	51 69 75
(1000 trials, Wallace)						
R_0: 1-,3-,5-MST	33 55 64	24 44 57	26 46 57	18 30 41	14 33 43	15 36 46
S: 1-,3-,5-MST	20 43 57	25 49 57	46 62 66	40 52 53	49 55 61	55 62 71

Table 3 above suggests that the Cross-Count test might outperform the Scaled Within-Group Count test in low dimensions, but the Scaled Within-Group Count test becomes increasingly dominant as dimension increases.

B. EXPLORING THE EFFECTS OF INCREASED SUBGRAPH DENSITY ON STATISTICAL TEST POWER

After refining and verifying the reproducibility of the power results of CF2017, we explored the effects of increasing subgraph density on statistical test power. Specifically, we simulated the same data from three tables discussed above; however, instead of building a 1-, 3-, and 5-MST, we investigated the value of building a 25-MST. Note that a 25-MST corresponds to 50% of the available edges of the complete graph on 100 observations, since a 25-MST has $25\times(100-1)$ edges, while the total number of edges is $\binom{100}{2} = 50\times(100-1)$ edges. Table 4 below suggests that the Cross-Count test outperforms the Scaled Within-Group Count test for normal location alternations and also that increasing graph density improves test power in both cases for normal location alternatives. Next, Table 5 suggests that the Scaled Within-Group Count test outperforms the Cross-Count test for normal scale alternatives. This table also seems to show that increasing graph density improves power for the Scaled Within-Group Count test but is inconclusive for the Cross-Count test under normal location alternatives. Finally, Table 6

suggests that the Scaled-Within Group Count test generally outperforms the Cross-Count test for lognormal location alternatives. Additionally, this table seems to show that increasing graph density improves power for the Scaled Within-Group Count test but is inconclusive for the Cross-Count test under lognormal location alternatives.

Table 4: Power estimates for normal mean alternatives using the Cross-Count test (R_0) and the Scaled Within-Group Count test (S) with different graph densities (1-, 3-, 5-, and 25-MST). *See Table 1 for comparison.*

Location alternatives							
(1000 trials, Wallace)							
dim	2	10	30	50	70	90	100
Δ	0.6	0.8	1.1	1.4	1.7	2	2
R_0: 1-,3-,5-MST	17 31 37	18 35 43	23 42 52	33 57 67	49 75 84	59 88 92	52 85 93
R_0: 25-MST	**65**	**66**	**76**	**87**	**97**	**99**	**99**
S: 1-,3-,5-MST	9 21 29	13 23 31	17 28 37	20 41 56	30 59 69	38 75 84	39 72 82
S: 25-MST	**55**	**52**	**62**	**77**	**92**	**99**	**97**

Table 5: Power estimates for normal scale alternatives using the Cross-Count test (R_0) and the Scaled Within-Group Count test (S) with different graph densities (1-, 3-, 5-, and 25-MST). *See Table 2 for comparison.*

Scale alternatives				
(1000 trials, Wallace)				
dim	2	5	10	20
σ	1.4	1.25	1.2	1.15
R_0: 1-,3-,5-MST	17 22 30	11 17 22	11 18 **25**	7 15 **21**
R_0: 25-MST	**40**	**26**	22	16
S: 1-,3-,5-MST	14 43 61	34 59 66	55 73 78	66 82 83
S: 25-MST	**72**	**80**	**91**	**95**

Table 6: Power estimates for lognormal location parameter alternatives using the Cross-Count test (R_0) and the Scaled Within-Group Count test (S) with different graph densities (1-, 3-, 5-, and 25-MST). *See Table 3 for comparison*.

Log location alternatives						
(1000 trials, Wallace)						
dim	2	10	30	50	70	90
Δ	0.8	1	1.3	1.3	1.5	1.7
R_0: 1-,3-,5-MST	33 55 64	24 44 57	24 46 **57**	18 30 **41**	14 33 **43**	15 36 **46**
R_0: 25-MST	**84**	**75**	51	26	26	29
S: 1-,3-,5-MST	20 43 57	25 49 57	46 62 66	40 52 53	49 55 61	55 62 71
S: 25-MST	**79**	**73**	**74**	**59**	**71**	**77**

Figure 12 shows the comparison of algorithm runtimes for a 1-, 3-, 5-, and 25-MST. Note that the x-axis corresponds to the total sample size (N) and the y-axis corresponds to the logarithm of the computational times.

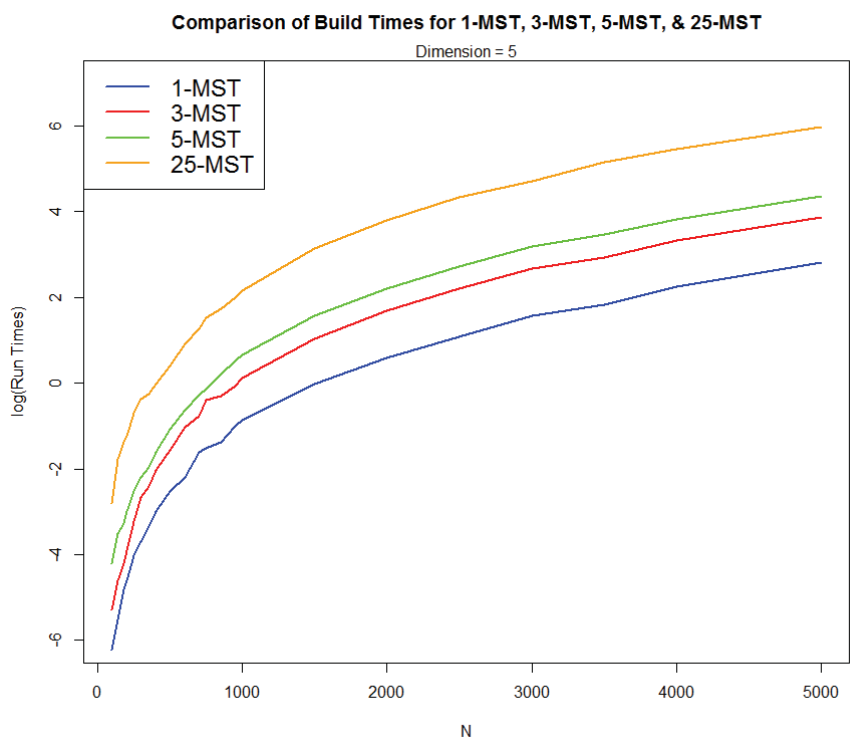


Figure 12: Comparison of relative computational times needed to build a 1-, 3-, 5-, and 25-MST

Contrary to the conjecture of CF2017, increasing graph density does seem to increase test power with no counter-information degradation, all the way up to including $\frac{1}{2}\times\binom{N}{2}$ edges, especially when using the Scaled Within-Group Count Test. As expected, however, increasing graph density also increases computational costs, and these costs become more pronounced as sample sizes increase. Ultimately, test users must take into consideration such factors as *the need for test power*, *sample sizes under consideration*, and *available computing resources* when determining the "optimal" graph density to use. As we will see shortly, our exploratory work aims to improve test power, decrease computational costs, and reduce user intervention.

C. EXAMINING Ru2014 AND CF2017 IN COMPETITION UNDER THE SAME TEST CONDITIONS

We now turn to examine two recent tests with strong demonstrated performance characteristics, Ruth (2014) and Chen and Friedman (2017), in competition under the same conditions. Ru2014 and CF2017 each report impressive test power results but under different conditions. The goal here is to assist users in their choice among these two tests for use in application. Recall that Ru2014 recommends building minimum-weight spanning subgraphs (MWSS), while CF2017 recommends building successive minimum spanning trees (MST). Also, recall that Ru2014's two-sample test is based on the number of cross-group edges, whereas CF2017's test is based on the scaled number of within-group edges for each group. Table 7, Table 8, and Table 9 below compare the simulated power results for these two tests. The data used in these tests are from the same specified distributions discussed previously (first is normal location alternative, second is normal scale alternative, third is lognormal location alternative). In order to put these tests on a level playing field, we have chosen optimal subgraph densities that have approximately the same number of graph edges. For example, in the case of 100 observations, a 10-MWSS has $10\times\frac{N}{2} = 10\times\frac{100}{2} = 500$ edges and a 5-MST has $5\times(N-1) = 5\times 99 = 495$ edges. These two numbers are close enough to infer that the two tests are using roughly the same "amount of dissimilarity information". In the tables below, a bolded number indicates which case has the best power for that particular scenario.

Table 7: Comparison of power estimates for normal mean alternatives using the Cross-Count test (R_0) with Minimum-Weight Spanning Subgraphs (Ru2014 approach) and the Scaled Within-Group Count test (S) with Minimum Spanning Trees (CF2017 approach).

Location alternatives							
(1000 trials, Wallace)							
dim	2	10	30	50	70	90	100
Δ	0.6	0.8	1.1	1.4	1.7	2	2
R_0: 2-,10-,50-MWSS	20 47 **63**	27 54 **68**	31 63 **82**	38 79 **92**	55 91 **98**	70 98 **100**	69 98 **100**
S: 1-,5-,25-MST	9 29 55	13 31 52	17 37 62	20 56 77	30 69 92	38 84 99	39 82 97

Table 8: Comparison of power estimates for normal scale alternatives using the Cross-Count test (R_0) with Minimum-Weight Spanning Subgraphs (Ru2014 approach) and the Scaled Within-Group Count test (S) with Minimum Spanning Trees (CF2017 approach).

Scale alternatives				
(1000 trials, Wallace)				
dim	2	5	10	20
σ	1.4	1.25	1.2	1.15
R_0: 2-,10-,50-MWSS	20 37 13	16 23 13	13 17 14	9 16 13
S: 1-,5-,25-MST	14 61 **72**	34 66 **80**	55 78 **91**	66 83 **95**

Table 9: Comparison of power estimates for lognormal location parameter alternatives using the Cross-Count test (R_0) with Minimum-Weight Spanning Subgraphs (Ru2014 approach) and the Scaled Within-Group Count test (S) with Minimum Spanning Trees (CF2017 approach).

Log location alternatives						
(1000 trials, Wallace)						
dim	2	10	30	50	70	90
Δ	0.8	1	1.3	1.3	1.5	1.7
R_0: 2-,10-,50-MWSS	39 71 **86**	30 60 **89**	27 60 **89**	22 50 **82**	24 60 **89**	29 66 **93**
S: 1-,5-,25-MST	20 57 79	25 57 73	46 66 74	40 53 59	49 61 71	55 71 77

Figures 13, 14, and 15 show three comparisons of graph build times for comparable densities of MSTs and MWSSs. All of the MSTs were constructed in R using the function “mst” from the package “igraph”. The “mst” function from this package finds the minimum spanning trees of complete, undirected graphs using Prim’s algorithm. All of the MWSSs were constructed in R using the function “rRegMatch” from the package “AcrossTic”. In general, the problem of finding MWSSs of a complete, undirected graph can be solved using binary integer linear

programming. Since this particular falls into a special class of combinatorial optimization problems called “fractional b-matchings”, the binary constraint may be relaxed and, with certain light conditions on r, the linear relaxation guarantees binary solutions. Thus, these algorithm runtimes do not necessarily represent the theoretical fastest runtimes, but instead the amount of time that an R user would have to wait for these graphs to be built. In these figures, note that the x-axis corresponds to the total sample size (N) and the y-axis corresponds to the logarithm of the algorithm runtimes.

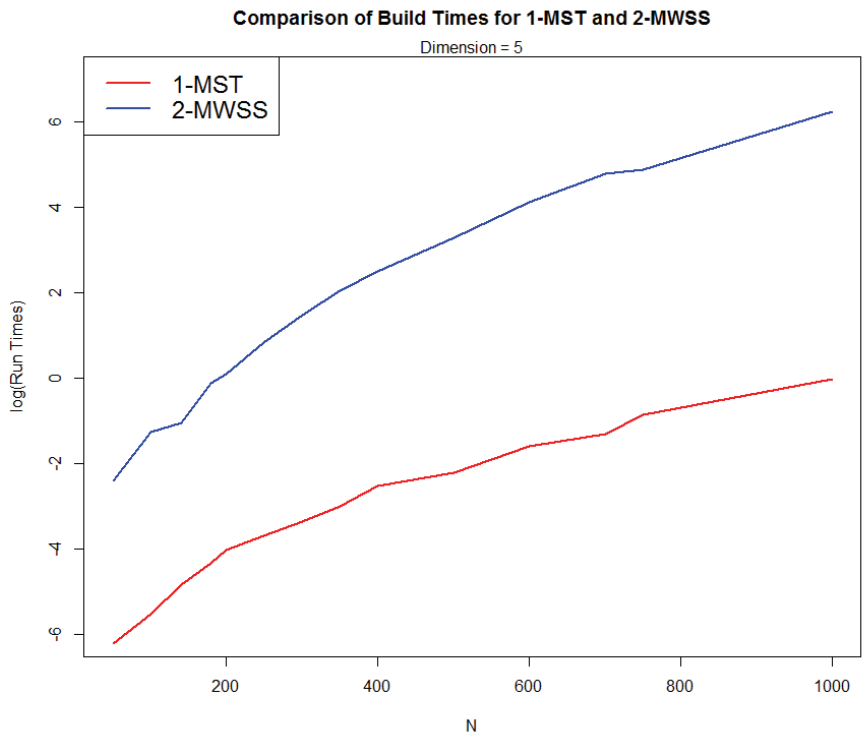


Figure 13: Comparison of run times required to build a 1-MST and a 2-MWSS

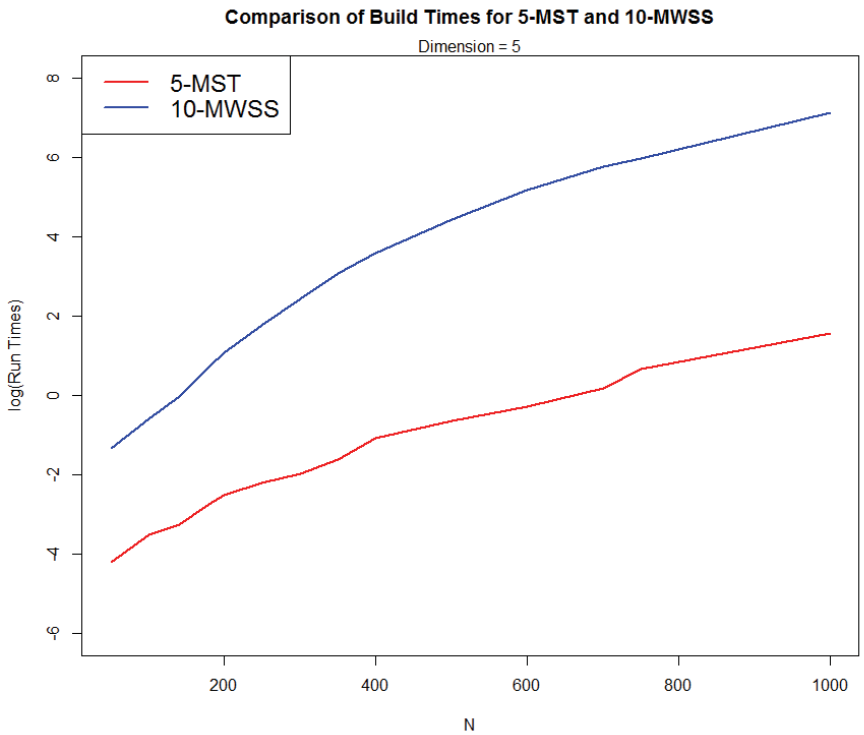


Figure 14: Comparison of run times required to build a 5-MST and a 10-MWSS

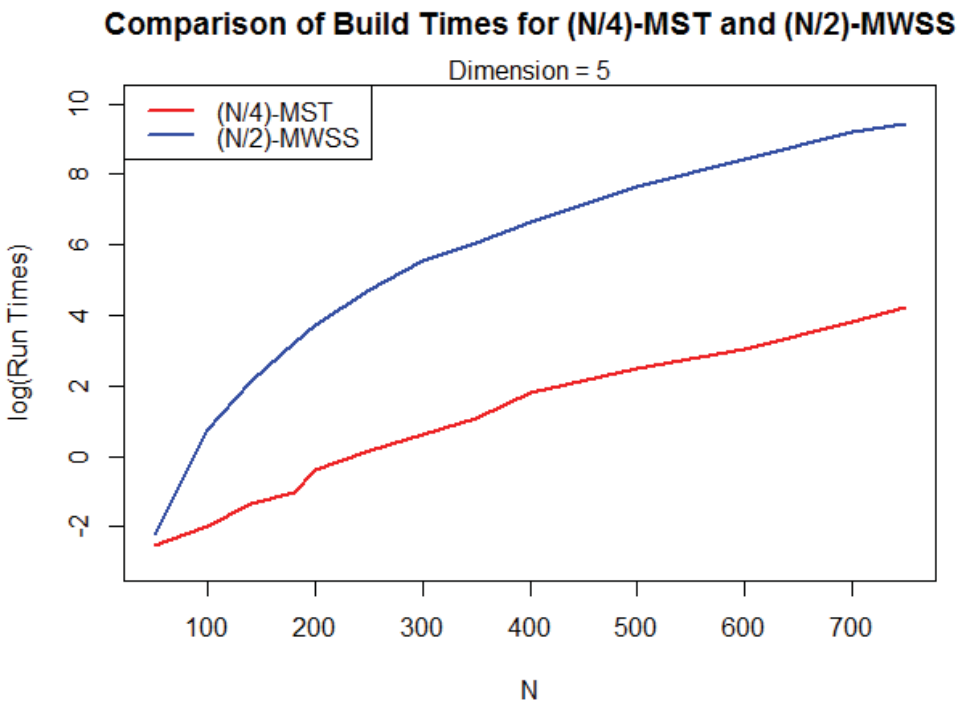


Figure 15: Comparison of run times required to build a (N/4)-MST and a (N/2)-MWSS

As these figures show, CF2017 has much faster graph-building ("edge-gathering") times than Ru2014. Based on the results in the tables above, Ru2014 seems to be the best choice if the user is looking for a location difference and if sample sizes are reasonably small. However, CF2017 seems to be the best choice if the user is looking for a scale difference or if sample sizes are large. This finding confirms the well-known fact that cross-count techniques, such as the one used in Ru2014, suffer against scale alternatives. If the user does not know what kind of difference might be present between the two groups, CF2017 seems to be the fastest, most robust option. However, the Ru2014 convention of using $\frac{1}{2}\times\binom{N}{2}$ edges seems best in all cases.

D. EXPLORING A NEW, ALTERNATIVE OPTIMAL SUBGRAPH: 50% SHORTEST EDGES

We now turn to our exploratory work in finding alternative optimal subgraphs that enable finding group differences at lower computational costs and possibly even with higher power. So far, from the previous simulation studies conducted, we have concluded that using the Scaled Within-Group Count Test (from CF2017) on a 25-MST (or, more generally, a *k*-MST that contains roughly 50% of the total edges from the complete undirected graph) is the most robust test-graph combination for maximizing test power. As such, those power results (S: 25-MST) serve as a benchmark for our exploratory work. In the previous section ("Motivation for Proposed Improvements"), we suggested building an optimal subgraph by collecting the 50% shortest edges from the complete undirected graph. Table 10 below compares the simulated power results of using the Scaled Within-Group Count Test on a 25-MST with the simulated power results of using the Scaled Within-Group Count Test on the 50% shortest edges. Observe that the power results for all scenarios are very similar. In terms of test power, neither approach seems to outperform the other with any significance.

Table 10: Comparison of power estimates for all scenarios previously studied using the Scaled Within-Group Count Test (S) with Minimum Spanning Trees (CF2017 approach) and 50% Shortest Edges. *Exploratory work.*

Location alternatives							
(1000 trials, Wallace)							
dim	2	10	30	50	70	90	100
Δ	0.6	0.8	1.1	1.4	1.7	2	2
S : 25-MST	55	52	62	77	92	99	97
S : 50% Shortest	52	49	59	77	92	98	96

Scale alternatives				
(1000 trials, Wallace)				
dim	2	5	10	20
σ	1.4	1.25	1.2	1.15
S : 25-MST	72	80	91	95
S : 50% Shortest	82	85	95	98

Log location alternatives						
(1000 trials, Wallace)						
dim	2	10	30	50	70	90
Δ	0.8	1	1.3	1.3	1.5	1.7
S : 25-MST	79	73	74	59	71	77
S : 50% Shortest	79	69	67	59	67	75

However, Figure 16 shows the comparison of build times for similar densities of MWSSs, MSTs, and shortest edges. Note that the x-axis corresponds to the total sample size (N) and the y-axis corresponds to the logarithm of the computational times. Combinatorially, the number of edges in a $\frac{N}{2}$-MWSS is approximately equal to number of edges in a $\frac{N}{4}$-MST, both of which are approximately equal to the $\frac{N}{4} \times (N-1)$ shortest edges. Moreover, each of these densities roughly corresponds to 50% of the total number of edges in the complete undirected graph on N observations. Observe that collecting the shortest edges (using a sorting algorithm) takes significantly less time than building other optimal subgraphs (using integer programming). Based on these results, our approach here (using the 50% shortest edges) appears to retain the test power characteristics of the best existing tests while substantially cutting down on computational costs.

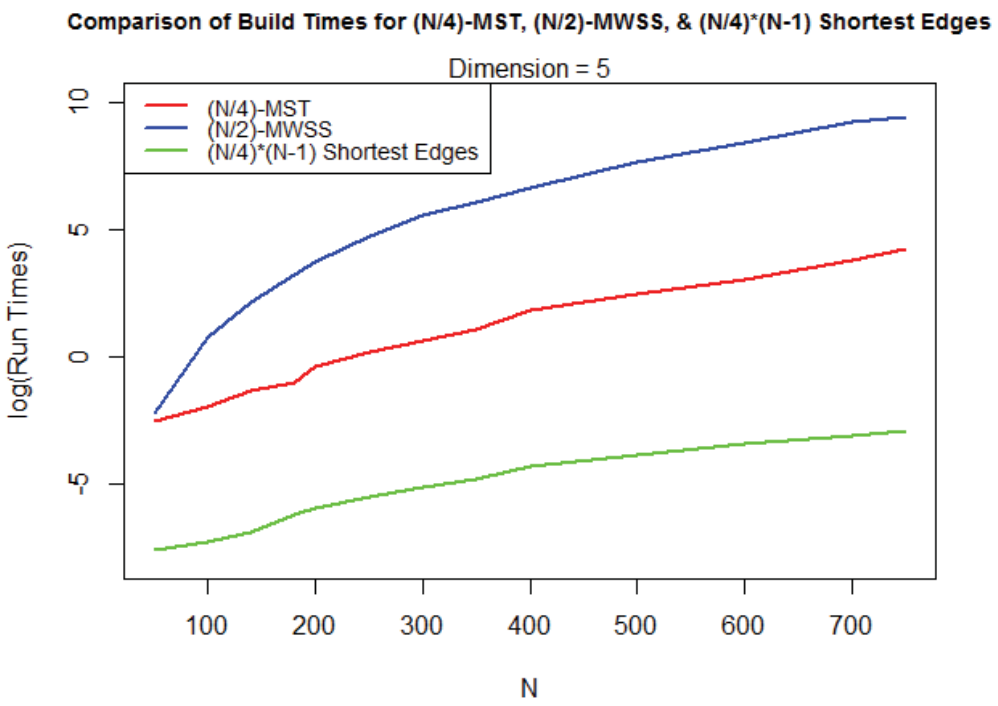


Figure 16: Comparison of computational times needed to build MWSSs & MSTs of comparable densities and to gather the equivalent number of shortest edges from the complete graph.

E. REAL DATA EXAMPLE: EVALUATING THE PERFORMANCE OF TREECLUST ON THE CLEVELAND HEART DISEASE DATA SET

All of the examples so far have come with "obvious" (or at least assumed) distance measures. We now turn to an exploration of treeClust, a newly-proposed dissimilarity measure for mixed data. Mixed data present challenges to determining interpoint dissimilarity. As discussed in the "Background" section, treeClust is a new measure to compute dissimilarity using classification and regression trees.

The UCI Machine Learning Repository has documented heart disease diagnosis data for 303 patients at the Cleveland Clinic foundation, plus 75 other attributes. We consider 5 quantitative and 8 categorical explanatory variables with separate binary response (presence of heart disease; evident in 139 of 303 patients). Table 11 below provides an example of 4 observations from the data set. Note that purple indicates quantitative explanatory variables and green indicates categorical explanatory variables.

Table 11: Example of 4 observations from the Cleveland Heart Disease data set (quantitative explanatory variables in purple; categorical explanatory variables in green).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
1	63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0
2	67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0
3	67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0
4	37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0

Using these real-world observations, the two-sample problem becomes: *Are the two diagnosis groups statistically different with respect to the explanatory variables?* For these data, it turns out that the two groups are not very difficult to differentiate (even a two-sample *t*-test on most of the individual variables would suffice). To test our dissimilarity approach, we should make the problem more challenging by adding noise to the data; this will make the groups more difficult to differentiate. For all-numeric, nicely-scaled, noise- and outlier-free data, existing techniques (e.g. Euclidean, Gower) are hard to beat for classification applications. However, for noise, incomplete, mixed data – as is often encountered in real life – the treeClust dissimilarity measure does very well. To compare Gower and treeClust, we do the following:

- Randomly permute some fraction of diagnosis labels, then estimate test power for each fraction.
- Add noise: Append a permuted copy of each of the 13 columns of data to the original data; do this 1, 20, 50, 100, and 150 times; resultant data frames have 26, 273, 663, 1313, and 1963 columns.

Figures 17, 18, 19, 20, and 21 compare the resilience of the well-known, existing Gower distance measure to the new treeClust distance measure. The *x*-axis corresponds to the number of unshuffled labels; this can be thought of as increasing the magnitude of the "distance" between the two groups (i.e. making the group different more apparent). The *y*-axis corresponds to estimated test power. When no noise is present, the existing Gower approach very slightly outperforms the new treeClust approach. However, when increasing amounts of noise are added, treeClust significantly outperforms Gower.

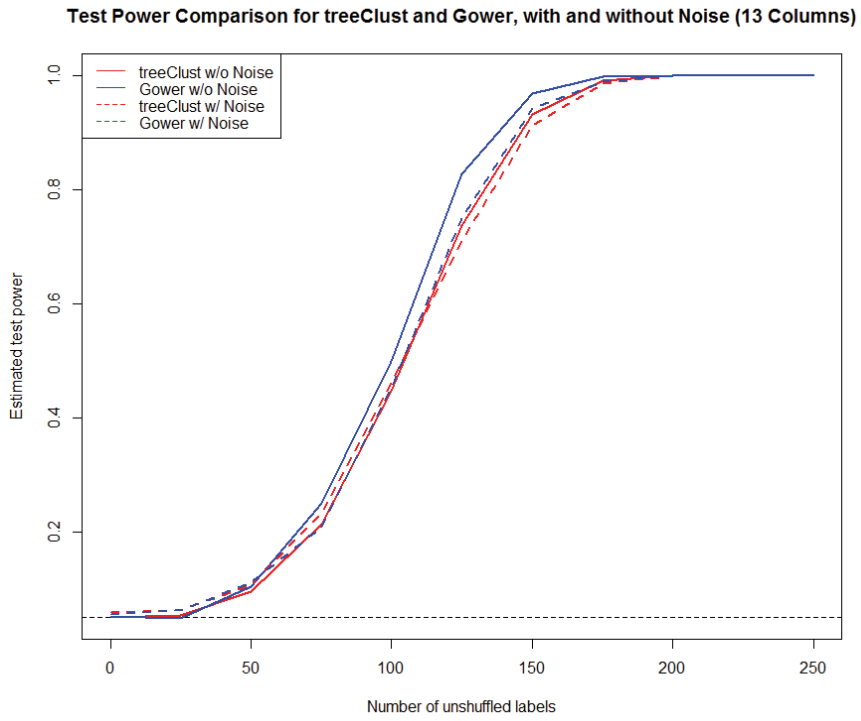


Figure 17: Comparison of power curves for treeClust distance measure (red dotted line) and Gower distance measure (blue dotted line), with ***13 columns of noise added.***

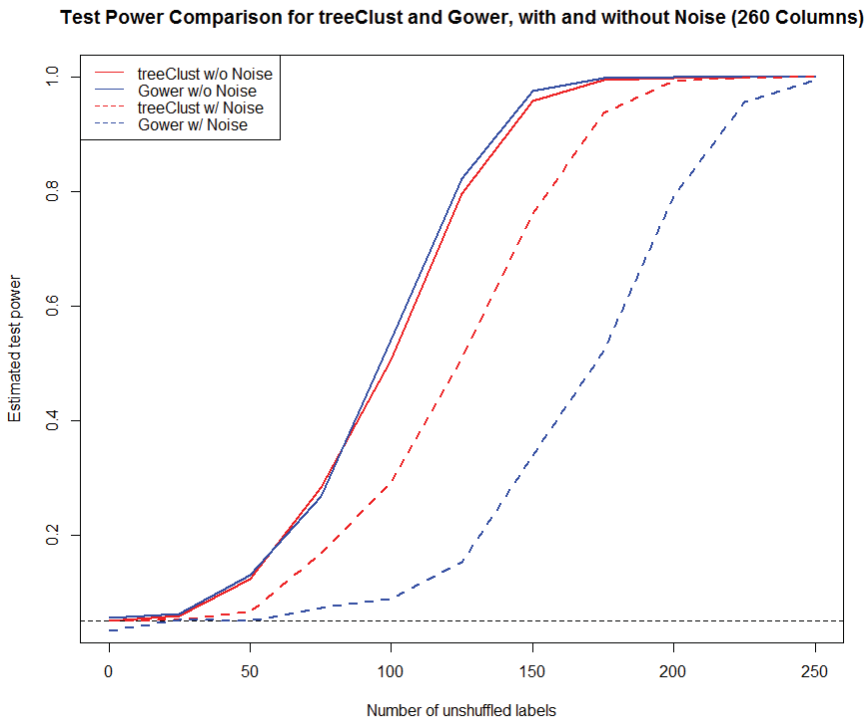


Figure 18: Comparison of power curves for treeClust distance measure (red dotted line) and Gower distance measure (blue dotted line), with ***260 columns of noise added.***

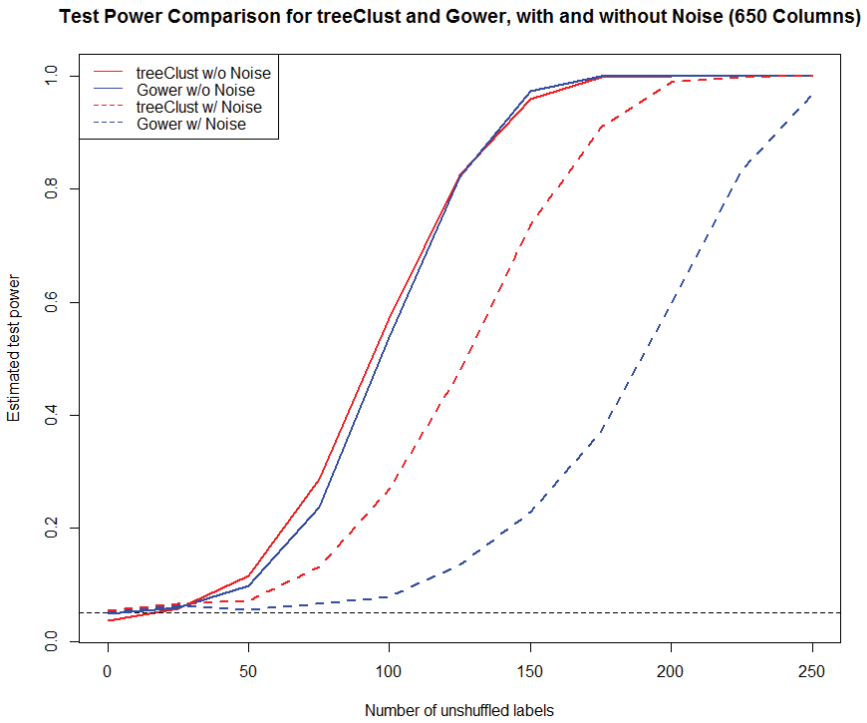


Figure 19: Comparison of power curves for treeClust distance measure (red dotted line) and Gower distance measure (blue dotted line), with ***650 columns of noise added.***

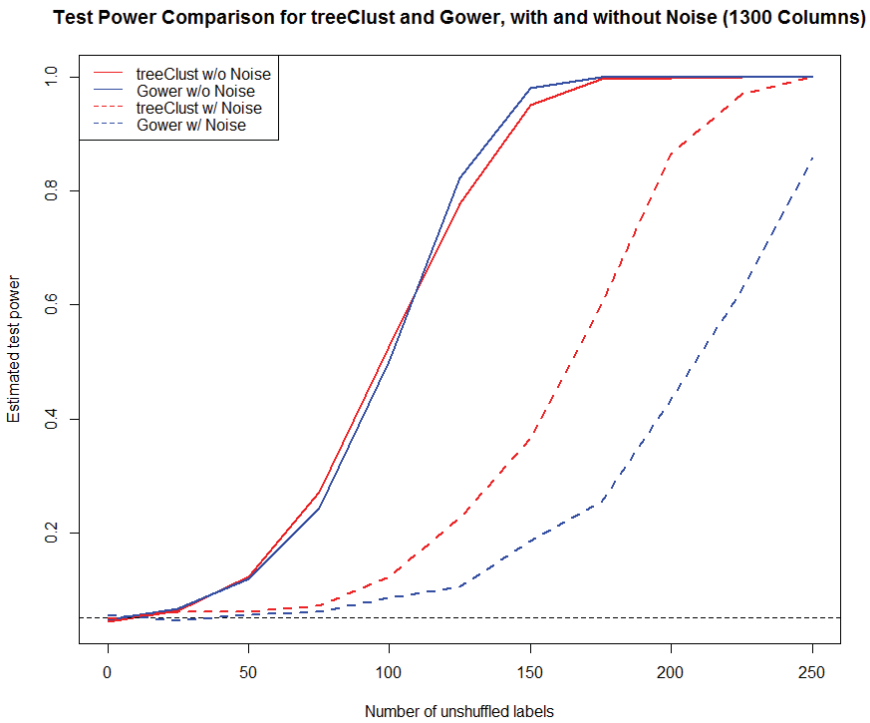


Figure 20: Comparison of power curves for treeClust distance measure (red dotted line) and Gower distance measure (blue dotted line), with ***1,300 columns of noise added.***

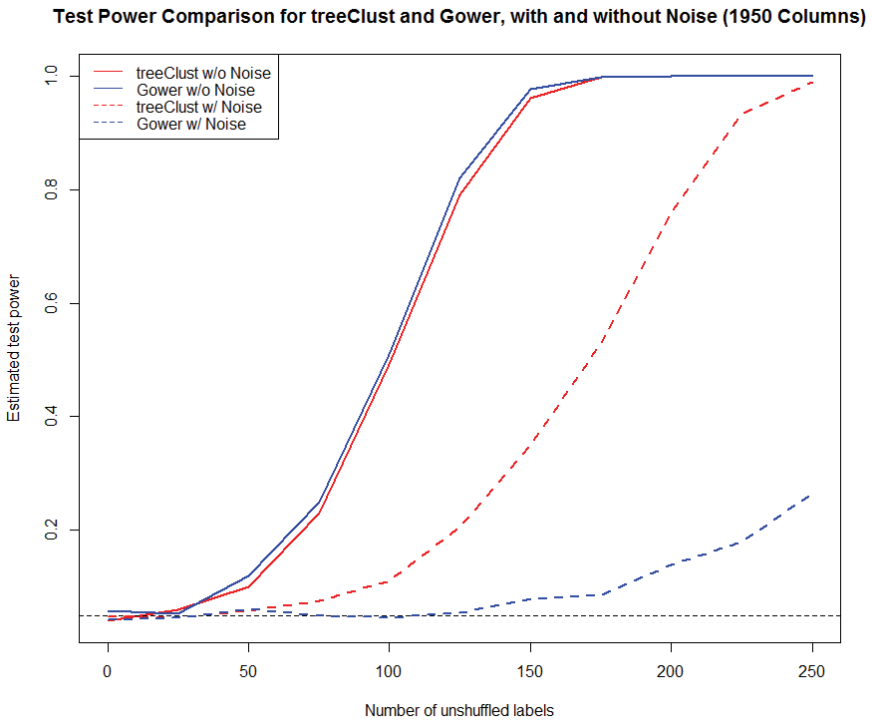


Figure 21: Comparison of power curves for treeClust distance measure (red dotted line) and Gower distance measure (blue dotted line), with ***1,950 columns of noise added.***

These figures suggest that dissimilarity measures using tree clustering are useful in the context of mixed data, and especially so when noise is present. treeClust retained impressive power characteristics even when 1,950 columns of noise were added to 13 columns of original data. At this time, the use of tree clustering for dissimilarity measures is not very well-studied; this serves as a potentially fruitful area for future work.

V. THE CUMULATIVE CROSS-COUNT (CCC) TEST

The findings presented thus far suggest that using denser subgraphs increases test power, but makes the test statistic null distribution difficult to characterize and is often more computationally complex. After comparing several different edge-counting approaches and test statistics, it is clear that there is an inherent tradeoff between test power (detecting a difference when a difference exists), mathematical complexity (characterizing the test statistic null distribution), computational costs (building optimal subgraphs and performing permutation tests), and user-friendliness (determining which test to use is a given scenario and optimal subgraph densities). We propose a promising new test that seems to provide a very competitive balance between the objectives above: the Cumulative Cross-Count (CCC) test.

A. CCC TEST METHODOLOGY AND OVERVIEW OF T_{CCC}

Instead of selecting a subset of edges with a very time-consuming optimality step and then counting the crossing edges on $\mathcal{G}^*$, the CCC test uses *all* edges for counting in the following manner:

- Rank the edges with respect to weight (based on some dissimilarity measure)
- Count *the accumulation of cross-counts with respect to edge-weight order*.
- Consider large deviations from expected counts as evidence of a group difference.

In the ensuing discussion, we will use the following notation throughout:

Let

$$X_{(i)} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ ranked edge is a crossing edge,} \\ 0 & \text{otherwise,} \end{cases}$$

and let

$$S_k = \sum_{i=1}^{k} X_{(i)}.$$

So, S_k is the total number of accumulated cross counts up to the kth ranked (shortest) edge. For our test statistic, we are interested in using the maximum deviation of S_k from the expected value of S_k under the null distribution, denoted μ_k, over all values of $k \in \left\{1, \dots, \binom{N}{2}\right\}$. Mathematically, this may be expressed as follows:

$$T_{\mathrm{CCC}} = \max_k |S_k - \mu_k|.$$

The Cumulative Cross-Count test rejects the null hypothesis of homogeneity for large T_{CCC}.

Figures 22-28 provide a visual description of the Cumulative Cross-Count test. Figure 22 shows all of the edges in a particular undirected, complete graph lined up based on rank and color-coded based on whether they are within-group edges or crossing edges. In this example, the group labels are randomly shuffled so that there is no group difference. The figure confirms the notion that, under the null hypothesis (i.e. when there is no group difference), each edge is equally likely to be a within-group edge or a crossing edge.

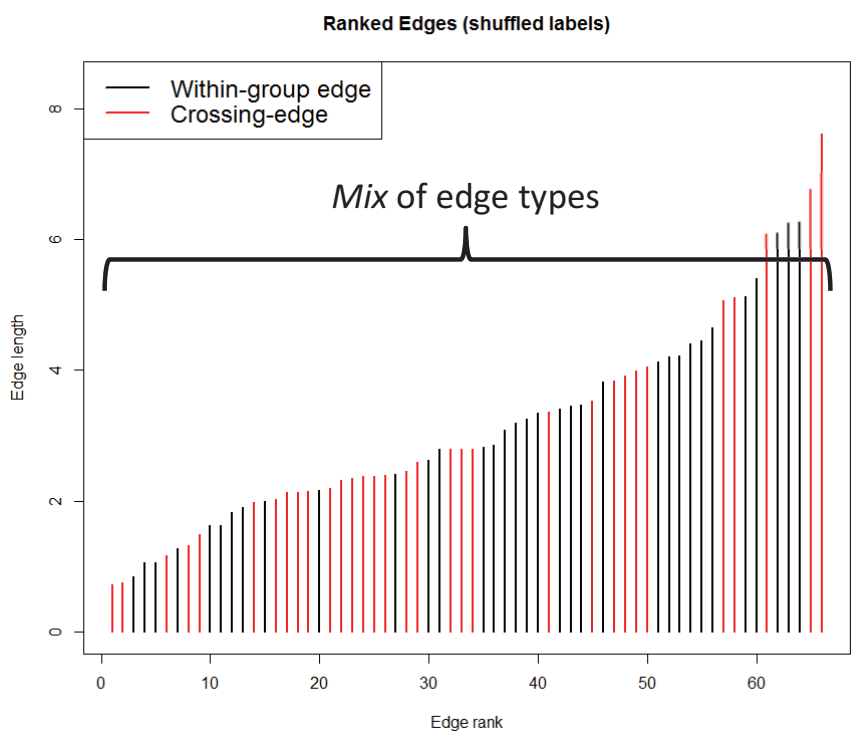


Figure 22: Graph of edges lined up by rank and colored black if within-group edges and red if crossing edges. No evidence of a group difference.

On the other hand, in Figure 23, the original group labels are preserved so that there is a group difference. This figure confirms the notion that, when a group difference exists, shorter edges are more likely to be within-group edges and longer edges are more likely to be crossing edges. Therefore, *crossing edges initially accumulate at a slower rate when a group difference actually exists compared to when there is no group difference*. This is the key idea behind the Cumulative Cross-Count test.

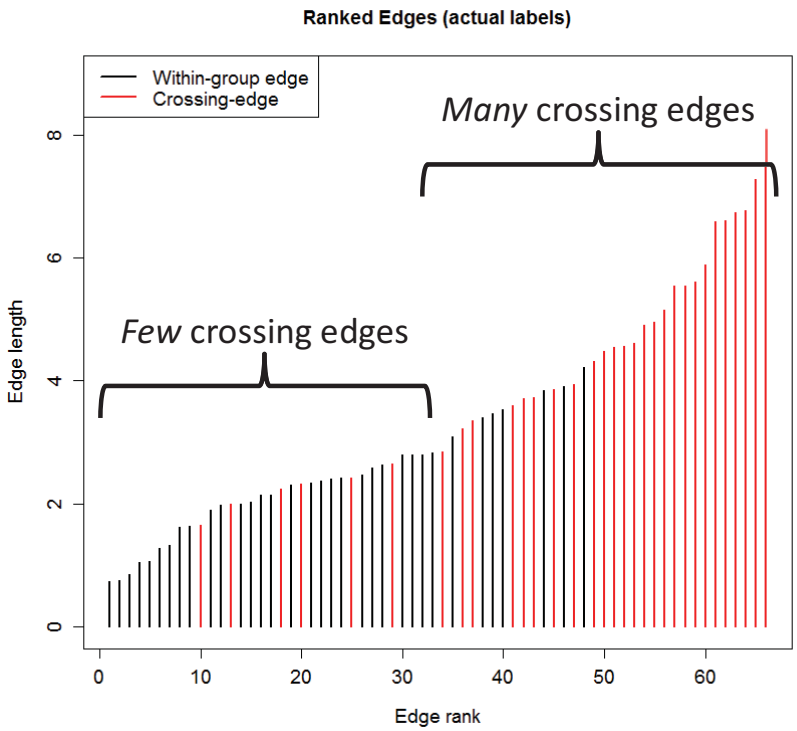


Figure 23: Graph of edges lined up by rank and colored black if within-group edges and red if crossing edges. Strong evidence of a group difference.

Figure 24 shows a single simulated trajectory that shows the accumulation of crossing edges (i.e. cross-counts) as a function of edge rank for a graph with permuted labels. Observe that the slope appears to be roughly constant with a little bit of variability. Since each edge is equally likely to be a crossing edge when there is no group difference, cross-counts are accumulated at a constant rate as edge rank increases. Thus, the figure below shows one simulated trajectory of what we would expect the graph of cumulative cross-counts to look like when there is no statistically significant difference between two groups.

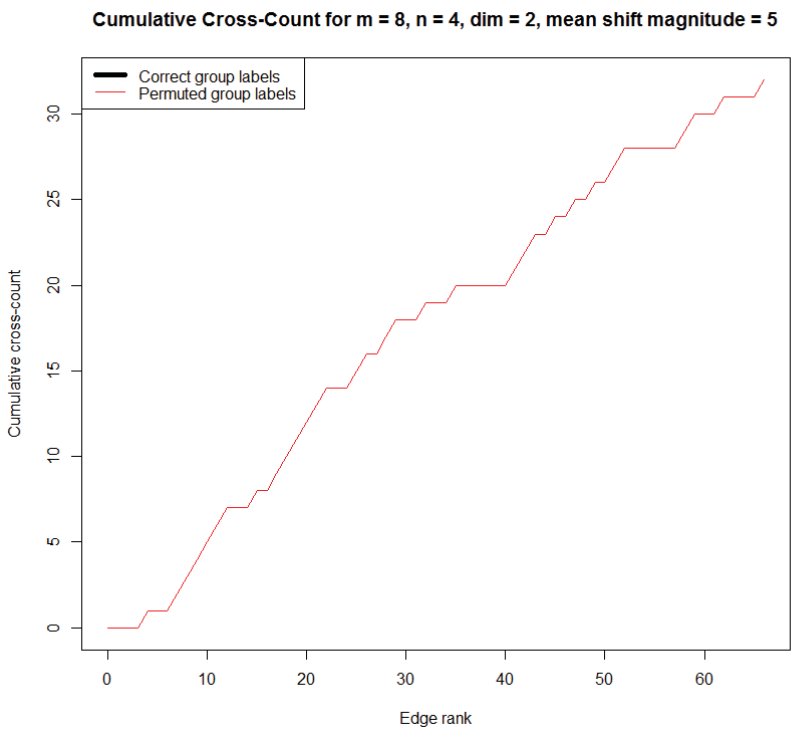


Figure 24: Plot of a single simulated cumulative cross-count trajectory. Permuted group labels.

Figure 25 shows ten simulated cumulative cross-count trajectories on graphs with permuted labels, instead of just one as in the previous figure. By displaying several trajectories of graph-label combinations for which there is no group difference, this figure provides a rough qualitative description of the null distribution of S_k, the cumulative cross-count, for a particular case where m = 8, n = 4, and dim = 2.

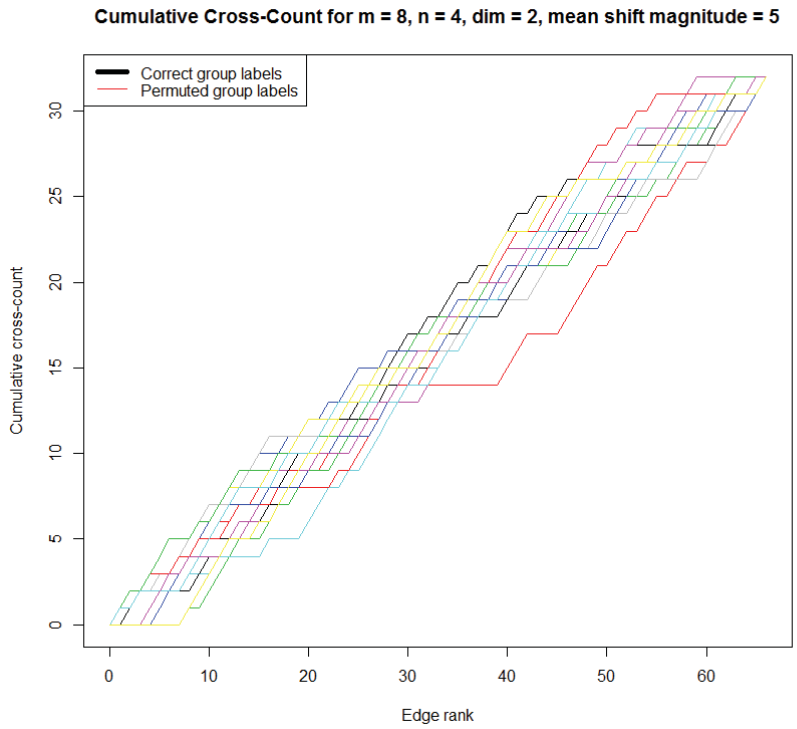


Figure 25: Plot of ten simulated cumulative cross-count trajectories. Permuted group labels. Rough qualitative estimation of the null distribution of S_k.

Figure 26 shows the same ten cross-count trajectories (for permuted group labels) from the previous figure as well as one cross-count trajectory when there is a mean shift magnitude of 5 and the actual group labels are preserved. As mentioned in the discussion for Figure 23, this figure reveals how cross-counts initially accumulate at a slower rate when there is an actual group difference compared to when there is no group difference. For our test statistic, T_{CCC}, we are interested in the maximum deviation between the expected cross-count under the null distribution and the actual cross-count.

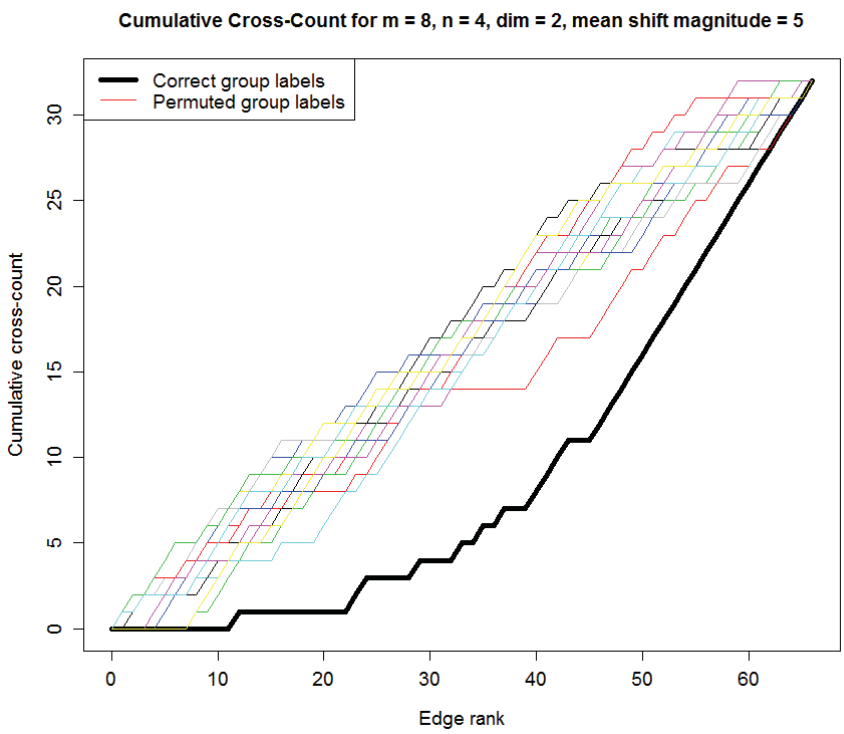


Figure 26: Plot of ten simulated cumulative cross-count trajectories with permuted group labels and one simulated cumulative cross-count trajectory with correct group labels. Note how cross-counts initially accumulate at a slower rate when there is an actual group difference (correct labels) compared to when there is no group difference (permuted labels).

Figure 27 provides a visual representation of the Cumulative Cross-Count test statistic, T_{CCC}. Instead of plotting the cumulative cross-count (as we did in the previous figures) which has a straight diagonal trajectory under the null distribution, here we instead plot the deviation from the expected cumulative cross-count (which, under the null distribution, has a horizontal trajectory with mean 0). In this figure, it is easy to see the maximum deviation from the expected cumulative cross-count occurs at the 39th ranked edge and has an absolute value of 11.9. This value is our test statistic, T_{CCC}.

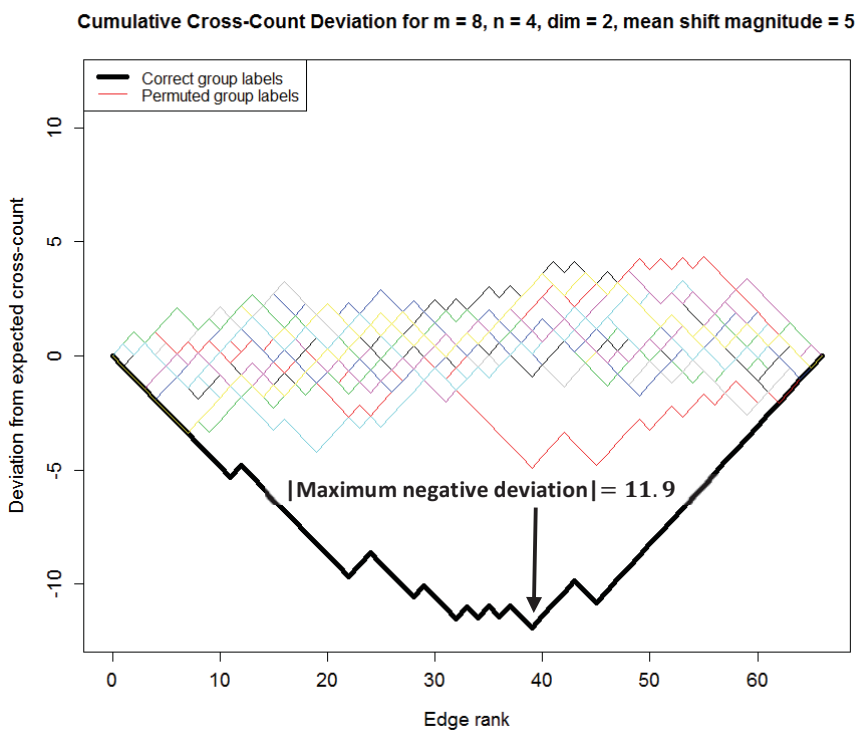


Figure 27: Plot of ten simulated cumulative cross-count deviation trajectories with permuted group labels and one simulated cumulative cross-count deviation trajectory with correct group labels. Note that T_{CCC} corresponds to the absolute value of the maximum negative deviation.

While figure 27 depicts a relatively simple example where the total sample size is only 12 (i.e. N = 12) and the observations are only 2-dimensional (i.e. *dim* = 2), Figure 28 shows a much more complex example where the total sample size is 200 (i.e. N = 200) and the observations are 100-dimensional (i.e. *dim* = 100). Instead of considering less than 70 total ranked edges as in the previous example, we now must consider almost 20,000 edges. Even so, the overall notion of the Cumulative Cross-Count test remains the same: cross-counts initially accumulate at a slower rate when there is an actual group difference compared to when there is no group difference. Collectively, the colored lines provide a rough estimate of the null distribution of S_k because the group labels were permuted and randomly assigned. The test statistic, T_{CCC}, allows us to characterize exactly how different the cumulative cross-count trajectory on our actual data is from what we would expect if no group difference existed. A larger test statistic provides stronger evidence of a group difference. In this case, the test statistic T_{CCC} corresponds to the deviation at the ~12000th ranked edge.

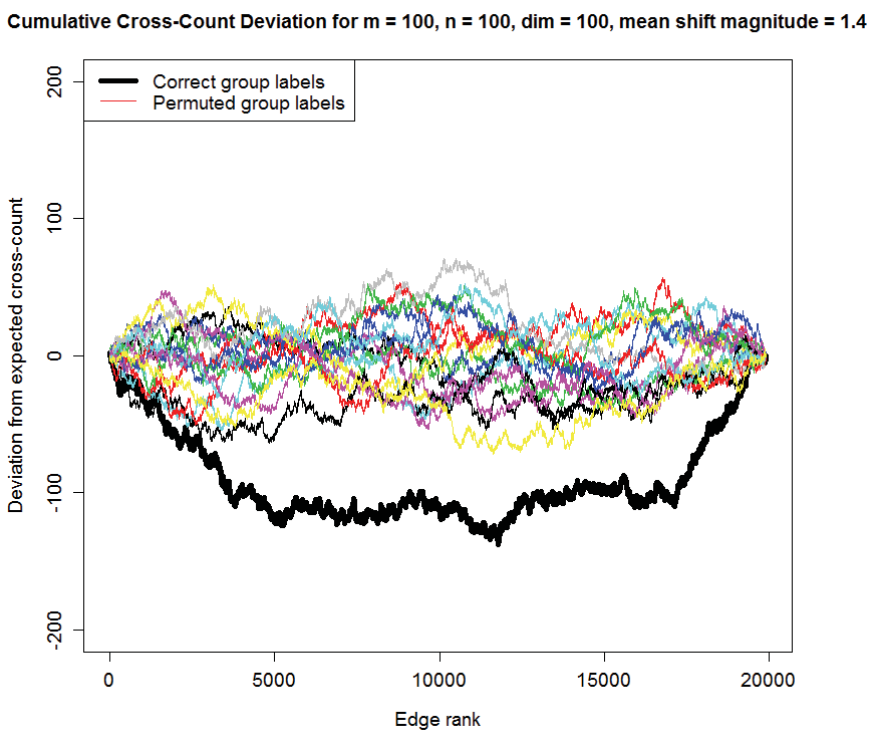


Figure 28: Visual representation of the Cumulative Cross Count (CCC) test. Two samples of 100 observations were drawn from 100-variate normal distributions with a mean shift magnitude of 1.4. The black line indicates deviation from the expected cumulative cross-count vs. edge rank when group labels are correctly assigned to their respective groups. The colored lines indicate deviation from the expected cross-count vs. edge rank when group labels are permuted and randomly assigned to the two groups (20 permutations shown above).

B. DERIVATION OF THE MEAN AND VARIANCE OF S_k

Recall that

$$X_{(i)} = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ ranked edge is a crossing edge,} \\ 0 & \text{otherwise,} \end{cases}$$

The expected value, variance, and covariance of $X_{(i)}$ may be derived as follows:

$$p \equiv E[X_{(1)}] = E[X_{(i)}] = P(X_{(i)} = 1) = \frac{mn}{\binom{N}{2}} \quad \forall i \in \{1, \dots, \binom{N}{2}\}.$$

since each edge is equally likely to be any rank under the under the hypothesis of homogeneity.

$$\sigma^2 \equiv \mathrm{Var}[X_{(1)}] = \mathrm{Var}[X_{(i)}] = p(1-p) = \frac{2mn(m(m-1)+n(n-1))}{N^2(N-1)^2}.$$

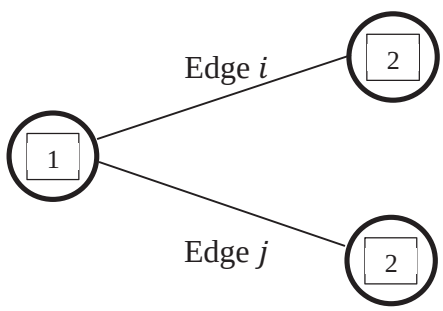
As in the FR1979 and CF2017 cases, the variance of the statistic of interest, S_k, depends on the number of adjacent edge pairs, C_k, among the k shortest edges. That is, define

$$C_k = \text{the number of adjacent edge pairs among the } k \text{ shortest edges.}$$

To compute this variance, it is necessary to compute $\mathrm{Cov}[X_{(i)}, X_{(j)} | C_k]$ for each (i, j).

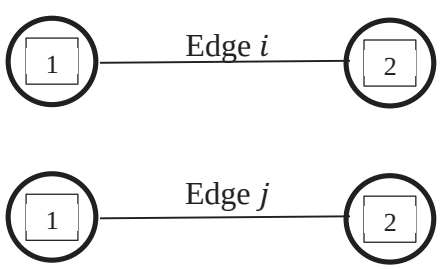
Thus, we must consider two cases:

<u>Case 1.</u> Edges (i) and (j) are incident.



$$p_{ij|1} = P\left(X_{(i)}X_{(j)} = 1 \middle| \text{ Case 1}\right) = \frac{m\binom{n}{2} + n\binom{m}{2}}{N\binom{N-1}{2}} = \frac{mn}{N(N-1)}.$$

<u>Case 2.</u> Edges (i) and (j) are non-incident.



$$p_{ij|2} = P\left(X_{(i)}X_{(j)} = 1 \middle| \text{ Case 2}\right) = \frac{4mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)}.$$

The mean of S_k is

$$E[S_k] = \sum_{i=1}^{k} E[X_{(i)}] = \sum_{i=1}^{k} p = kp = \frac{kmn}{\binom{N}{2}} \quad \forall k \in \left\{1, \dots, \binom{N}{2}\right\}.$$

The conditional variance of S_k can be found as follows:

$$\begin{aligned}\mathrm{Var}[S_k|C_k] = \sum_{i=1}^{k} \mathrm{Var}\left[X_{(i)}\middle|C_k\right] + 2\sum_{j=2}^{k}\sum_{i=1}^{j-1} \mathrm{Cov}\left[X_{(i)}, X_{(j)}\middle|C_k\right] = \sum_{i=1}^{k} \mathrm{Var}\left[X_{(i)}\middle|C_k\right] \\ + 2\sum_{j=2}^{k}\sum_{i=1}^{j-1}\left\{\mathrm{Cov}\left[X_{(i)}, X_{(j)}|C_k \text{ and Case 1}\right]P(\text{Case 1}) \right. \\ \left. + \mathrm{Cov}\left[X_{(i)}, X_{(j)}|C_k \text{ and Case 2}\right]P(\text{Case 2})\right\}.\end{aligned}$$

In the first term, $X_{(i)}$ is independent of C_k, so

$$\mathrm{Var}\left[X_{(i)}\middle|C_k\right] = \mathrm{Var}\left[X_{(i)}\right] = \frac{2mn(m(m-1)+n(n-1))}{N^2(N-1)^2}.$$

The second term reduces to

$$2\left(\sum_{\substack{\text{adjacent}\\ \text{edge pairs}}} \mathrm{Cov}\left[X_{(i)}, X_{(j)}\middle|C_k \text{ and Case 1}\right] + \sum_{\substack{\text{non-adjacent}\\ \text{edge pairs}}} \mathrm{Cov}\left[X_{(i)}, X_{(j)}\middle|C_k \text{ and Case 2}\right]\right)$$

$$\begin{aligned}&= 2\left[C_k\left[\frac{mn}{N(N-1)} - \left(\frac{mn}{\binom{N}{2}}\right)^2\right] + \left(\binom{k}{2} - C_k\right)\left[\frac{4mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} - \left(\frac{mn}{\binom{N}{2}}\right)^2\right]\right] \\ &= 2\frac{4mn}{N(N-1)}\left[C_k\left(\frac{1}{4} - \frac{(m-1)(n-1)}{(N-2)(N-3)}\right) + \binom{k}{2}\left(\frac{(m-1)(n-1)}{(N-2)(N-3)} - \frac{mn}{N(N-1)}\right)\right].\end{aligned}$$

Hence, the conditional variance of S_k is

$$\begin{aligned}\mathrm{Var}[S_k|C_k] = \frac{2kmn(m(m-1)+n(n-1))}{N^2(N-1)^2} \\ + 2\frac{4mn}{N(N-1)}\left[C_k\left(\frac{1}{4} - \frac{(m-1)(n-1)}{(N-2)(N-3)}\right) + \binom{k}{2}\left(\frac{(m-1)(n-1)}{(N-2)(N-3)} - \frac{mn}{N(N-1)}\right)\right].\end{aligned}$$

Figure 29 provides evidence that the above result is correct. Samples of size 10 and 30 (i.e. $m = 10$ and $n = 30$) are both drawn from a 5-variate standard normal distribution. The k shortest edges are computed and the theoretical variance (red line) is plotted over the corresponding sample variance (black dots), resulting from 2,000 node label permutations. It is clear that this theoretical curve (derived above) well captures the observed result from simulation.

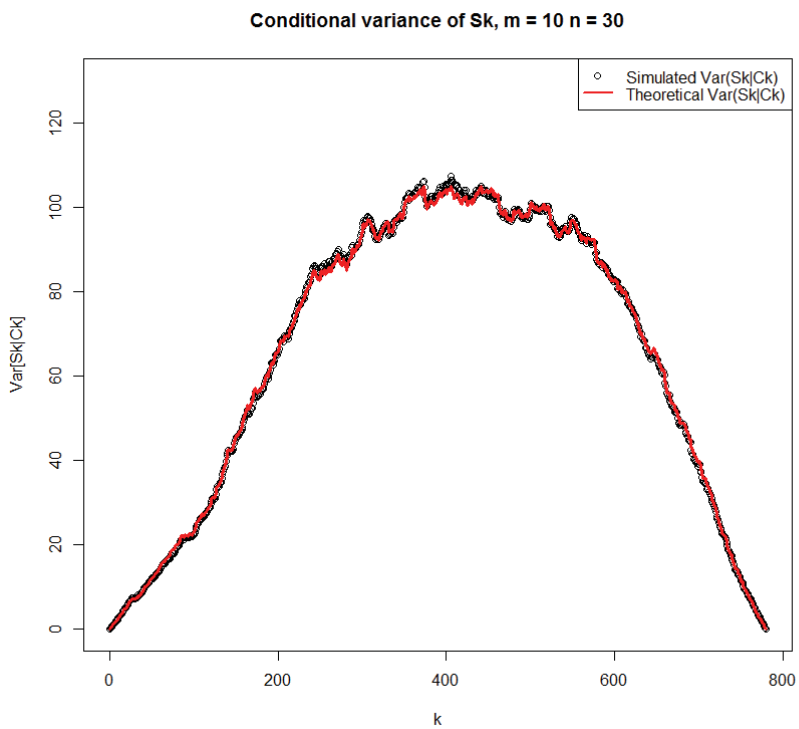


Figure 29: Overlaid plots of the theoretical and simulated conditional variance of S_k for two groups of size 10 and 30, respectively, drawn from a 5-variate standard normal distribution.

C. COMPARING CCC TEST POWER TO EXISTING STATE-OF-THE-ART TESTS USING SIMULATED DATA FROM CF2017

We now compare the power of the Cumulative Cross-Count test to that of the best existing parametric and nonparametric approaches using the familiar simulated data from CF2017. In Table 12, we present CCC test power estimates for detecting normal location group differences. We also include the test power estimates of Hotelling's T^2 (the asymptotically most powerful test based on normal theory if assuming equal covariance matrices), S: 25-MST (the CF2017 approach

with optimal subgraph density improvements as suggested in this paper), and R_0: 50-MWSS (the Ru2014 approach with optimal subgraph density improvements as discussed in this paper).

Table 12: Comparison of CCC test power estimates for detecting normal location group differences to the best available parametric and nonparametric tests. $n = m = 50$. 2000 simulations. Approximate p-values found using permutation testing.

Location alternatives							
dim	2	10	30	50	70	90	100
Δ	0.6	0.8	1.1	1.4	1.7	2	2
Best Parametric Approach (1940s)							
Hotelling's T^2	**77**	**71**	74	76	70	26	--
Best Chen and Friedman Approach (2017, *JASA*) with Subgraph Density Improvements							
S : 25-MST	55	52	62	77	92	99	97
Best Ruth Approach (2014, *JSDA*) with Subgraph Density Improvements							
R_0: 50-MWSS	63	68	**82**	92	**98**	**100**	**100**
Cumulative Cross-Count Test (2018)							
CCC Test	61	69	81	**93**	**98**	**100**	**100**

Note in the above table that the Hotelling's T^2 test performs very well in low-to-moderate dimensions since all assumptions for the Hotelling's T^2 test are satisfied. However, as the dimension increases, the power of the Hotelling's T^2 test diminishes. In higher dimensions, the new CCC test demonstrates very impressive power, outperforming the best available CF2017 approach and on par with the best available Ru2014 approach. These results suggest that the CCC test may be the best choice if the user is looking for a location difference between groups, especially in high dimensions.

In Table 13, we present CCC test power estimates for detecting normal scale group differences. We include test power estimates for the same approaches as in the previous table, except instead of Hotelling's T^2 (which is very weak for scale alternatives) we use GLR (the asymptotically most powerful test based on normal theory if not assuming equal covariance matrices).

Table 13: Comparison of CCC test power estimates for detecting normal scale group differences to the best available parametric and nonparametric tests. $n = m = 50$. 2000 simulations. Approximate p-values found using permutation testing.

Scale alternatives				
dim	2	5	10	20
σ	1.4	1.25	1.2	1.15
Best Parametric Approach (1930s-1940s)				
GLR	69	42	28	12
Best C&F Approach (2017) with Density Improvements				
S: 25-MST	**72**	**80**	**91**	**95**
Best Ruth Approach (2014) with Density Improvements				
R_0: 50-MWSS	13	13	14	13
Cumulative Cross-Count Test (2018)				
CCC Test	54	51	62	67

Note in the above table that the best available CF2017 approach outperforms all other approaches in every scenario. Even so, we see that the CCC test outperforms the best available parametric approach ("GLR"), except in very low dimensions. Moreover, the CCC test does not suffer nearly as much as the best Ru2014 approach against scale alternatives. These results suggest that the CCC test, although not the single most powerful test in these scenarios, retains acceptable power for detecting scale differences between groups.

In Table 14, we present CCC test power estimates for detecting lognormal location group differences. We include test power estimates for the same approaches as in Table 12, because Hotelling's T^2 is the best parametric test for finding a group difference under these conditions.

Table 14: Comparison of CCC test power estimates for detecting product lognormal location group differences to the best available parametric and nonparametric tests. $n = m = 50$. 2000 simulations. Approximate p-values found using permutation testing.

Log location alternatives						
dim	2	10	30	50	70	90
Δ	0.8	1	1.3	1.3	1.5	1.7
Best Parametric Approach (1930s-1940s)						
Hotelling's T^2	82	81	79	52	39	20
Best Chen and Friedman Approach (2017) with Density Improvements						
S: 25-MST	79	73	74	59	71	77
Best Ruth Approach (2014) with Density Improvements						
R_0: 50-MWSS	**86**	**89**	**89**	**82**	**89**	**93**
Cumulative Cross-Count Test (2018)						
CCC Test	85	79	83	74	79	88

Note in the above table that the best available Ru2014 approach outperforms all other approaches in every scenario. Even so, we see that the CCC test outperforms the best available parametric and CF2017 approaches, and its underperformance against the best Ru2014 approach is not considerable. These results suggest that the CCC test, although not the single most powerful test in these scenarios, retains impressive power for detecting group differences in non-normal data.

D. COMPARING ALGORITHM RUNTIMES BETWEEN CCC TEST AND EXISTING METHODS

We showed in the previous section that, in terms of test power, the Cumulative Cross-Count test is a peer competitor to even the most state-of-the-art two-sample statistical testing approaches. However, test power is only one of the factors that a user must consider when choosing the most appropriate test for his or her application. Another very important factor in these graph-theoretic approaches is the amount of time required to build the desired optimal subgraph. For instance, even the most powerful test can be useless if it takes an absurd amount of time or computing resources to run. Table 15 shows a comparison of algorithm run times needed to build various densities of minimum spanning trees, minimum-weight spanning subgraphs, as

well as to run the CCC test for different sample sizes. The dashes in the table below indicate that the particular time trial was manually terminated for taking an unreasonable amount of time.

Table 15: Comparison of computational times (in seconds) needed to build various densities of minimum spanning trees (MSTs) and minimum-weight spanning subgraphs (MWSSs), as well as to run the Cumulative Cross-Count (CCC) test. All algorithm run times are represented in seconds. These times do not necessarily represent the theoretical fastest run times, but the amount of time an ordinary R user would expect to wait (based on available packages and normal computing power).

Type of Optimal Subgraph	Sample Size (*N*)							
	100	**250**	**500**	**750**	**1000**	**5000**	**10000**	**15000**
1-MST	<0.01	0.01	0.06	0.19	0.34	14.79	99.92	--
3-MST	<0.01	0.05	0.2	0.67	1.06	45.11	--	--
5-MST	0.02	0.08	0.33	1.06	1.83	74.79	--	--
(*N*/4)-MST	0.05	0.62	6.93	28.8	67.22	13357.01	--	--
2-MWSS	0.28	2.31	27.02	132.20	506.47	--	--	--
6-MWSS	0.33	3.65	41.12	226.77	861.77	--	--	--
10-MWSS	0.38	5.82	83.88	390.47	1262.61	--	--	--
(*N*/2)-MWSS	2.09	112.61	2081.19	12406.44	--	--	--	--
CCC Test	<0.01	<0.01	0.02	0.08	0.18	1.15	4.02	10.19

The improvement in computational times of the CCC test over existing methods is incredible. For example, building a 375-MWSS on 750 observations (which uses 50% of the available edges from the complete, undirected graph) takes over 12,000 seconds, whereas the CCC test takes less than one-tenth of a second on 750 observations. In addition, running the CCC test on 10,000 observations takes just over 4 seconds; on the other hand, using any other sort of existing graph-theoretic approach (even a 1-MST) on 15,000 observations is not possible for an ordinary R user due to the very large amount of computer memory needed to store such a complex graph. Also note that, because of its design, the CCC test does not require the user to specify an optimal subgraph density. Instead, the CCC test considers every edge from the complete, undirected graph and does so with extreme efficiency. Based on its comparable power to existing state-of-the-art tests, its unmatched speed, and its overall user-friendliness, the Cumulative Cross-Count test appears to be among the best of its kind.

E. REAL DATA EXAMPLE: SENSORLESS DRIVE DIAGNOSIS DATA

We now turn to explore the performance of the Cumulative Cross-Count test on a real-world large-sized data set. The UCI Machine Learning Repository has documented features extracted from electric current drive signals from a drive which has intact and defective components. This Sensorless Drive Diagnosis quantitative data set includes 5,319 observations on 48 attributes (*dim* = 48), across 11 classes to be compared. We specifically investigate Class 2 and Class 6, although the results below hold generally for any comparison of two classes in this data set. In Figure 30, we show the results of two-sample testing on a small subset of observations ($m = n = 15$) from the data set. The black lines represent average p-values over 1,000 simulations using the univariate two-sample *t*-test on each individual attribute, the blue line represents the average p-value over 1,000 simulations using the multivariate CCC test on all attributes, and the red dashed line represents the significance level.

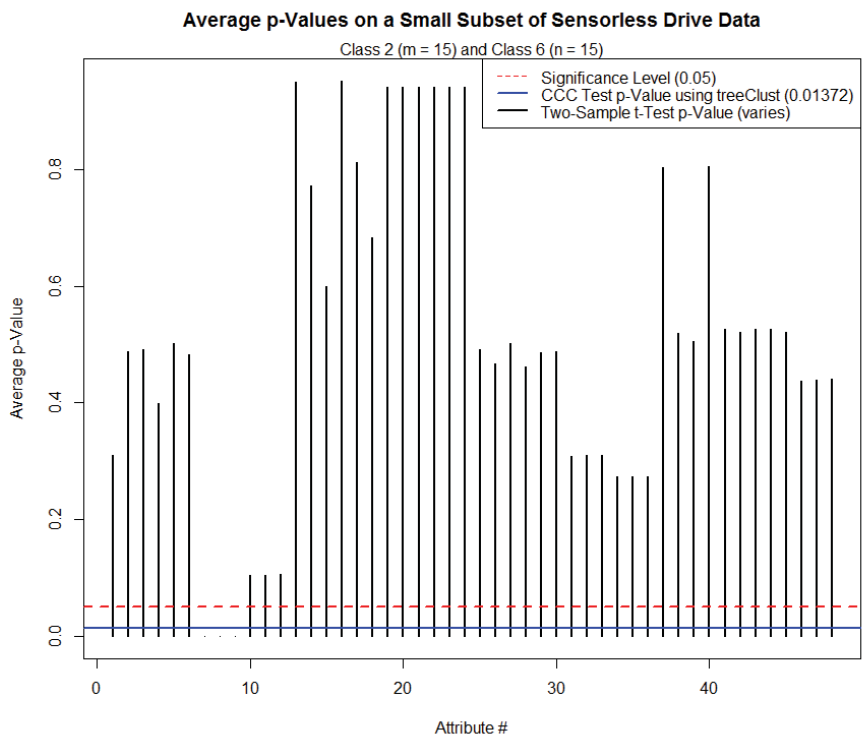


Figure 30: Average p-values over 1,000 simulations on a small subset of observations ($m = n = 15$) from the Sensorless Drive Data. The black lines represent the average p-values on each individual attribute using the univariate two-sample *t*-test. The blue line represents the average p-value on the entire set of attributes using the multivariate CCC test. The red dashed line indicates the significance level (0.05). Note that only 3 of the 48 *t*-test average p-values fall below the significance level of 0.05, and the single CCC test average p-value is 0.01372.

This result highlights a few notable advantages of the CCC test over existing parametric tests. The first advantage has to do with sample size limitations. In the above example, where the dimension (*dim* = 48) is larger than the sample size (*N* = 30), the multivariate Hotelling's T^2 test is not even an option. Instead, we are left with performing a univariate *t*-test on each of the individual attributes. This leads to the multiple testing problem, which occurs when one considers a set of statistical inferences simultaneously. The idea is that the more inferences are made, the more likely erroneous inferences are to occur. So, if we actually want to perform 48 univariate *t*-tests simultaneously, this technique would require a stricter significance level (i.e. lower than 0.05) for individual comparisons, so as to compensate for the number of inferences being made. In the CCC test, we avoid the multiple testing problem because we are only performing a *single* two-sample test, one which simultaneously considers all of the attributes. In Figure 31 below, we perform the same tests as in Figure 30 but instead consider the entire set of observations (*m* = *n* = 5319).

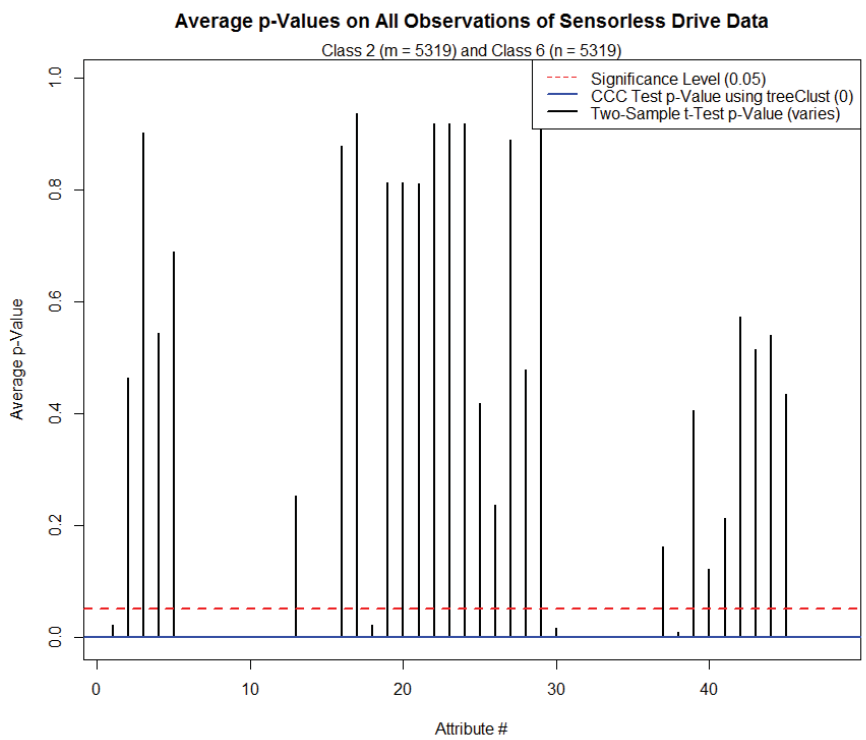


Figure 31: P-values on the entire set of observations (*m* = *n* = 5319) from the Sensorless Drive Data. The black lines represent the p-values on each individual attribute using the univariate *t*-test. The blue line represents the p-value on the entire set of attributes using the multivariate CCC test. The red dashed line indicates the significance level (0.05). Note that only 22 of the 48 *t*-test average p-values fall below 0.05, and the single CCC test average p-value is 0.

This result also highlights some additional advantages of the CCC test over existing parametric tests. In this case, where the sample size (N = 10,638) is larger than the dimension (dim = 48), we should be able to perform the multivariate Hotelling's T^2 test. The problem here is that, to find the Hotelling's T^2 test statistic, it is necessary to calculate the covariance matrix (which is a pooled matrix) and then invert it. When there is a high degree of correlation in the variables (as is present in the Sensorless Drive Data), the covariance matrix could be close to a singularity (the determinant near zero) and non-invertible. If the covariance matrix is non-invertible, we cannot perform the Hotelling's T^2 test. There are some methods for "cleaning up" the data to make it suitable for the Hotelling's T^2. One option is simply to remove the highly-correlated variables from the data matrix and to calculate the covariance matrix again until the matrix can be inverted; another option is to perform some sort of principal component analysis on the data to convert the correlated variables into a set of values of linear uncorrelated variables called principal components. The techniques are, at best, cumbersome and, at worst, beyond the scope of most individuals who use two-sample tests in their fields. On the other hand, the CCC test (along with, in this case, the treeClust dissimilarity measure) does not get impeded by correlated data. Since the multivariate Hotelling's T^2 test is not an option without data manipulation, most users are left with performing the standard univariate *t*-test on each of the attributes. This brings us back to the multiple testing problem. Ultimately, the two-sample problem on this large-sized set of real-world data demonstrates not only the power of the CCC test but also its remarkable user-friendliness. The R code for the Cumulative Cross-Count test may be found in Appendix 4.

VI. CONCLUSION & OPPORTUNITIES FOR FUTURE WORK

In this paper, we propose innovations to increase the power of state-of-the-art graph-theoretic two-sample statistics tests. These innovations yield promising results which strongly suggest that, in many cases, they outperform existing methods. Through a variety of simulation studies, we demonstrate that increasing graph density can lead to impressive improvements in test power. In addition, we show that using the shortest graph edges (which is computationally "cheap") produces very similar power results to using other optimal subgraphs (which are computationally "expensive"). Our simulated competition between Ru2014 and CF2017 also provides useful insights on which test is best to use for a given scenario. We explore the noteworthy resilience of a new interpoint dissimilarity measure, treeClust, using a real-world data set with added noise. Lastly, we propose a promising new test, the Cumulative Cross-Count (CCC) test, with remarkable potential and derive some of its moment information.

With this project coming to its conclusion, the body of work invites several possibilities for further exploring and extension. One opportunity involves investigating the impact of increasing subgraph density on two groups with unequal sample sizes. Thus far, our recommendations concerning "optimal" subgraph densities have only been tested on data with equal group sizes. Future work might explore how these recommendations hold up to data with unequal sample sizes and, if necessary, adjust or generalize the recommendations appropriately.

Another opportunity lies in deriving additional moment information for T_{CCC}. In this paper, we derive the mean and conditional variance of S_k. However, recall that the test statistic for the Cumulative Cross-Count test is $T_{\mathrm{CCC}} = \max_k |S_k - \mu_k|$. While it is very unlikely that an exact null distribution for the CCC test statistic, T_{CCC}, can be found, a possible acceptable alternative to finding an exact distribution would be finding a result which bounds tail distribution properties for T_{CCC}. Relating the CCC test statistic to properties of a Brownian bridge appears to be a promising starting point.

In some cases, the decision of which dissimilarity measure to use is even more important than the decision of which statistical test to use. In this paper, we show the impressive capability of treeClust to handle mixed, noisy, and correlated data. At this time, the use of tree clustering for dissimilarity measures is not very well-studied; this serves as a potentially fruitful area for future work.

Finally, in our two real-world data sets, the Cleveland Heart Data and the Sensorless Drive Diagnosis Data, the group differences were pretty substantial. In many cases, even a simple univariate two-sample *t*-test was sufficient for detecting a difference. To make this program more challenging, we added noise and/or limited the number of observations under consideration. It would be interesting and useful to apply our approaches, such as the CCC test with treeClust, to other real-world scenarios but where a group difference is much more subtle. Areas such as healthcare clinical trials, industrial quality assurance, and marketing campaign analysis have readily available data and provide excellent opportunities for future work.

APPENDIX 1: BASIC GRAPH THEORY DEFINITIONS AND DISCUSSION OF MINIMUM SPANNING TREES

We will begin by reviewing some basic terms from graph theory. In mathematics, ***graphs*** are mathematical structures used to model pairwise relations between objects. A ***vertex*** or ***node*** is the fundamental unit of which graphs are formed. Although the terms may be used interchangeably, we will exclusively refer to this unit as a vertex (or plural, ***vertices***). An ***undirected graph*** consists of a set of vertices and a set of ***edges*** (unordered pairs of vertices), while a ***directed graph*** consists of a set of vertices and a set of ***arcs*** (ordered pairs of vertices). All of the graph-theoretic approaches discussed herein use undirected graphs only. We say that an edge ***links*** the two vertices defining it and that it is ***incident*** on both of them. The ***degree*** of a node is the number of edges incident on it.

The earliest and most well-known graph-theoretic approaches to the two-sample problem make use of a particular type of graph known as a minimum spanning tree. We will define some additional terms that are relevant to minimum spanning trees. A ***path*** between any two specified vertices is an alternating sequence of vertices and edges with the specified vertices as first and last elements, all other vertices distinct, and each edge linking the two vertices adjacent to it in the sequence. A ***connected graph*** has a path between any two distinct vertices. A ***cycle*** is a path that begins and ends with the same vertex. A ***tree*** is a connected graph with no cycles. A ***subgraph*** of a given graph is a graph which has all of its vertices and edges in the given graph. A ***spanning subgraph*** of a given graph is a subgraph which contains every vertex of the given graph. A ***spanning tree*** of a graph is a spanning subgraph that is a tree. Because of its structure, there is a unique path between every two vertices in a tree. Thus, a spanning tree of a given connected graph features a path between any two vertices of the given graph.

An ***edge-weighted graph*** is a graph with a real number (weight) assigned to each edge. A ***minimum spanning tree*** (MST) of an edge-weighted graph is a spanning tree for which the sum of edge weights is a minimum. ***Orthogonal minimum spanning trees*** are simply additional minimum spanning trees of the same edge-weighted graph such that none of the edges from previous minimum spanning trees are reused.

In the two-sample problem, we begin with the complete graph, which consists of the N pooled sample data points (observations) in $\mathbb{R}^{dim}$ (where dim is the dimension size, or number

of covariates) as vertices, and edges linking all pairs of observations. By its combinatorics arrangement, the complete, undirected graph has $\binom{N}{2} = \frac{N(N-1)}{2}$ edges. If we consider the weight associated with each edge to be Euclidean distance, which is the ordinary "straight-line" distance between two points in space, or a generalized dissimilarity between the two vertices defining it, then the minimum spanning tree of this graph is the subgraph of minimum total distance (dissimilarity value) that provides a path between every two vertices.

Minimum spanning trees have two notable properties that make them especially useful for application to the two-sample problem. First of all, because of its structure as a spanning tree, a minimum spanning tree always contains $N - 1$ edges. Secondly, because the sum of its edge weights are a minimum, the vertex pairs defining the edges represent observations that tend to be close to each other. These properties prove to be useful in developing test statistics and deriving moments of the test statistic null distributions.

We build all of our minimum spanning trees in R, which is one of the most widely used languages and environments for statistical computing and graphics, using the function "mst" from the package "igraph". The "mst" function from this package finds the minimum spanning trees of undirected graphs using Prim's algorithm. Prim's algorithm is a heuristic (greedy) algorithm that operates by building the tree one vertex at a time, beginning with an arbitrary starting vertex, and at each step adding the cheapest possible link from the tree to another vertex. The output of this algorithm is a minimum spanning tree of the original weighted undirected graph.

We can describe Prim's algorithm informally as performing the following steps:

1. Arbitrarily choose a single vertex from the graph. This initializes the tree.
2. Grow the tree by one edge. Of all the possible edges that connect the tree to vertices outside of the tree, find the minimum-weight edge and transfer it to the tree.
3. Repeat Step 2 until every vertex is in the tree.

In building an optimal solution, Prim's algorithm uses the greedy principle, which involves making the choice that is best at the current state. Although the greedy principle in general is not guaranteed to provide an optimal solution, it does in the case of Prim's algorithm.

This is true because Prim's algorithm maintains a set E with the following condition: Prior to each iteration, E is a subset of edges of some minimum spanning tree. Beginning with an arbitrary initial starting vertex, we determine the shortest (lowest cost) edge that connects that

vertex to another vertex. This edge is the first element of E. At each subsequent iteration, we determine an edge (i, j), where i is a vertex inside the tree and j is a vertex outside the tree, that can be added to E such that the condition stated above is not broken. Thus, if E is a subset of edges of some minimum spanning tree before a given iteration, then $E \cup \{(i, j)\}$ is a subset of edges of some minimum spanning tree before the following iteration. Since every vertex is added to the tree in this manner, therefore the output of this algorithm is guaranteed to be the minimum spanning tree containing every vertex of the original weighted undirected graph. Hence, it is optimal.

APPENDIX 2: DERIVATION OF THE MEAN AND VARIANCE OF T_{FR} (FR1979)

To find the mean and variance of T_{FR} under the null hypothesis, we proceed using the following notation. Let $\mathcal{G}$ be the complete undirected graph $(\mathbb{Z}_N, E_N)$ where the vertex set $\mathbb{Z}_N$ consists of the indices $1, 2, \ldots, N$ and the edge set consists of all $\binom{N}{2} = \frac{N(N-1)}{2}$ pairs of vertices. Partition $\mathbb{Z}_N$ into two sets S and T, with $|S| = m$ and $|T| = n$, so $m + n = N$. By convention, we say that a node is labeled S if it is an observation from Group 1 and a node a labeled T if it is an observation from Group 2. Build a minimum spanning tree (MST) on $\mathcal{G}$; recall that this subset of E_N contains $N - 1$ edges. Number the $N - 1$ edges of the MST arbitrarily and define Y_i, $1 \leq i \leq N - 1$, as follows:

$$Y_i = 1 \text{ if the } i\text{th edge links nodes from different samples.}$$
$$= 0 \text{ otherwise.}$$

Let T_{FR} be the total number of cross-group edges in the MST plus one. This is equivalent to the number of subtrees remaining after we remove every cross-group edge from the MST. The FR1979 test is considered to be a multivariate generalization of the Wald-Wolfowitz runs test; as such, we also may think of T_{FR} as the number of runs in the MST.

Then,

$$T_{\mathrm{FR}} = \sum_{i=1}^{N-1} Z_i + 1 \text{ and } E[T_{\mathrm{FR}}] = \sum_{i=1}^{N-1} E[Z_i] + 1. \tag{1}$$

Note that

$$E[Z_i] = P(Z_i = 1). \tag{2}$$

Now, $P(Y_i = 1)$ is the probability that the two nodes defining this edge are labeled S and T or T and S. In other words, this is the probability that the ith edge in the MST is a cross-group edge. Since, under H_0, each edge is equally likely to be a cross-group edge, finding these probabilities is a simple combinatorics argument: the total number of cross-group edges in $\mathcal{G}$ divided by the total number of edges in $\mathcal{G}$. Thus, these probabilities are

$$P(ST) = P(TS) = \frac{mn}{N(N-1)},$$

so that

$$P(Y_i = 1) = P(ST) + P(TS) = \frac{2mn}{N(N-1)}, \tag{3}$$

and from (1)

$$E[T_{\mathrm{FR}}] = \sum_{i=1}^{N-1} \frac{2mn}{N(N-1)} + 1 = \frac{2mn}{N} + 1. \tag{4}$$

This is the same result as in the univariate case (Wald-Wolfowitz, 1940).

The variance of T_{FR} under H_0 can be calculated similarly. For the Y_i random variables, we have that the variance of their sum is equal to the sum of their covariances. Thus,

$$\mathrm{Var}[T_{\mathrm{FR}}] = \mathrm{Var}\left[\sum_{i=1}^{N-1} Y_i\right] = \sum_{i,j=1}^{N-1} \mathrm{Cov}\left[Y_i, Y_j\right] = \sum_{i=1}^{N-1} \mathrm{Var}[Y_i] + 2\sum_{i<j} \mathrm{Cov}\left[Y_i, Y_j\right]. \tag{5}$$

Note that the third equality comes from the fact that $\mathrm{Cov}[Y_i, Y_i] = \mathrm{Var}[Y_i]$.

The sum of the variances is computed directly as

$$\begin{aligned}\sum_{i=1}^{N-1} \mathrm{Var}[Y_i] &= (N-1)[E[Y_i^2] - (E[Y_i])^2] = (N-1)\left(\frac{2mn}{N(N-1)} - \frac{4m^2n^2}{N^2(N-1)^2}\right) \\ &= \frac{2mn}{N} - \frac{4m^2n^2}{N^2(N-1)}.\end{aligned} \tag{6}$$

Note that, since Y_i can only take on values of 0 or 1, $E[Y_i^2] = E[Y_i]$.

Now consider

$$\mathrm{Cov}\left[Y_i, Y_j\right] = E\left[Y_i Y_j\right] - (E[Y_i])^2. \tag{7}$$

Observe that

$$E\left[Y_i Y_j\right] = P\left(Y_i Y_j = 1\right). \tag{8}$$

This probability depends on whether the ith and jth edges are adjacent or disjoint. If they are adjacent, the two edges are defined by three nodes and there are two possible label sequences for which $Y_i Y_j = 1$: STS or TST with probabilities respectively

$$P(STS) = \frac{mn(m-1)}{N(N-1)(N-2)}$$

and

$$P(TST) = \frac{mn(n-1)}{N(N-1)(N-2)},$$

so that

$$E\left[Y_iY_j|\text{adjacent}\right] = P(STS) + P(TST) = \frac{mn\big((m+n)-2\big)}{N(N-1)(N-2)} = \frac{mn}{N(N-1)}. \tag{9}$$

If Y_i and Y_j are disjoint, there are four nodes defining the two edges and four possible labelings that lead to $Y_iY_j = 1$: $(ST)(ST), (ST)(TS), (TS)(ST), (TS)(TS)$, each with the same probability, so that

$$E\left[Y_iY_j|\text{disjoint}\right] = 4\frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)}. \tag{10}$$

Let C be the total number of adjacent edge pairs in the MST. The total number of edge pairs is $\binom{N-1}{2}$. Combining this with $(2-10)$, we get (after some algebraic simplification):

$$\begin{aligned}
\mathrm{Var}[T_{\mathrm{FR}}|C] &= \sum_{i=1}^{N-1}\mathrm{Var}[Y_i] \\
&+ 2\left[\frac{C}{\binom{N-1}{2}}\sum_{i<j}\mathrm{Cov}\left[Y_i, Y_j|\text{adjacent}\right] + \left(1-\frac{C}{\binom{N-1}{2}}\right)\sum_{i<j}\mathrm{Cov}\left[Y_i, Y_j|\text{disjoint}\right]\right] \\
&= \sum_{i=1}^{N-1}\mathrm{Var}[Y_i] + 2\left[\frac{C}{\binom{N-1}{2}}\sum_{i<j}\left(E\left[Y_iY_j\middle|\text{adjacent}\right] - (E[Y_i])^2\right)\right. \\
&\qquad \left. + \left(1-\frac{C}{\binom{N-1}{2}}\right)\sum_{i<j}\left(E\left[Y_iY_j\middle|\text{disjoint}\right] - (E[Y_i])^2\right)\right] \\
&= \frac{2mn}{N(N-1)}\left\{\frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)}\left[N(N-1)-4mn+2\right]\right\}.
\end{aligned} \tag{11}$$

The value of C depends upon the topology of the MST. Specifically, it is determined by the node degrees. In the univariate ($dim = 1$) case, there are always two nodes of degree one and $N-2$ nodes of degree two. In this case, $C = N-2$ and (11) reduces to the Wald-Wolfowitz result.

In the general case ($dim > 1$), MSTs with a variety of node degree values are possible and the variance of R under H_0 depends on the underlying probability distributions. In order to make this a distribution-free test, we may condition on the observed MST of the pooled sample points. In other words, we may build an MST on $\mathcal{G}$, determine C (which is fixed for a particular MST), and then calculate the variance conditioned on C.

APPENDIX 3: DERIVATION OF THE MEAN AND VARIANCE OF T_{Ru} (Ru2014)

For the following discussion, we assume N is even. If N is odd, we create a pseudo-observation $N+1$, with $d(i, N+1) = 0$ for $i = 1, \ldots, N$. Then, we optimally match the $N+1$ observations, and discard the one pair containing the pseudo-observation. This results in a matching with $\frac{N-1}{2}$ pairs that minimizes the total distance between all matchings of the original N observations into $\frac{N-1}{2}$ pairs which discard one observation. Rather than having separate notation for even and odd N, we adopt a convention such that all of the notation always refers to the case of even N, perhaps after discarding one observation (for odd N).

To find the mean and variance of T_{Ru} under the null hypothesis (H_0), we proceed using the following notation. Let $\mathcal{G}$ be the complete undirected graph $(\mathbb{Z}_N, E_N)$ where the vertex set $\mathbb{Z}_N$ consists of the indices $1, 2, \ldots, N$ and the edge set consists of all $\binom{N}{2} = \frac{N(N-1)}{2}$ pairs of vertices. By convention, we write the pairs with smaller vertex first, so $E_N = \{(i,j): 1 \leq i < j \leq N\}$. Partition $\mathbb{Z}_N$ into two sets S ("Group 1") and T ("Group 2"), with $|S| = m$ and $|T| = n$, so $m + n = N$. Denote $E_N^{(S,T)}$ as the set of all edges with one vertex in S and the other in T. We refer to these edges as crossing edges. Let X_{ij} be the random variable that indicates whether the edge (i,j) is included in a minimum-weight r-regular subgraph, $\mathcal{G}_r^*$, with $1 \leq r \leq N-2$, where r is the node degree of each vertex. By the r-regularity of $\mathcal{G}_r^*$, for each $i \in \mathbb{Z}_N$ we have

$$r = \sum_{j=1}^{i-1} X_{ij} + \sum_{j=i+1}^{N} X_{ij}.$$

So,

$$r = E[r] = E\left[\sum_{j=1}^{i-1} X_{ij} + \sum_{j=i+1}^{N} X_{ij}\right] = \sum_{j=1}^{i-1} E[X_{ij}] + \sum_{j=i+1}^{N} E[X_{ij}],$$

since the expected value of the sum of random variables is equal to the sum of their individual expected values by the linearity of expectation, (1)

$$= \sum_{j=1}^{i-1} P(X_{ij} = 1) + \sum_{j=i+1}^{N} P(X_{ij} = 1),$$

since X_{ij} can only assume values of 0 or 1.

But under H_0, each edge is equally likely to be included in $\mathcal{G}_r^*$, so the summations above simplify to $r = (N-1)P(X_{12}-1)$. Therefore, for all $(i,j) \in E_N$,

$$E[X_{ij}] = P(X_{ij} = 1) = \frac{r}{N-1} \tag{2}$$

and

$$\begin{aligned} \mathrm{Var}[X_{ij}] &= P(X_{ij} = 1)P(X_{ij} = 0) = \big(P(X_{ij} = 1)\big)\big(1 - P(X_{ij} = 1)\big) \\ &= \left(\frac{r}{N-1}\right)\left(1 - \frac{r}{N-1}\right) = \frac{r(N-1-r)}{(N-1)^2}. \end{aligned} \tag{3}$$

The total cross-count, A_{Ru}, may be written

$$A_{\mathrm{Ru}} = \sum_{(i,j)\in E_N^{(S,T)}} X_{ij} \tag{4}$$

Let T_{Ru} be the total cross-count, scaled by the degree of the minimum-weight r-regular subgraph. Thus,

$$E[T_{\mathrm{Ru}}] = \frac{1}{r}E[A_{\mathrm{Ru}}] = \frac{1}{r}E\left[\sum_{(i,j)\in E_N^{(S,T)}} X_{ij}\right] = \frac{mn}{r}E[X_{ij}], \tag{5}$$

since there are mn total crossing edges in the complete undirected graph,

$$\left(\frac{mn}{r}\right)\left(\frac{r}{N-1}\right) = \frac{mn}{N-1}.$$

Finding the variance of T_{Ru} is slightly more complicated. For the X_{ij} random variables, we have that the variance of their sum is equal to the sum of their covariances. Thus,

$$\begin{aligned} \mathrm{Var}[A_{\mathrm{Ru}}] &= \mathrm{Var}\left[\sum_{(i,j)\in E_N^{(S,T)}} X_{ij}\right] = \sum_{(i,j)\in E_N^{(S,T)}} \sum_{(k,l)\in E_N^{(S,T)}} \mathrm{Cov}[X_{ij}, X_{kl}] \\ &= \sum_{(i,j)\in E_N^{(S,T)}} \mathrm{Var}[X_{ij}] + \sum_{\substack{(i,j),(k,l)\in E_N^{(S,T)} \\ (i,j)\neq(k,l)}} \mathrm{Cov}[X_{ij}, X_{kl}]. \end{aligned} \tag{6}$$

Note that the third equality comes from the fact that $\mathrm{Cov}[X_{ij}, X_{ij}] = \mathrm{Var}[X_{ij}]$.

By substituting (3), the sum of variances is determined to be

$$\sum_{(i,j)\in E_N^{(S,T)}} \mathrm{Var}\left[X_{ij}\right] = mn\mathrm{Var}\left[X_{ij}\right] = \frac{mnr(N-1-r)}{(N-1)^2}. \tag{7}$$

The sum of covariances may be broken up into terms that include pairs of adjacent edges and terms that include non-adjacent edges (*graphic for each term to be inserted*):

$$\sum_{\substack{i,k\in S\\ j,l\in T\\ (i,j)\neq(k,l)}} \mathrm{Cov}\left[X_{ij},X_{kl}\right] = \sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} \mathrm{Cov}\left[X_{ij},X_{il}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} \mathrm{Cov}\left[X_{ij},X_{kj}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j,l\in T\\ j\neq l}} \mathrm{Cov}\left[X_{ij},X_{kl}\right]. \tag{8}$$

For any two adjacent edges (i,j) and (k,l),

$$P\left(X_{ij}X_{kl}=1\right) = P\left(X_{kl}=1\middle|X_{ij}=1\right)P\left(X_{ij}=1\right) = \frac{(r-1)r}{(N-2)(N-1)} = E\left[X_{ij}X_{kl}\right]. \tag{9}$$

So,

$$\sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} \mathrm{Cov}\left[X_{ij},X_{il}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} \mathrm{Cov}\left[X_{ij},X_{kj}\right]$$

$$= \sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} E\left[\left(X_{ij}-E\left[X_{ij}\right]\right)\left(X_{il}-E\left[X_{il}\right]\right)\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} E\left[\left(X_{ij}-E\left[X_{ij}\right]\right)\left(X_{kj}-E\left[X_{kj}\right]\right)\right],$$

by using the linearity property of expectations, this can be simplified to the expected value of their product minus the product of their expected values, (10)

$$= \left(mn(n-1)+m(m-1)n\right)\left(E\left[X_{ij}X_{kl}\right]-E\left[X_{ij}\right]E\left[X_{kl}\right]\right)$$

$$= \left(mn(n-1)+m(m-1)n\right)\left(\frac{(r-1)r}{(N-2)(N-1)} - \left(\frac{r}{N-1}\right)^2\right)$$

$$= -\frac{mnr(N-1-r)}{(N-1)^2}.$$

For any two non-adjacent edges (i,j) and (k,l),

$$P\left(X_{ij}X_{kl}=1\right) = P\left(X_{kl}=1\middle|X_{ij}=1\right)P\left(X_{ij}=1\right) = \frac{(r(N-4)+2)}{(N-3)(N-2)}\frac{r}{(N-1)} = E\left[X_{ij}X_{kl}\right]. \tag{11}$$

So,

$$\sum_{\substack{i,k\in S\\ i\neq k\\ j,l\in T\\ j\neq l}} \mathrm{Cov}\left[X_{ij},X_{kl}\right] = \left(m(m-1)n(n-1)\right)\left(E\left[X_{ij}X_{kl}\right]-E\left[X_{ij}\right]E\left[X_{kl}\right]\right) \tag{12}$$

$$= m(m-1)n(n-1)\left(\frac{(r(N-4)+2)}{(N-3)(N-2)}\frac{r}{(N-1)} - \left(\frac{r}{N-1}\right)^2\right)$$

$$= \frac{2m(m-1)n(n-1)(N-1-r)r}{(N-3)(N-2)(N-1)^2}.$$

Combining (7), (10), and (12) yields

$$\mathrm{Var}[A_{\mathrm{Ru}}] = \sum_{\substack{i\in S\\ j\in T}} \mathrm{Var}\left[X_{ij}\right] + \sum_{\substack{i\in S\\ j,l\in T\\ j\neq l}} \mathrm{Cov}\left[X_{ij}, X_{il}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j\in T}} \mathrm{Cov}\left[X_{ij}, X_{kj}\right] + \sum_{\substack{i,k\in S\\ i\neq k\\ j,l\in T\\ j\neq l}} \mathrm{Cov}\left[X_{ij}, X_{kl}\right]$$

$$= \frac{mnr(N-1-r)}{(N-1)^2} - \frac{mnr(N-1-r)}{(N-1)^2} + \frac{2m(m-1)n(n-1)(N-1-r)r}{(N-3)(N-2)(N-1)^2} \tag{13}$$

$$= \frac{2m(m-1)n(n-1)(N-1-r)r}{(N-3)(N-2)(N-1)^2}.$$

Therefore, the variance of the scaled total cross-count may be expressed as

$$\mathrm{Var}[T_{\mathrm{Ru}}] = \frac{1}{r^2}\mathrm{Var}[A_{\mathrm{Ru}}] = \frac{2m(m-1)n(n-1)(N-1-r)}{r(N-3)(N-2)(N-1)^2}. \tag{14}$$

Note that, when $r = 1$, the first and second moment results in (5) and (14) match the results in Ro2005, as we would expect.

Simulation suggests that the null distribution of T_{Ru} is negatively skewed. Figure 32 shows a particular example of this negative skewness, using a histogram with an overlaid standard normal distribution curve and a Normal Q-Q plot. Similar behavior has been observed over a variety of other conditions. This behavior suggests that using a normal approximation might be inappropriate because the probability of getting a value of 0.05 in the normal approximation is actually greater than 0.05 (i.e. greater probability of making type I errors). Even so, simulations in Ru2014 suggest that for sufficiently large N and possible certain conditions on r, the null distribution of T_{Ru} is asymptotically normal, independent of distribution function F.

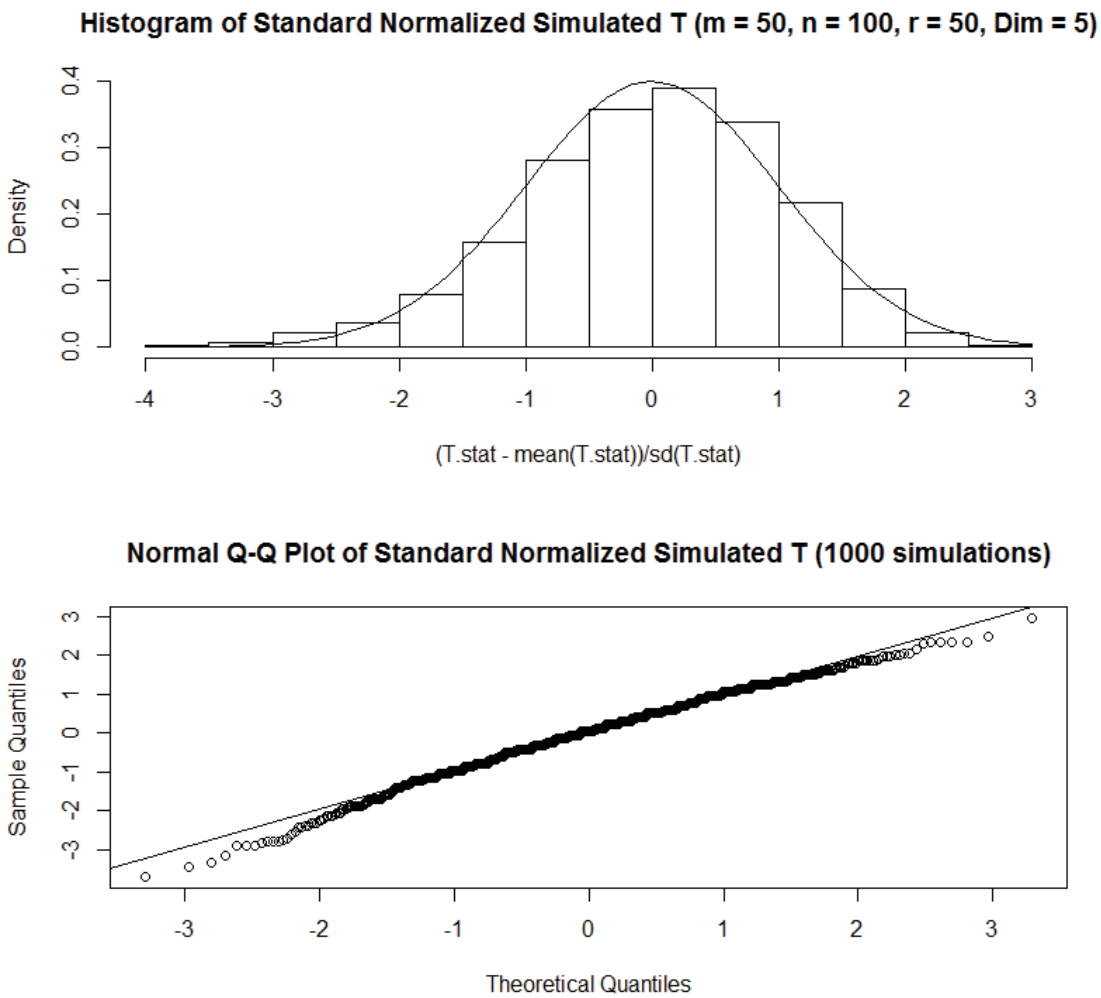


Figure 32: Histogram of standard normalized simulated T_{Ru} (with overlaid standard normal distribution curve) and Normal Q-Q plot. This suggests that the null distribution of T_{Ru} is negatively skewed.

APPENDIX 4: R CODE FOR THE CUMULATIVE CROSS-COUNT (CCC) TEST

```
# Performs the cumulative cross-count test.
# Returns p-value and other info
# Last update: 3/21/18

CCC.test = function(D, group.labels, lambda = c(1), nperms = 1000,
                    keep.perms = FALSE){
  N = attributes(D)$Size
  num.edges = choose(N,2)                          # total number of edges
  groups = unique(group.labels)                    # group identifiers
  num.groups = length(groups)                      # number of groups
  m = numeric(num.groups)
  in.group.jj = matrix(0,nrow = N, ncol = num.groups)
  in.group.jj.perm = in.group.jj

  # Find number in each group and build matrix whose columns are group
    indicators.
  # (This facilitates identifying cross-group edges.)
  for (ii in 1:num.groups) {
    m[ii] = sum(group.labels==groups[ii])
    in.group.jj[,ii] = (group.labels == groups[ii])
  }

  # Find cross-group edges:
  cross.mat = dist(in.group.jj)>0
  num.crossing.edges = sum(cross.mat)

  # Rank-order edges:
  o = order(D)

  # Compute cumulative cross-count and test statistic:
  cum.cross.count = cumsum(cross.mat[o])
  K = cum.cross.count - (num.crossing.edges/num.edges) *(1:num.edges)
  Tc.max = max(-K)

  # Now do permutation test:
  Tc.max.perm = numeric(nperms)
  if (keep.perms){
    Sk.mat = matrix(0,nrow = nperms, ncol = num.edges)
    K.mat = Sk.mat
    }
  for (jj in 1:nperms){
    perm.labels = sample(group.labels)
    for (ii in 1:num.groups) {
      in.group.jj.perm[,ii] = (perm.labels == groups[ii])
    }
    cross.mat.perm = dist(in.group.jj.perm)>0
```

```
    cum.cross.count.perm = cumsum(cross.mat.perm[o])
    K.perm = cum.cross.count.perm-num.crossing.edges/num.edges*(1:num.edges)
    Tc.max.perm[jj] = max(-K.perm)
    if (keep.perms){
      Sk.mat[jj,] = cum.cross.count.perm
      K.mat[jj,] = K.perm
    }
  }

  pvalc.max = mean(Tc.max.perm>=Tc.max)
  if (keep.perms){
    return(list(statistic = Tc.max, pvalc.max = pvalc.max, K = K, Sk.mat =
                Sk.mat, K.mat = K.mat, dists = D, cross.vec = cross.mat,
                edge.order = o, N = N, group.labels = group.labels,
                num.edges = num.edges, num.crossing.edges =
                num.crossing.edges))
  }

  else{
    return(list(statistic = Tc.max, pvalc.max = pvalc.max, Sk =
                cum.cross.count, K = K, dists = D, cross.vec = cross.mat,
                edge.order = o, N = N, group.labels = group.labels,
                num.edges = num.edges, num.crossing.edges =
                num.crossing.edges))
  }
}
```

LIST OF REFERENCES

Bhattacharya, B. (2017), “A General Asymptotic Framework for Distribution-Free Graph-Based Two-Sample Tests,” arXiv:1508.07530v4. arXiv.org.

Buttrey, S. and Whitaker, L. (2015), “treeClust: an R package for tree-based clustering dissimilarities” *The R Journal*, Vol. 7, No. 2, pp. 227-236.

Chen, H. and Friedman, J. (2017), “A New Graph-Based Two-Sample Test for Multivariate and Object Data,” *Journal of the American Statistical Association*, Vol. 112, No. 517, pp. 397-409.

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Friedman, J. and Rafsky, L. (1979), “Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests,” *The Annals of Statistics*, Vol. 7, No. 4, pp. 697–717.

Gordon, A. and Klebanov, L. (2010), “On a paradoxical property of the Kolmogorov-Smirnov two-sample test,” *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, Vol. 7, pp. 70-74.

Rosenbaum, P. (2005), “An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency,” *Journal of the Royal Statistical Society Series B*, Vol. 67, No. 4, pp. 515–530.

Ruth, D., and Koyak, R. (2011). “Nonparametric Tests for Homogeneity Based on Non-Bipartite Matching,” *Journal of the American Statistical Association*, Vol. 106, No. 496, pp. 1615-1625.

Ruth, D. (2014), “A new multivariate two-sample test using regular minimum-weight spanning subgraphs,” *Journal of Statistical Distributions and Applications*, Vol. 1, No. 22, pp. 1–12.

Weiss, M. (1978), “Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data,” *Journal of the American Statistical Association*, Vol. 73, No. 364, pp. 872-875.