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)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
05-09-2014		

4. TITLE AND SUBTITLE

Applications of Graph-Theoretic Tests to Online Change Detection

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Bodgan, Colin Edward

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. 424 (2014)

12. DISTRIBUTION / AVAILABILITY STATEMENT

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Detecting change in a stochastic process is a central problem in statistics. This project explores nonparametric graph-theoretic approaches to solving online change-point problems. The foundation for our methodology is the Ensemble Sum of Pair-Maxima (ESPM) Test, a powerful offline test developed by Ruth and Koyak (2011). Our work investigates the efficacy of the ESPM Test in a variety of offline settings, and ultimately extends that test to online settings through a novel modification of recently developed multiple testing procedures designed to control false discovery rate. When tested against simulated and pseudo real-world data, this modified procedure maintains the desired overall test level while achieving impressive power and useful advanced warning times in many scenarios. This method is not limited to the ESPM test and holds much promise for adapting other powerful offline techniques to online scenarios.

15. SUBJECT TERMS

Online change detection; Ensemble Sum of Pair-Maxima test (ESPM); Graph-theoretic procedure; Nonbipartite matching; Benjamini-Hochberg Procedure; False Discovery Rate

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		57	**19b. TELEPHONE NUMBER** *(include area code)*

U.S.N.A. --- Trident Scholar project report; no. 424 (2014)

APPLICATIONS OF GRAPH-THEORETIC TESTS TO ONLINE CHANGE DETECTION

by

Midshipman 1/c Colin E. Bogdan
United States Naval Academy
Annapolis, Maryland

_____
(signature)

Certification of Adviser Approval

CDR David M. Ruth, USN, Permanent Military Professor
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

Given a sequence of observations, has a change occurred in the underlying probability distribution with respect to observation order? How well can such a change be detected if the sequence is being monitored in real-time? The problem of detecting change, and detecting it with minimal delay, is an important one in a wide variety of real-world situations. For example, one might monitor a complicated multivariate system (such as a military helicopter or a human being) with the goal of detecting subtle change in order to provide advance warning of system failure.

Change-point problems may be classified as "online" or "offline." In offline problems, all data under consideration are on-hand at the time of analysis and the goal is to determine if, and perhaps when, a change occurred in the observation sequence. This leads to problems if a fatal result is encountered in the middle of the process from which data is being collected because it is too late for detection to do any good. In online problems, data are collected in real-time with the goal of identifying a change as soon as possible after it occurs and thus as far as possible in advance of death.

This project explores nonparametric graph-theoretic approaches to solving online change-point problems. The foundation for our methodology is the Ensemble Sum of Pair-Maxima (ESPM) Test, a powerful offline test developed by Ruth and Koyak (2011). Our work investigates the efficacy of the ESPM Test in a variety of offline settings, and ultimately extends that test to online settings through a novel modification of recently developed multiple testing procedures designed to control false discovery rate. When tested against simulated and pseudo real-world data, this modified procedure maintains the desired overall test level while achieving impressive power and useful advanced warning times in many scenarios. This method is not limited to the ESPM test and holds much promise for adapting other powerful offline techniques to online scenarios.

KEYWORDS: Online change detection; Ensemble Sum of Pair-Maxima test (ESPM); Graph-theoretic procedure; Nonbipartite matching; Benjamini-Hochberg Procedure; False Discovery Rate

ACKNOWLEDGEMENTS

Thank you to CDR Ruth for his unwavering faith and inspiring mentorship as my advisor on both the good and bad days. His patience and understanding are truly endless.

TABLE OF CONTENTS

I. INTRODUCTION

Life is all about making decisions. Many decisions are made as a result of observing change. A bus driver stops at a traffic light because it has turned from green to red; a stock broker decides to sell a majority of his positions due to a change in the markets; a child grabs a snack because he has become hungry; an alarm clock goes off to wake someone up because the correct amount of time has passed; a linguistics computer program awards credit to a student for correctly inflecting his voice. People (or things) that are better able to identify a change will be better equipped to make decisions. Humans are naturally good at recognizing changes limited in complexity which are of a large-scale and/or predictable like a traffic light. However in the real-world, clear-cut situations like this are not common. The real-world is intricate and at times very ambiguous. The ability of humans to discern subtle change across a complex system with many variables and do so in advance of a negative result is very difficult and imperfect. This project seeks to further develop this foresight in complex real-world systems.

The ability to detect change in a stochastic process is a central problem in statistics that has great practical importance. Consequently, robust solutions to this problem are highly sought after and widely used throughout the world. For example, consider these four real-world situations where change detection is used to make better decisions starting with a simplified one-variable situation:

- *Univariate Quality Control*. A tire factory wants to ensure that the tires it produces are of the highest quality and will not degrade too quickly. Quality control supervisors measure the tread depths of tires rolling off the production line and apply change detection methods in order to prevent tires from being made improperly. They hope to detect subtle, yet significant, changes that foreshadow imminent problems with tread depth production quality so they can be prevented.

- *Multivariate Quality Control*. Over time, the same tire factory wanting to improve upon its quality control method decides that tracking and analyzing tread depth measurements is insufficient to ensure quality tire production. Therefore, in addition to tread depth, features such as sidewall thickness, inner diameter, and aspect ratio are also measured. These four possibly-related numbers taken together may be compared to the measurements collected

from other tires produced. Now it may be possible to detect more subtle process deviations, based on individual values and relative values.

- *Multivariate Machine Health.* The Marine Corps wants to ensure its helicopters are being maintained and repaired before accidents occur. During operation, a helicopter vibrates due to its mechanical design and aviation maintenance experts determine that the best way to track helicopter health is through tracking its operating vibration frequency. Aircraft maintainers record frequencies with respect to time during flight at various locations throughout the aircraft. When an aircraft returns from a flight, the recorded multivariate data is analyzed for subtle evidence of health degradation to determine if repairs are necessary to prevent future mechanical failure.

- *Multivariate Biosurveillance*. Health officials want to anticipate (and possibly deter) disease outbreaks. These situations represent significant changes from the normal health of a population and so might seem easy to detect, but the complex world of human health is home to many subtle changes invisible to the naked eye. Doctors through the use of change detection methods are able to monitor many variables on public health at once and detect when subtle changes in certain measurements signal high risk of or even foretell an outbreak of contagious disease and vaccinate properly.

These few examples only provide a quick glimpse at a small portion of problems where change detection is important: others include image analysis, structural damage assessment, crime investigation, and environmental field analysis. Our work offers a new tool for change detection that can be employed in real-time in very general multivariate settings.

In the formal mathematical study of change detection, the problem of detecting a change in a stochastic process is known as a "change-point problem." Each observation, $\mathbf{X}$, of this process is said to be drawn from an underlying probability distribution, F. The change point, $\mathbf{X}_\varphi$, where $\varphi \in \{2, 3, \dots, i\}$ refers to the first point in a sequence of observations $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_i)$ at which the underlying probability distribution for that observation F_φ differs from that of prior observations$\left(F_1 = F_2 = \cdots = F_{\varphi-1} \neq F_\varphi\right)$.

Within change-point problems, one may differentiate between offline and online problems. In offline problems, testing for change is done only at a predetermined point when the system is turned off; no new data are collected as change-testing occurs. Such an approach is inadequate if a fatal change occurs in the middle of a process. In online problems, testing for

change occurs as data are collected. Generally, multiple tests for change are conducted in sequence that as much data about system performance as possible is considered for each test and change is detected in time for the appropriate decisions to be made.

The nature of such changes may be quite general. For instance, change could occur as a jump in data distribution mean at one point in time as seen for a univariate example in Figure 1.

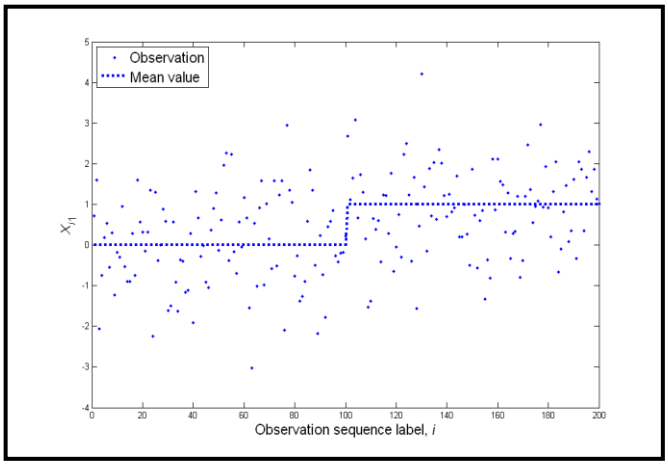

Figure 1: Jump in distribution mean at point 101 of 200

But, change may also occur as a gradual drift in data distribution mean, or change may occur as a jump or gradual drift in other parameters of the underlying probability distribution such as variance, scale, shape, et cetera as seen for another univariate example in Figure 2.

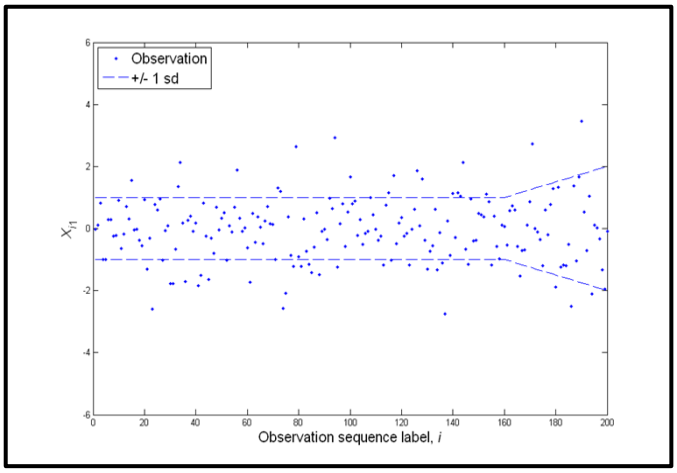

Figure 2: Drift in distribution variance at point 161 of 200

Simplistic in nature, the previous examples consider only a single variable. Figure 3 shows five dimensions of 200 sequential observations on some system. Prior to a change-point

at 101, the observations (plotted in black) follow a particular multivariate distribution; after the change point, the observations (plotted in red) follow a different multivariate distribution. This change is by no means obvious to the eye (in fact, a jump change in the distribution mean of magnitude 1 occurs at the change point); we seek a test that effectively detects such changes as soon as possible after they occur.

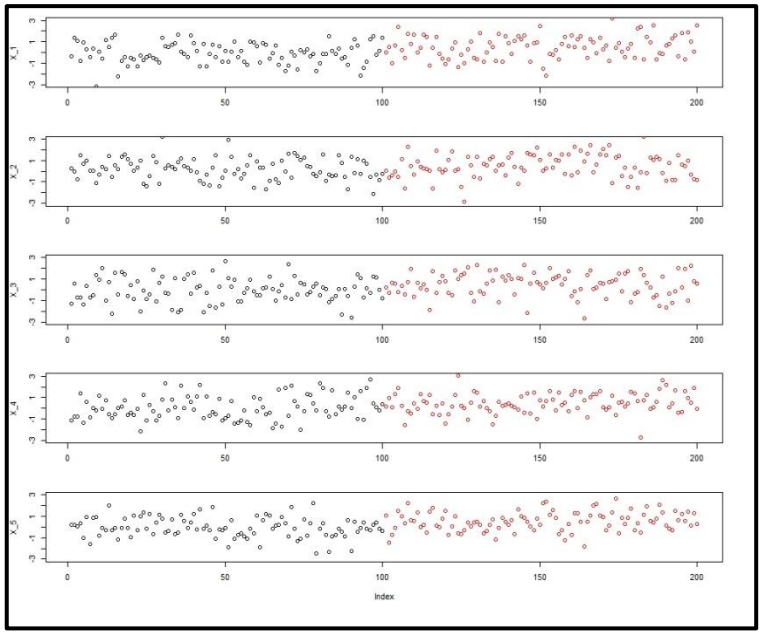

Figure 3: Dimensional breakdown of 200 observations of 5-variate data

The problem of change detection has a long history. In the relatively recent era of "big data," multivariate methods are of great interest. Most current approaches involve fairly strong assumptions about the distribution of the monitored observations. When the assumptions fail to be met, these approaches often suffer. Also, many existing techniques test only for a specific type of change (such as an abrupt jump in distribution mean), or require the change-point to be pre-determined. Such constraints limit the extent to which these test may be applied to real-world situations.

In this project, we specifically consider the problem of detecting change in the underlying distribution of a sequence of observations based on monitoring the observations online, with the dual objectives of: 1) identifying correctly that a change has occurred, and 2) maximizing the time between when a change is detected and the time when a negative result will occur (i.e. machine death through mechanical failure) or a positive chance squandered (not buying stock

early enough to take advantage of a rise in the stock market). Detected change is only useful if it is detected in time to be useful in a decision-making process.

We are interested in finding a way to detect change –perhaps subtle change of a quite general nature – within multivariate data without any *a priori* assumptions about underlying probability distributions: that is, ours is a *nonparametric* test. Our approach is graph-theoretic and relies upon the idea of *matching*, which involves pairing observations together based on interpoint distances. As a starting point in this project, we extend a recently developed offline nonparametric change detection test formulated by Ruth and Koyak (2011) for use in online situations.

Our work towards these ends is organized as follows: In Section II, we classify sequential change-point problems into two categories, offline and online, with discussion on how each type of problem relates to real-world scenarios and is framed. Because our new test is an extension of the Ensemble Sum of Pair-Maxima (ESPM) test of Ruth and Koyak (2011), we then explore the theoretical underpinnings of that through a review of relevant key literature and the main tenets of graph-theoretic matching. In Section III, we investigate the change detection capabilities of the ESPM Test in multivariate offline settings by varying both the amount of data available for testing and the location of the change within the data. This section culminates with an extension of the ESPM test to multivariate online situations using multiple testing theory and the Benjamini-Hochberg Procedure for controlling the False Discovery Rate. Section IV demonstrates the performance of the "telescope" multiple testing format among various change locations through a simulation study built to mimic online scenarios. In Section V, we apply our test to the (simulated) real-world data from the 2008 Prognostics and Health Management Challenge (PHM) and speak to the characteristics and challenges of real-world data sets in regards to graph-theoretic tests. Of central importance is the introduction of "horizons" to extend the Benjamini-Hochberg Procedure to online situations. Section VI summarizes all our findings and highlights opportunities for further work within the field and on the ESPM test in particular.

II. PROBLEM BACKGROUND

A. PROBLEM FORMULATION

Given a multivariate stochastic process can we detect departures from homogeneity in real time? In other terms, can we tell if the underlying probability distribution from which a sequence of multivariate observations is drawn has *changed*? Due to the widespread and unceasing nature of change, these same questions arise in many real-world applications and are commonly referred to as the "change-point problem" as previously stated.

1. Change Points

In the field of change detection, the term **change point** may be defined as follows: Given a sequence of random vectors $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N)\ \forall\ i \in \{1, \dots, N\}$, let D_i represent the probability distribution of $\mathbf{X}_i$. The **change point** $\varphi \in \{2, \dots, N\}$ is the first point in the sequence starting from the left where $D_i \neq D_{i-1}$. In our setting, once the initial change occurs, there is no return to the original distribution. For example, the distribution change at φ could be the result of an abrupt mean change, or "mean jump", where $D_1 = D_2 = \cdots = D_{\varphi-1} \neq D_\varphi = D_{\varphi+1} = \cdots = D_N$. It is also possible that D_i varies with i for $i \geq \varphi$: for instance, the distribution change beginning at φ could be a gradual mean change, or "mean drift", that is implemented incrementally and described as $\boldsymbol{\mu}_{D_{i\geq\varphi}} = \boldsymbol{\mu}_{D_{\varphi-1}} + \mathbf{z}(i - \varphi + 1)$ where $\boldsymbol{\mu}_D$ is the average value of D_1 and each component of $\mathbf{z}$ is the rate at which the associated mean component changes. More complex forms for $D_{i\geq\varphi}$, like those intrinsic to real-world scenarios, are allowed as well.

In a hypothesis-testing context, the general change-point problem with respect to observations $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$ involves defining the null hypothesis

(2.1) $$H_0: D_1 = D_2 = \cdots = D_N$$

against the alternative hypothesis

(2.2) $$H_1: \quad D_1 = D_2 = \cdots = D_{\varphi-1} \neq D_\varphi$$

and $D_{j\geq\varphi} \neq D_1\ \exists\ \varphi \in \{2, \dots, N\}$ and $\forall\ j \in \{\varphi, \dots, N\}$.

As a result of its general nature, this alternative hypothesis fails to be inclusive of every possible change situation. Its framework is suitable for many change situations, but does not cover brief departures from homogeneity such as:

$$H_1: \quad D_1 = D_2 = \cdots = D_{\varphi-1} = D_{\varphi+1} = \cdots = D_N \neq D_\varphi,$$

or periodic departures from homogeneity that cause there to be a cyclic pattern of change away from the original distribution for an interval of time and then change back to the original distribution for an interval of time. These scenarios can develop when performing image analysis to detect short-lived change like a car passing through an area or a brief hand command to a robot. Our new test is not designed to find change in these situations. It is designed to find change in situations where prior to the change point φ all observations come from the same initial distribution and then once at the change point all observations come from distributions different from the initial one.

2. Dichotomy of Change-Point Problems

Various possible taxonomies can be used when studying change-point problems. Our work naturally separates change-point problems into two main categories: *online* and *offline*. This system centers on the difference in how new observations are incorporated into testing. For *online* problems, data collection and testing occur concurrently. As new observations are added to the data set, the null hypotheses are tested repeatedly in the following manner:

1) With $N-1$ observations $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{N-1})$ collected and available, add $\mathbf{X}_N$.
2) Test for a change-point $\varphi \in \{2, \ldots, N\}$
    a. If a change-point is detected, then take appropriate action.
    b. If not, then return to step (1) with $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$ available.

Classic cases of online problems include the monitoring of operating mechanical systems, where system measurements believed to indicate system health are sampled while the system is operating and tested in sequence to identify if change has occurred in the process. The goal is to detect change in time to be useful in preventing a negative result (i.e. machine death) or a positive opportunity from being squandered (i.e. bullish stock market). Because there may be no limit to how large the set of observations will be if and when a change-point is detected though,

the set of observations has the potential to become prohibitively large and therefore difficult to work with depending on the specific scenario.

Alternatively in o*ffline* problems, all data are collected in advance of any testing; that is, testing is conducted upon a finite sequence of observations whose length is known prior to testing. Observations are not collected during the testing period and vice versa. *Offline* problems can be further delineated into cases of Two Sample Tests or Simultaneous Tests. For a Two Sample Test, the change-point $\varphi \in \{2, \dots, N\}$ tested for in the sequence of multivariate observations $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N)$ is known or assumed. The sequence of observations is then split into two samples at this predetermined point, $\{\mathbf{X}_1, \dots, \mathbf{X}_{\varphi-1}\}$ and $\{\mathbf{X}_\varphi, \dots, \mathbf{X}_N\}$, and then the analyst tests the null hypothesis $H_0: D_1 = D_2 = \cdots = D_N$ against the alternative hypothesis $H_1: D_1 = D_2 = \cdots = D_{\varphi-1} \neq D_\varphi = D_{\varphi+1} = \cdots = D_N$. A well-known Two Sample Test is the univariate Kolmogorov-Smirnov Test for Distributional Homogeneity, though it is not normally associated with change-point problems. This nonparametric test regularly used by scientists examines the maximum difference between the empirical distribution functions of associated data sets in order to determine if they are both drawn from the same distribution. An example of this problem type is seen in clinical trials where two groups of subjects are drawn from the general population. The first is designated a control group and given a placebo, while the other is designated a test group and administered a treatment. The groups are then studied to determine whether or not the treatment has had some kind of effect. As a whole, these tests are not well-built for real-world use beyond scenarios where the change point is easily assumed. If the predetermined change-point chosen by the analyst unknowingly differs from the unknown actual change-point by a significant amount or by a couple crucial data points, the test could easily misidentify change when there is none or fail to detect when there actually is. In real-world situations, the actual change-point is often nearly impossible to identify. The need to make assumptions can cause the effectiveness of this test to suffer heavily. As a consequence, an extremely attractive feature of the ESPM test is its nonparametric nature. It does not require any limiting assumptions to be made concerning the underlying probability distribution of the data or the predetermination of a change point for it to be applied. This universality cements it as a strong foundation for solution approaches to numerous types of change-point problems including ours.

For simultaneous tests, there is no predetermined change-point φ. Instead, the test is performed upon the entire sequence of observations $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$ available; thus, all possible change-points are tested simultaneously. The null (2.1) and alternative (2.2) hypotheses tested in this problem are as stated in the previous subsection. From a utility standpoint, the absence of an assumption concerning change-point location makes these tests perfect candidates for application to such real-world situations, where change points are not known in advance. One example of such a test is environmental image analysis. The abuse natural vegetation sustains as a result of training activities is a significant issue on many military installations. Consider a large tract of land on a military base monitored a by satellite specially designed for ground-level imagery where each image taken of the land represents a multivariate observation and each pixel within an image represents a variable. Natural resource managers are unable to measure key statistics like land cover physically so they rely on innovative environmental management and monitoring tools to identify areas of land cover change through the use of satellite imagery during their periodic check.

As mentioned previously in Section I, the goal of this research is to extend the ESPM test developed by Ruth and Koyak (2011), which is offline, nonparametric, and simultaneous, for use in online problems. Consequently, our research has focused on conducting successful simultaneous testing in online situations. A survey of literature in the field of change detection reveals that there are few powerful nonparametric simultaneous tests for multivariate change-point problems, particularly online problems. We now proceed to review various graph-theoretic approaches to change detection concluding with a discussion about how these approaches comprise the theoretical origin of the ESPM test. This will lay the foundation for the new online extension of the ESPM test in the next chapter.

B. GRAPH-THEORETIC APPROACHES TO CHANGE DETECTION

Graph-theoretic ideas provide an innovative approach to change-point problems. Of late, methods using graph-theoretic ideas such as minimum spanning trees, nearest-neighbor algorithms, and clustering methods have proved interesting to many due to advances in computational capacity which makes them realistic to implement. The ESPM test is itself based

heavily upon *matching*. To fully understand what *matching* is, one must know something about the basic definitions and background of graph theory, so that is where we begin.

1. Graph Theory

These definitions come from Chartrand and Zhang (2005). A **graph** is an ordered pair $G = (V, E)$ consisting of a finite nonempty set of vertices V connected by edges e, which are two-element unordered subsets of V. A graph $G_1 = (V_1, E_1)$ is called a **subgraph** of $G = (V, E)$ if $V_1 \in V$ and $E_1 \in E$; if $V_1 = V$, then G_1 is a **spanning subgraph** of G. Two distinct vertices, v_1 and v_2, are **adjacent vertices** if they are joined by an edge $\{v_1, v_2\}$. Two distinct edges are **adjacent edges** if they share a vertex such as $\{v_1, v_2\}$ and $\{v_2, v_3\}$. A **complete graph** is a graph in which all vertices are adjacent. An **undirected** graph is one in which the edges have no orientation- that is, edge $\{v_1, v_2\}$ is identical to $\{v_2, v_1\}$; a **directed** graph is one in which edges have a direction associated with them making edge $\{v_1, v_2\}$ distinct from edge $\{v_2, v_1\}$. Vertex v_1 and edge $\{v_1, v_2\}$ can be referred to as **incident** with each other, and the **degree** of vertex v_1 is the number of edges incident with v_1.

A $u - v$ **walk** in graph G is a sequence of vertices in G beginning with u and ending with v such that consecutive vertices within the sequence are adjacent; if $u = v$, then the walk is **closed**. A walk in which no edge is used more than once is called a $u - v$ **trail.** A **circuit** is a closed trail that includes at least three distinct vertices; a circuit that repeats no vertex except the first and last is a **cycle**. If there is a $u - v$ **walk** for every pair of vertices in graph G, then G is said to be **connected**.

A graph G is called **acyclic** if it has no cycles and is a **tree** if it is both acyclic and connected. A **spanning tree** of G is a spanning subgraph that is also a tree. If a real number expressing some form of interpoint cost is assigned to each edge in G, then G becomes a

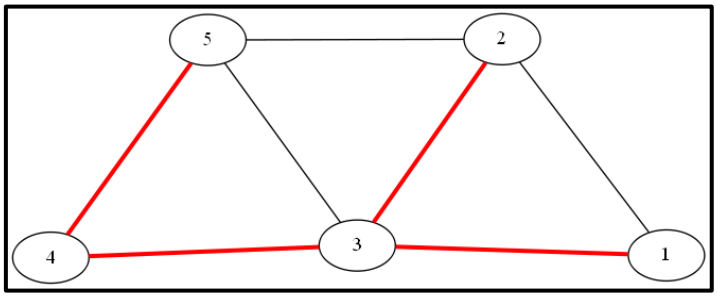

Figure 4: Undirected complete graph on 5 vertices; subgraph highlighted in red is a spanning tree

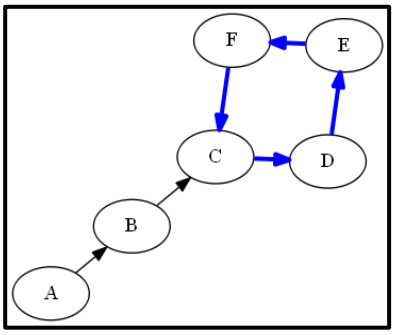

Figure 5: Directed graph with cycle (C,D,E,F) highlighted in blue

weighted graph and the sum of all edge weights known as the **weight** of the graph. The spanning tree of weighted graph G whose weight is the least among all possible spanning trees is the **minimum spanning tree** (MST) of G. Figures 4 and 5 illustrate some of these terms. From this point forward, every graph discussed is a complete undirected graph.

In a pioneering paper, Friedman and Rafsky (1979) considered various change-detection test statistics based on the relational information contained within MSTs in order to test if two samples came from the same distribution. They begin with two sets of observations, S_1 and S_2, and define $V = S_1 \cup S_2$. Next, an MST is constructed with respect to some interpoint cost function on V. They then remove each edge in the MST which connects a point in S_1 to a point in S_2, and count the number of disjoint trees created as a result of edge removal. This count tends to be lower when S_1 and S_2 come from different distributions, although this test is not particularly powerful.

2. Matching

To begin our review of matching-based approaches to change detection, we present a few additional definitions to help develop later ideas. A subset of edges $E_1 \in E$ is **independent** if no two edges in E_1 are adjacent. A **matching** in a graph $G = (V, E)$ is an independent set of edges in G. The amount of possible matches in a graph depends upon its size. A **maximum matching** in G is a matching that has at least as many edges as any other potential matching in G. Since the ESPM test utilizes maximum matchings and our new test is an extension of it, all matchings discussed from this point forward in the paper are maximum matchings. A **perfect matching** in G is a matching that includes every vertex in G. Perfect matchings are inherently maximum matchings, although not vice versa; additionally, they are only possible if a graph has an even number of vertices.

Matching-based approaches have been used in a variety of problem areas such as organ donation and large-scale logistics. These methodologies commonly seek to find a matching which minimizes some contextually relevant interpoint cost function. In these problems, two main types of matching occur: **bipartite**, where the vertices of the graph are split into two unique subsets of observations, S_1 and S_2, and each edge of the graph consists of a vertex from each subset (a vertex can only be paired with a vertex from the other subset), and **nonbipartite**, where matching does not depend on any previous partitioning of the vertices (a vertex can be paired with any vertex other than itself). Each type of matching has a different effect upon which edges make up the minimum-weight matching.

For this research, we take interest in minimum-weight, or minimum-cost, nonbipartite matchings (MNBMs). The general interpoint cost function used in change-point problems is defined by Ruth and Koyak (2011) as follows: Given sample space S, $c: S \times S \rightarrow [0, \infty)$ is a cost function if it satisfies

(2.3) $$c(x, x) = 0 \; \forall \; x \in S$$

and

(2.4) $$c(x, y) = c(y, x) \; \forall \; x, y \in S.$$

Let c_{ij} denote the cost $c(\mathbf{X}_i, \mathbf{X}_j)$, or in more general terminology the weight of the edge connecting $\mathbf{X}_i$ and $\mathbf{X}_j$. The function c will be referred to as a distance function if it satisfies the triangle inequality

(2.5) $$c(x, z) \leq c(x, y) + c(y, z) \; \forall \; x, y, z \in S$$

in addition to (2.3) and (2.4). This general definition gives the flexibility needed to accommodate all types of data (discrete and continuous, univariate and multivariate, etc.). However, this does not mean there is no need for adjustment. Interpoint cost functions should generally be tailored to the context of the problem and the pertinent data types. In nonbipartite minimum matching, the assignment of pairs relies entirely upon the choice of cost function. Context might clearly point towards one measure of cost over others; it might require the choice to become a matter of debate. Commonly used cost functions include computing interpoint Euclidean distance or

(2.6) $$d_{ij}^{ED} = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' (\mathbf{X}_i - \mathbf{X}_j)},$$

Mahalonobis distance, in order to account for measurement scale and correlation between data components

(2.7) $$d_{ij}^{MD} = \sqrt{\left(\mathbf{X}_i - \mathbf{X}_j\right)^T S^{-1} \left(\mathbf{X}_i - \mathbf{X}_j\right)}$$

where S is the covariance matrix, and Manhattan distance, but there are many others.

Rosenbaum (2005) developed a test statistic derived from MNBMs to compare two multivariate distributions. Using simple interpoint distances to construct a MNBM of multivariate observations, his cross-match statistic is the number of pairs comprised of one observation from the first distribution and one observation from the second distribution. Distributions that are very different will cause few cross-matches, whereas distributions that are very similar will cause many cross-matches. Impressively, his cross-match statistic has a known exact distribution and is nonparametric. This test strongly motivated Ruth and Koyak (2011) in their development of the ESPM test.

In other related work, Lu *et al.* (2001) demonstrate the ability to use optimal nonbipartite matching in a multivariate real-world scenario and achieve solid analysis of the problem in question. Through an observational study on the media campaign against drug use, optimal non-bipartite matching was used to pair teenage subjects who were demographically similar but had extremely different levels of exposure to the media campaign. The stated intentions of the subjects in relation to illegal drug use were used to assess the effectiveness of the campaign.

Another successful application of optimal nonbipartite matching appears in Lu and Rosenbaum (2004) where they investigate if a localized minimum-wage increase is at all associated with depressed low-wage employment rates in the same area. Faced with the problem of needing to compare one test group to two control groups, they convert the natural tripartite problem into a nonbipartite matching problem and provide relevant analysis on the topic in question.

In the vein of minimizing global interpoint cost like Friedman and Rafsky's (1979) MST test and Rosenbaum's (2005) cross-match test, the cornerstone of the ESPM test is the computation of a MNBM to pair observations so as to minimize the sum of all distances between paired observations. Unlike the others though, the ESPM test computes many MNBMs. We elaborate in the following section.

3. Ensembles

Intuition suggests that if a single MNBM contains information about distributional homogeneity in a sequence of observations, then additional information might be contained in additional matchings where each successive matching is the next-best matching independent of the first. This idea has been thoroughly explored by many people and found to be true. More specifically, the power of graph-theoretic tests is known to be enhanced by considering collections of orthogonal subgraphs called *ensembles*. Friedman and Rafsky (1979) originally suggest that ensembles of MSTs with $k \ll N/2$ orthogonal matchings be used to refine the sensitivity of their multivariate runs test where N is the number of observations in the data set. In examining that same test, Friedman and Rafsky (1979) show it has higher power when used in higher dimensions, but more importantly augments general test power by computing their test statistic on an ensemble of orthogonal MSTs, where two MSTs are orthogonal if they do not share any edges in common.

Although the ESPM test makes use of a different type of subgraph (a matching), Ruth and Koyak (2011) make use of a similar idea in order to increase test power. To describe their adaptation, they define the term orthogonally successive optimal matchings (OSOMs) to refer to matchings constructed through the following process: compute an optimal nonbipartite matching, then find the next best matching that is orthogonal to the first, then the next best matching that is orthogonal to both the first and second, and so on until there are no more orthogonal matches left to be made. Figure 6 below displays OSOMs on $N = 6$ points. Each color represents a different

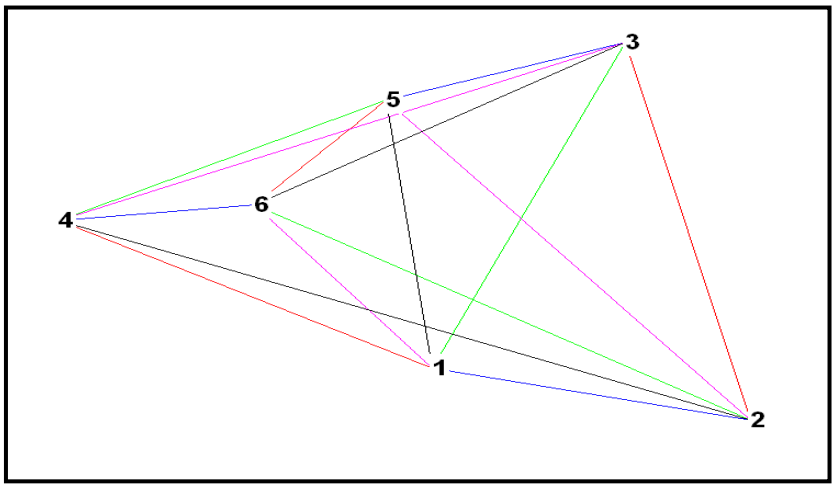

Figure 6: Orthogonal, successive, optimal, non-bipartite, maximum matchings on $N = 6$ points.

matching where the best is blue, then red, green, purple, and black. Orthogonality is shown in the graph by how no edge is reused by two different colors. Given $N = 2n$ observations and an associated interpoint cost function, the OSOM procedure guarantees that at least $N/2$ matchings may be obtained from the data set.

At its core, the ESPM test centers on the idea that MNBMs (more than one) based on interpoint distances will tend to result in pairings that are closer in sequence order than would be the case if all observations came from the same distribution. Until now though, the application of this idea has been limited to offline problems. We now propose a new test that is simultaneous, multivariate, and distribution-free which will extend the methodology of the ESPM test for use in online situations.

III. TEST DISCUSSION

We introduce a new change detection test for use in online change-point problems. As stated previously, this new test builds off the methodology of the powerful, nonparametric, multivariate, offline Ensemble Sum of Pair-Maxima (ESPM) test created by Ruth and Koyak (2011). Our online extension applies Multiple Testing Theory through a novel extension of the False Discovery Rate procedure of Benjamini-Hochberg (1995).

For the purpose of presenting our test, assume a sequence of $N = 2n \geq 4$ observations ordered with respect to time. Our goal is to test if change occurs with respect to this ordering. For instance, change may be a jump or a drift in some distributional parameter at some unknown point in the sequence. The requirement that N be even is not strict, but it simplifies description of the test and we later describe how to handle odd N accordingly. For a simultaneous test where F_i is the underlying probability distribution of the i^{th} observation ($\mathbf{X}_i$) within this sequence, the null hypothesis of distributional homogeneity asserts that $F_1 = F_2 = \cdots = F_N$. The alternative hypothesis asserts that there exists a change point $\varphi \in \{2, \dots, N\}$ such that $F_1 = F_2 = \cdots = F_{\varphi-1} \neq F_\varphi$. As in the case of the ESPM test, we compute a minimum nonbipartite matching of n pairs, $M = \left\{\{\mathbf{X}_{i_1}, \mathbf{X}_{i_2}\}, \{\mathbf{X}_{i_3}, \mathbf{X}_{i_4}\}, \dots, \{\mathbf{X}_{i_{N-1}}, \mathbf{X}_{i_N}\}\right\}$, with respect to some cost function. The ordering of these pairs is arbitrary. We use Euclidean distance (2.6) unless otherwise specified. As in Ruth and Koyak (2011), let $R_{i_{2q-1}} = i_{2q-1}$ and $R_{i_{2q}} = i_{2q}$ be the sequence labels for the q^{th} pair $\left\{\mathbf{X}_{i_{2q-1}}, \mathbf{X}_{i_{2q}}\right\}$. For example, if the second pair in M is $\{\mathbf{X}_7, \mathbf{X}_3\}$, then $q = 2$ and so $R_{i_{2q-1}} = R_{i_3} = i_3 = 7$ and $R_{i_{2q}} = R_{i_4} = i_4 = 3$. Now order each individual pair as (U_q, Y_q), where U_q and Y_q are correspondingly the minimum and maximum value of the ordering variable:

$$U_q = \min\left\{R_{i_{2q-1}}, R_{i_{2q}}\right\} \text{ and } Y_q = \max\left\{R_{i_{2q-1}}, R_{i_{2q}}\right\}, \text{ where } q \in \{1, 2, \dots, n\}. \tag{3.1}$$

Continuing the $q = 2$ example, $U_2 = 3$ and $Y_2 = 7$. These ideas are illustrated in the context of successive matching in Figure 7 and Table 1 for a case with N = 6 points. With this foundation in place, we are now ready to discuss the test statistics which we will employ for sequential testing.

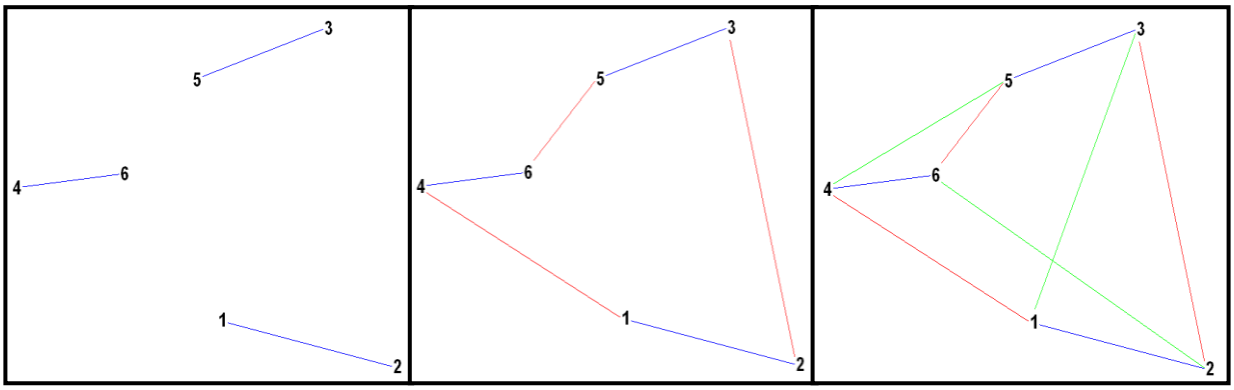

Figure 7: Three successive minimum nonbipartite matchings on $N = 6$ points.

Table 1: Maximum and minimum values of pairs from matchings shown in Figure 7

First Matching (Blue)			Second Matching (Red)			Third Matching (Green)		
Pair	U_q, Minimum	Y_q, Maximum	Pair	U_q, Minimum	Y_q, Maximum	Pair	U_q, Minimum	Y_q, Maximum
(1,2)	1	2	(1,4)	1	4	(1,3)	1	3
(3,5)	3	5	(2,3)	2	3	(2,6)	2	6
(4,6)	4	6	(5,6)	5	6	(4,5)	4	5

A. ENSEMBLE SUM OF PAIR-MAXIMA TEST

The foremost job of any change detection test is to properly detect change regardless of other concerns. Thus, the decision of which test statistic to use as the determining factor of whether or not change has occurred is of paramount importance. A desirable test statistic is proportionally sensitive to the occurrence, or lack thereof, of change; bad ones are completely insensitive or overly sensitive to the change attempting to be detected. Our new test is powered by the three-level test statistic used in the ESPM test where each proceeding level is used to help determine the current level's numeric value.

1. The Sum of Pair-Maxima, T_N

If the alternative hypothesis is true – that is some form of distributional change has occurred within a sequence of observations – it is expected that a minimum nonbipartite matching based on interpoint Euclidean distances would pair observations that are closer together in sequence than would be the case under the null hypothesis of no change. Drawing upon this, Ruth and Koyak (2011) devise a test statistic based on summing the differences

$Y_i - U_i$ in each pair and rejecting the null hypothesis if the sum is less than some critical value. They consider, T_N, an equivalent test statistic labeled as the *Sum of Pair-Maxima* (SPM) test statistic built upon this sum:

(3.2) $$T_N = \sum_{i=1}^{n} Y_i.$$

This is equivalent to the sum of sequence label differences seen by the identity:

(3.3) $$\sum_{i=1}^{n} Y_i = \frac{1}{2}n(2n+1) + \frac{1}{2}\sum_{i=1}^{n}(Y_i - U_i).$$

The null hypothesis is rejected for small values of this sum.

The mean and variance of T_N are derived using Fristedt and Gray's (2004) definition for exchangeable sequences of random variables. A sequence $(Y_1, Y_2, \dots, Y_j)$ of random variables is exchangeable if for any permutation β of indices $\{1, 2, \dots, j\}$ the joint probability distributions of $(Y_1, Y_2, \dots, Y_j)$ and $\left(Y_{\beta_{(1)}}, Y_{\beta_{(2)}}, \dots, Y_{\beta_{(j)}}\right)$. This definition is easily fit by the case of the null hypothesis since the ordering among the pairs in the matching is arbitrary. There are $N!$ possible permutations of sequence labels, all of which are equally probable, and the random variables $(Y_1, Y_2, \dots, Y_n)$ are exchangeable. This leads to the mean of T_N being expressed as:

(3.4) $$\mu_N = E[T_N] = n * E[Y_i] = \frac{N(N+1)}{3}$$

and the variance as:

(3.5) $$\sigma_N^2 = \mathrm{Var}[T_N] = n * \mathrm{Var}[Y_1] + n(n-1)\mathrm{Cov}[Y_1, Y_2] = \frac{N(N-1)(N+1)}{180}.$$

Theorem 1 of Ruth and Koyak (2011) shows that T_N has a limiting normal distribution:

(3.6) $$P\left(\frac{T_N - \mu_N}{\sigma_N} \leq t\right) \rightarrow \Phi(t) \text{ as } N \rightarrow \infty$$

where t represents all real numbers and Φ is the standard normal cumulative distribution function. It is important to note though that this is not done through the classical central limit theorems. Despite each Y_i in T_N having identical marginal distributions because they are exchangeable, they are not independent; therefore (3.6) cannot be proved using classical central limit theorems. To overcome this, (3.6) is proved via a technique known as *Stein's method*. Stein's method (Stein, 1972, 1986) relies on a differential equation that describes the normal distribution and a process known as "coupling," by which auxiliary random variables similar to

the variables being studied are constructed. The results establish bounds on the distance from normality for selected cases of dependency, including this case. Implementation of the method and proof of asymptomatic normality for T_N is seen in Section III of Ruth (2009).

The normal approximation for T_N is improved using an Edgeworth expansion to diminish error present in cases where n is small or moderately sized. This procedure approximates the distribution of interest by starting with a normal distribution and then adding in higher order corrections for non-zero moments of the third order or above. In Ruth (2009), the third central moment of T_N is shown to be:

$$E[(T_N - \mu_N)^3] = E[T_N^3] - 3\mu_N\sigma_N^2 - \mu_N^3 = \frac{-n(n-1)(2n+1)(2n+3)}{945} = \kappa_3. \tag{3.7}$$

For all $n > 1$, $\kappa_3 < 0$. Thus, it follows that T_N is negatively skewed for all cases of interest. The Edgeworth approximation is then given by:

$$P\left(\frac{T_N-\mu_N}{\sigma_N} \leq t\right) \approx \Phi(t) + c_0 \frac{N+3}{N\sqrt{(N-2)(N+1)}}(t^2-1)e^{-t^2/2} \tag{3.8}$$

where $c_0 = \frac{\sqrt{45}}{126\sqrt{2\pi}} \approx 0.06$. A detailed derivation is found in Section III of Ruth (2009).

Many theoretical properties of interest for a test statistic have to do with its properties under the null hypothesis. In contrast, a test statistic is said to be *consistent* if it has the property that its power approaches 1 as sample size increases without limit for any level of significance and any departure from the null hypothesis (no matter how small). While the null properties of the tests such as SPM may be straightforward to establish, consistency is often difficult to prove. The consistency of the generalized runs test of Friedman and Rafsky (1979) that used minimum spanning trees was only proven by Henze and Penrose twenty years after the test was introduced. Rosenbaum's cross-match test has only been proven consistent under less general conditions. Without specification of a change point, Ruth (2009) showed that the SPM test is consistent against general jump alternatives under the same conditions for which the cross-match test is consistent. Simulations conducted during our research suggest that the SPM statistic is consistent across drift and other alternatives as well.

We stated earlier that the requirement for N to be even was not strict. If the sample size is odd, nonbipartite matching will leave one observation unpaired. In order to overcome this problem without leaving out the last observation, Ruth and Koyak (2011) create a dummy

observation which is assigned a distance zero from every other observation; this ensures that every observation will be part of a matching pair. These situations do not affect the asymptomatic normality discussed earlier, and the mean and variance are given by:

(3.9) $$\mu_N = E[T_N] = \frac{(N-1)(N+1)}{3},\ \sigma_N^2 = \mathrm{Var}[T_N] = \frac{N(N-1)(N+1)}{180}.$$

We consider a short graphical example of the T_N test statistic for illustration purposes. Figure 8 below shows a randomly generated sample of N = 20 bivariate observations which are plotted by their sequence numbers and then summarized further below in Table 2. By design, we drew these observations from unequal distributions. From the SPM test, we calculate $T_N = 122$. The null hypothesis tells us that the expected value and standard deviation of T_N, given by (3.4) and (3.5), are $\mu_N = \frac{(20)(21)}{3} = 140$, $\sigma_N = \sqrt{\frac{(10)(9)(21)}{45}} = 6.48$. For an $\alpha = 0.05$ level test, the critical value stemming from (3.8) is 129. $T_N = 122 < 129$ suggesting that the underlying probability distributions of the sequential observations displayed in Figure 8 have undergone some sort of change, as indeed is the case.

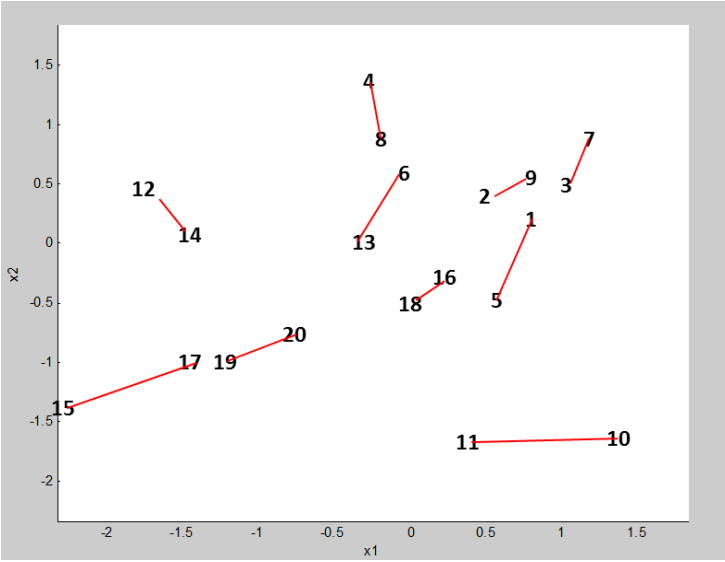

Figure 8: A minimum nonbipartite matching on N = 20 bivariate observations.

Table 2: SPM test calculations for data shown in Figure 5

Pair	Y_i, Maximum
(1,5)	5
(2,9)	9
(3,7)	7
(4,8)	8
(6,13)	13
(10,11)	11
(12,14)	14
(15,17)	17
(16,18)	18
(19,20)	20
$T_N = \sum_{i=1}^{n} Y_i = 122$	

2. Cumulative Sum of Pair-Maxima, $S_{N,k}$

The power of the SPM test is significantly increased through the use of a collection of orthogonal matchings, or "ensemble". This relies upon the idea that if a single matching provides information about distributional homogeneity in a sequence of observations, then subsequent optimal pairwise orthogonal matchings on the same observations will provide additional information. A *k-ensemble* of matchings is considered to be complete if $k = N - 1$ orthogonal matchings of the data are possible. Complete ensembles have many practical applications such as being solutions to round robin scheduling problems for a sports tournament.

Ruth and Koyak (2011) construct these ensembles recursively; that is they compute the first MNBM on the data, then compute the next best MNBM on the data that is pairwise orthogonal to the first, and so on until no more MNBMs are possible. This procedure can often fail to produce a complete ensemble of MNBMs, but Anderson (1972) shows that it will always yield at least a half ensemble ($k = N/2$), which is sufficient for this test.

Each of these matchings has a sum of pair-maxima statistic. Let $T_{N,i}$ signify the sum of pair-maxima statistic for the i^{th} best successive orthogonal matching. The n-ensemble version of the SPM test is then expressed as:

(3.10) $$S_{N,k} = \sum_{i=1}^{n} T_{N,i}.$$

We define $S_{N,k}$ as the cumulative sum of pair-maxima over the first k matchings. Just as T_N takes on lower values under the alternative hypothesis than under the null hypothesis, $S_{N,k}$ is also expected to drop below its mean value when under alternative hypotheses. Theorem 2 of Ruth and Koyak (2011) shows that each $T_{N,i}$ has identical univariate marginal distributions and moments up to the second order when under the null hypothesis of homogeneity. It is natural then that:

(3.11) $$\mu_{S_{N,k}} = E\left[S_{N,k}\right] = \sum_{i=1}^{n} E\left[T_{N,i}\right]. = k * \mu_N = k * \frac{N(N+1)}{3},$$

and

$$\mathrm{Cov}\left(S_{N,v}, S_{N,w}\right) = \left(\frac{v}{N-1}\right)\left(1 - \frac{w}{N-1}\right) c_N^2,$$

where $1 \le v \le w \le N-1$ and $c_N^2 = N(N+1)(N-1)^2/180$. An expression for the variance of $S_{N,k}$ is desired in order to find its exact distribution, but hard to determine due to its dependence upon the covariance between $T_{N,i}$, which is also difficult to determine analytically. Simulation has suggested that $\mathrm{Cov}\left(T_{N,i}, T_{N,j}\right) = \mathrm{Cov}(T_{N,1}, T_{N,2})$ for all $i \neq j$, so for the sake of analysis this is assumed to be true. Under this assumption, it follows that:

(3.12) $$\underset{\sim}{\sigma}^2_{N,k} = Var\left[S_{N,k}\right] = \frac{kN(N+1)(N-k-1)}{180},$$

where the underscore-tilde denotes that the equality depends upon the covariance assumption.

3. Ensemble Sum of Pair-Maxima, K_N

The asymptotic normality of T_N suggests that $S_{N,k}$ is also asymptotically normal. Furthermore, the covariance structure of $S_{N,k}$ is the same as that of a Brownian bridge. This leads to a choice of test statistic:

(3.13) $$K_N = \max_{k \in (1,2,\ldots,N/2)} \left(\frac{\mu_{N,k} - S_{N,k}}{c_N}\right),$$

where from (3.11), $c_N = \sqrt{c_N^2} = (N-1)\sqrt{\frac{N(N+1)}{180}}$.

The exact distribution of K_N is known for normal $S_{N,k}$; Ruth and Koyak (2011) call this the Ensemble Sum of Pair-Maxima (ESPM) statistic. They demonstrate through simulation that $S_{N,k}$

does not appear to be asymptomatically normal for low dimensions. As a result, the exact null distribution of K_N remains unknown and is an open problem. Fortunately, simulation-based critical values from Ruth and Koyak (2011) enable the use of K_N as a test statistic.

These approximated critical values for K_N are varied for N, α, and the dimensionality of the observations being tested. They can be found in Table 2 of Ruth and Koyak (2011).

4. ESPM Performance Characteristics

Part of our proposed work was to independently verify ESPM performance characteristics (Table 3) as published in Ruth (2009). The original simulations suggest this test is powerful against a wide-range of change-point alternatives. A total of 1000 samples were generated for each of 30 vignettes which vary by dimensionality (p = 5, 20), type of change (jump or linear drift), multivariate distribution type (normal, normal mixture, or Weibull), the parameter θ affected by change (mean vector, covariance matrix, or scale parameter), and lastly, the magnitude of change (Δ). In each vignette, the true change point is located at k = 101. For each vignette in which a mean-change takes place, all samples begin with a mean vector of zero and end with the mean vector having magnitude Δ, either as a jump or linear drift starting at the change point. For vignettes in which the covariance matrix or scale parameter changes, all samples start out with unit parameters and end with the parameter multiplied by a value of (1 + Δ), but only in the first variate. Multivariate normal mixtures generate observations with a zero mean and an identity covariance matrix with a probability 0.9, and observations with a zero mean and 16 times the identity covariance matrix with probability 0.1. Multivariate Weibull vectors are comprise of independent, univariate Weibulls with shape parameters = 1.5 and scale parameters = 1. All simulations were conducted with a nominal test level of α = 0.05; thus, power when Δ = 0 should be within simulation error of the nominal test level. Two-sided confidence intervals are computed as in Devore (2004).

Our results successfully recreate the published performance characteristics, and in some cases (highlighted in yellow), strongly suggest that the true power level of the ESPM test may be higher than previously thought. This reaffirms the impressive nature of graph-theoretic tests that use ensembles to elicit information from data and assigns great potential value to any effort seeking to bring these procedures into widespread usage in modern applications.

Table 3: Verification of simulated powers for ESPM test under different distributions and change scenarios based on a total of 1000 simulations, $\alpha = 0.05$, $N = 200$, change point $k = 101$, p = dimensionality. All confidence tests are at 95% level.

		Jump				Drift				
Δ		R/K Value	Tested Value	Low CI	High CI	R/K Value	Tested Value	Low CI	High CI	
				Multivariate normal, θ = mean, p = 5						
0		0.04	0.057	0.044	0.073	0.06	0.064	0.050	0.081	
0.5		0.60	0.624	0.594	0.653	0.27	0.280	0.253	0.309	
1		1.00	0.999	0.994	1.000	0.84	0.853	0.830	0.874	
				Multivariate normal, θ = mean, p = 20						
0		0.05	0.062	0.049	0.079	0.05	0.064	0.050	0.081	
0.5		0.33	0.352	0.323	0.382	0.13	0.135	0.115	0.158	
1		0.95	0.976	0.965	0.984	0.56	0.585	0.554	0.615	
				Multivariate normal, θ = covariance matrix, p = 5						
0		0.05	0.062	0.049	0.079	0.05	0.057	0.044	0.073	
0.5		0.97	0.973	0.961	0.981	0.52	0.537	0.506	0.568	
1		1.00	1.000	0.996	1.000	1.00	0.999	0.994	1.000	
				Multivariate normal mixture, θ = mean, p = 5						
0		0.04	0.064	0.050	0.081	0.06	0.057	0.044	0.073	
0.5		0.56	0.548	0.517	0.579	0.21	0.239	0.214	0.267	***
1		0.99	0.997	0.991	0.999	0.76	0.783	0.756	0.807	***
				Multivariate Weibull, θ = scale parameter, p = 5						
0		0.06	0.064	0.050	0.081	0.05	0.057	0.044	0.073	
0.5	***	0.70	0.771	0.744	0.796	0.35	0.426	0.396	0.457	***
1		0.99	0.996	0.990	0.998	0.86	0.901	0.881	0.918	***

***Indicate situations simulated where power level exceeds published power levels with statistically significant difference.

With the ESPM test thoroughly explored, we now transition over to new work to extend the ESPM test for use in online situations.

B. MULTIPLE TESTING THEORY

In our Introduction, we stated that the goals behind our new test are to identify change correctly when it occurs and detect that change as quickly possible after it does occur. The first goal is one common to all change detection tests; the second though comes about as a result of our focus on online situations where testing occurs as data is acquired. To detect change as quickly as possible, we make use of *multiple testing theory* and seek to test every time new observations are added to the data instead of waiting for all observations to be collected. A sample of N observations allows for as many as $N - 3$ sequential tests in order to detect any possible change up until all N observations have been collected given that the ESPM test assumes there are at least four observations to start with when first testing. Although this extra testing allows for potential early warning, it raises two questions: 1) how do we maintain overall test level and control the Type I error rate and 2) what data ought to be included in each test as new observations arrive?

Depending on the size of N, this approach can obviously lead to a high number of hypothesis tests and a problematic increase in Type I Error as seen in Table 4. For example, the monitoring of a certain machine may result in performing 200 separate hypotheses tests to find change. Through the use of a standard test level $\alpha = 0.05$, it is expected that 10 tests will be deemed "significant" and flag positive for change even when the null hypothesis, H_0, is actually true. In general form, there is probability α of a Type I error and complementarily probability $(1 - \alpha)$ that a false positive is avoided in any given test. It follows that if m independent hypothesis tests are performed, there is probability $(1 - \alpha)^m$ of no false positives in m tests.

Table 4: Breakdown of possible outcomes in hypothesis and their respective probabilities of occurrence.

Type I and II Errors		
	Actual Condition	
Decision	$\boldsymbol{H_0}$ **True**	$\boldsymbol{H_0}$ **False**
Do Not Reject $\boldsymbol{H_0}$	Correct Decision $(1 - \alpha)$	Incorrect Decision Type II Error β
Do Reject $\boldsymbol{H_0}$	Incorrect Decision Type I Error α	Correct Decision $(1 - \beta)$
$\alpha = P(Type\ I\ Error)$ and $\beta = P(Type\ II\ Error)$		

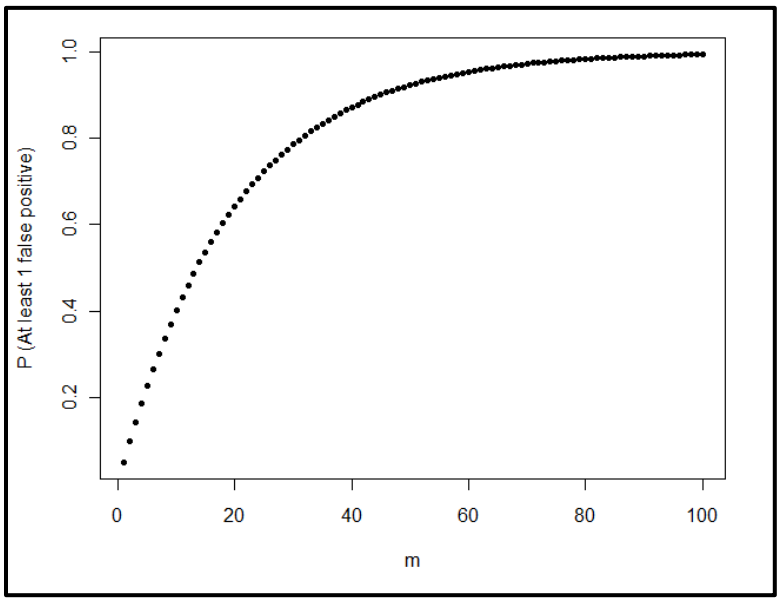

Figure 9: Type I error rate as the number of independent hypotheses (m) under test increases.

Thus, a probability $1 - (1 - \alpha)^{m}$ of making at least one Type I error exists when m independent tests are performed. Figure 9 demonstrates how quickly the Type I error rate, or chance of encountering at least one false positive, rises as m is increased.

Table 5 summarizes the situation in a traditional form (see for example Benjamini and Hochberg (1995)). There are m hypotheses to be tested, of which m_0 are true for the null hypothesis; **R** serves conversely as the number of hypotheses rejected. **R** is an observable random variable, whereas **U**, **V**, **S**, and **T** are unobservable random variables. **V** is the number of Type I errors. If each null hypothesis is tested individually at level α, then **R** increases as α increases. Use of equivalent lower case letters is done to signify the realized values of the same variables.

Table 5: Number of errors committed when testing m hypotheses

	Declared non-significant	*Declared significant*	*Total*
True null hypotheses	**U**	**V**	m_0
Non-true null hypotheses	**T**	**S**	m - m_0
	m - **R**	**R**	m

There are various methods for controlling the Type I error rate and maintaining overall test level. In terms of the random variables from Table 5, the *per-comparison error rate* (PCER) is the expected value of Type I errors to the number of tested hypotheses; that is, PCER $= E[\mathbf{V}]/m$. Testing each hypothesis individually at level α/m guarantees that $E[\mathbf{V}]/m \leq \alpha$, which means that PCER attempts to minimize, but not eliminate Type I errors. The *family-wise error rate* (FWER) is the probability of at least one Type I error occurring during testing, where FWER $= P(\mathbf{V} \geq 1)$. Testing each hypothesis individually at level α guarantees that $(\mathbf{V} \geq 1) \leq \alpha$. This method specifically seeks to guard against even one Type I error occurring. Lastly, the *false discovery rate* (FDR) is the expected proportion of Type I errors to the number of rejected hypotheses. Unlike many classical approaches which control FWER in the strong sense, our new test will use FDR to control the Type I error rate, which we will discuss in the following section.

In offline situations, testing is naturally done upon all the data collected as it would be detrimental to test efficacy to leave any observations out. In online situations, however, there is a decision to be made concerning which available data to include in each test. We investigated two techniques for applying multiple testing to online change detection situations before ultimately moving forward with one.

Consider a sample *R* of N *d*-dimensional observations. In *overlap testin*g, the first test of the sequence is a subgroup of f observations spanning from indices 1 to f. For each subsequent test, both the front and back index shift by a value of s to incorporate the newest available data without the subgroup being tested growing beyond f observations in size. In *telescope testing*, the first test of the sequence is again a subgroup of size f spanning from indices 1 to f. In this case though, for each subsequent test only the larger index shifts by a value of s to incorporate the newest data causing the tested subgroup to grow by s observations with every test. The subgroup of the final test has a size of N and is equal to *R*. Our new test uses *telescope testing*.

The answers to both of the questions we posed at the beginning of this section directly affect the nature of our online extension to the ESPM test and, in turn, ultimately determine its viability. We will elaborate further on the theory and motivation behind these decisions.

1. False Discovery Rate

Here we follow the groundbreaking work of Benjamini and Hochberg (1995). FDR is the expected value of the proportion of rejected null hypotheses that are falsely rejected. This proportion may be expressed by the random variable **Q** = **V** / (**V** + **S**). When **V** + **S** = 0, we define **Q** = 0, as in false rejection cannot occur if no null hypothesis is rejected in the first place. **Q** is an unobserved random variable because it is impossible to exactly know the values of v and s, and therefore $q = v/(v+s)$, even after all experimentation and data analysis is completed. The FDR is defined as the expected value of **Q**:

(3.14) $$\text{FDR} = E[\mathbf{Q}] = E[\mathbf{V}/\mathbf{R}] = E[\mathbf{V}/(\mathbf{V} + \mathbf{S})].$$

There are many properties of this error rate, but two specific ones are particularly fundamental, and coincidentally easy to show. Consider again a situation where m hypotheses are being tested:

1) If all null hypotheses are true, the FDR is equal to FWER. In this case, $s = 0$ and $v = r$ if any tests are declared significant, so if $v = 0$ then **Q** = 0, and if $v > 0$ then **Q** = 1, which results in $P(\mathbf{V} \geq 1) = E[\mathbf{Q}]$. Thus, control of the FDR is also in fact control of the FWER in an indirect and less effective manner.
2) In situations where $m_0 < m$, the FDR is less than or equal to the FWER. In these cases, if $v > 0$ then $v/r \leq 1$, and $P(\mathbf{V} \geq 1) \geq E[\mathbf{Q}]$. Therefore, any procedure that controls the FWER intrinsically controls the FDR. Yet, any procedure that controls the FDR only will likely be less stringent and, in turn, a gain in power will result. This becomes especially apparent the more non-true null hypotheses exist in the data. **S** tends to be larger and so does the difference between the error rates; ergo the potential for increases in power is larger the closer m_0/m is to 1.

To control the random variable **Q** at each hypothesis declared significant would be optimal. This is impossible though. If $m_0 = m$ and even a single null hypothesis is rejected, then $v/r = 1$ and **Q** cannot be controlled because every rejection is false. Adding in the extra condition ($\mathbf{V}/\mathbf{R} \mid \mathbf{R} > 0$) does not alleviate the problem, and neither does $E[\,\mathbf{V}/\mathbf{R} \mid \mathbf{R} > 0]$. Instead, FDR equally expressed as $E[\,\mathbf{V}/\mathbf{R} \mid \mathbf{R} > 0] \cdot P(\mathbf{R} > 0)$, which is possible to control.

The dominant influence for the method by which we control FDR in our new test is the famous Benjamini-Hochberg Procedure (1995), which for independent test statistics and any arrangement of non-true null hypotheses impressively controls FDR at α^* via a linear adjustment of test levels. Consider testing m hypotheses $(H_1, H_2, \dots, H_m)$ with corresponding p-values $(P_1, P_2, \dots, P_m)$. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered p-values and $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. They prove that the following Bonferroni-type multiple testing decision rule controls FDR at overall test level α^*:

(3.15) let k be the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha^*$;

then reject all $H_{(i)}$ for $i = \{1,2, \dots, k\}$.

2. Extension of Benjamini-Hochberg Procedure to Online Scenarios

Although the Benjamini-Hochberg Procedure (1995) controls the FDR, it is designed for use in offline situations only. Three of its basic assumptions make it unable to work in online scenarios in its current form. First, the procedure assumes that all hypotheses have been tested and have corresponding p-values before its implementation. In an online problem, where testing is ongoing as new observations are collected, this caveat would force the user to wait until all data was collected. Second, the procedure as it stands requires the number of tested hypotheses to be determined prior to its use. Online scenarios, by contrast, involve ongoing processes for which the end is not explicitly known. For example, a mechanical system being observed might be a "lemon" and experience system failure soon after it starts being used and tested, or it could give a decade of faithful uninterrupted service and provide a lot of observations for testing. This makes it nearly impossible to determine how many tests are going to be conducted prior to use. Thirdly, Benjamini and Hochberg (1995) stipulate that each hypothesis must be independent from the others. It turns out that in many real-world applications, dependency exists among the tested hypotheses. We will now proceed to discuss our solutions to these problems before giving an exact definition of our procedure extending control of the FDR to online situations.

Because online problems involve incremental accumulation of data, one may only test some subset of all null hypotheses at any given time. As a conservative approach, we treat all untested hypotheses as if they had been tested, with p-values all equal to 1. Once pertinent data are available to test a yet-untested hypothesis, we replace the "placeholder" p-value of 1 with the

actual value and then order all the p-values. The key to our procedure is the fact that the Benjamini-Hochberg Procedure evaluates *ordered* p-values, and our placeholding scheme simply ensures that that the p-values associated with untested hypotheses are guaranteed to be last in order. We outline this procedure in detail shortly.

To combat the problem of uncertain endpoints, we choose an observation horizon to determine the correct linear adjustment of test levels prior to applying the procedure. Based on the typical operating profile of the system being studied, this horizon H is set at N observations and allows all observations less than or equal to N in sequence label to be available for testing. Any observation greater than N in sequence label goes untested as seen in Figure 10, where the horizon for a general mechanical system is set at 150 observations, but the machine lasts for 200 observations causing the last 50 to go untested. In general, H ought to be set to include the entire range of interest for possible change detection, but not be excessively large (which can reduce test power). We discuss this H-selection issue more in our results section.

The solution for the problem of dependency among hypotheses comes from Benjamini and Yekutieli (2001). Realizing that dependent test statistics are encountered often when trying to control the FDR in practice, they develop a new procedure that controls the FDR for positively

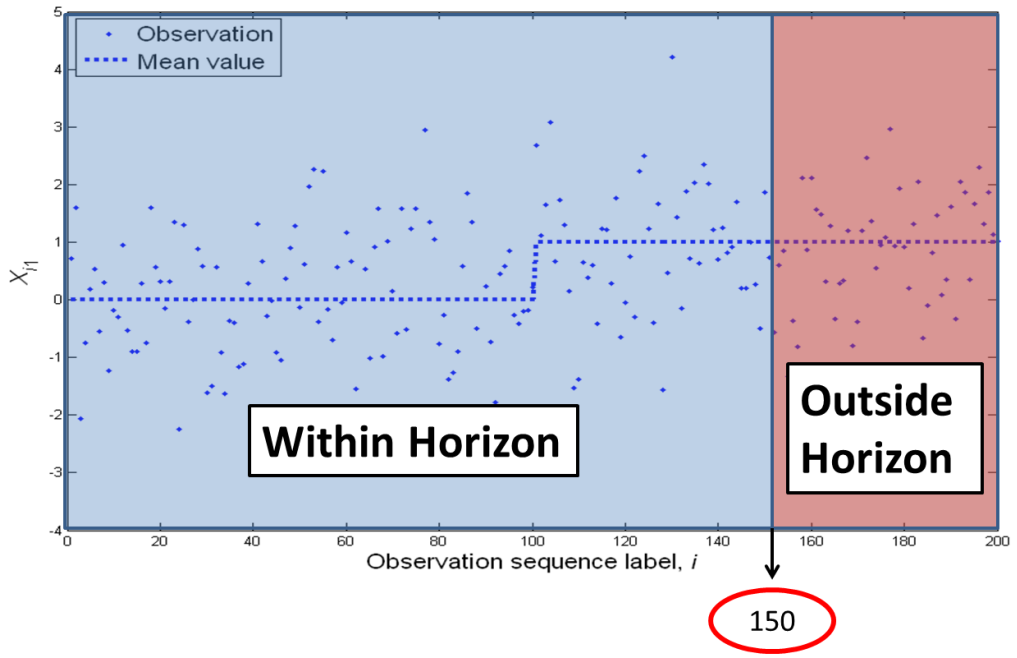

Figure 10: Horizon H set at 150 observations for a machine that happens to live for 200 observations.

dependent test statistics. For other cases of dependency, they prove that the same procedure can be easily modified to control the FDR, but the resulting procedure is more conservative. Since the exact distribution of K_N is not known, we are unable to formally prove that our test statistic meets the dependency conditions set forward by Benjamini and Yekutieli (2001) in order to use their procedure. But, the simulation study we conduct in Section IV points to the result that K_N meets these conditions because test efficacy under the alternative hypothesis is not unreasonably suppressed compared to the efficacy of the original offline test whereas the opposite would be expected if K_N did not meet the conditions. With nothing suggesting so far that it violates these dependency conditions in any important way, we make use of the procedure in our new test.

3. Online Extension Procedure

Consider testing m independent true null hypotheses $(H_1, H_2, \dots, H_m)$ with corresponding p-values $(P_1, P_2, \dots, P_m)$ on those observations. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$ be the ordered p-values and $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$ and let $\mathbf{P}_{(m)} = \left(P_{(1)}, P_{(2)}, \dots, P_{(m)}\right)$. Suppose these p-values are realized in sequence; that is, at step j only p-values P_1 through P_j are known. We propose the following on-line testing procedure for each step $j \in \{1, \dots, m\}$ for m sequential observations:

(1) Let α^* be the desired overall test level, and let $\mathbf{Q}_0 = (1,\ 1, \dots, 1)$.

(2) Let $\mathbf{Q}_j = \left(Q_1, Q_2, \dots, Q_j, 1, 1, \dots, 1\right)$ for $1 \leq j \leq m$, where $Q_i = P_i \ \ \forall i \leq j$.

(3) Let $\mathbf{Q}_{(j)} = \left(Q_{(1)}, Q_{(2)}, \dots, Q_{(j)}, 1, 1, \dots, 1\right)$ were $Q_{(i)}$ is the i^{th} ordered value in $\mathbf{Q}_j$.

(3.16)
- If there exists some $k \leq j$ and some $i \leq j$ such that $Q_k = Q_{(i)} \leq \frac{i}{m}\alpha^*$, then reject H_k and declare that some change has occurred in the distribution sequence.
- If $Q_{(i)} > \frac{i}{m}\alpha^* \ \forall i \leq j$, go back to step (2) with $j = j + 1$ until $j = m$.

That this procedure controls FWER at the desired level is established in the following result:

Theorem: For m independent test statistics and m independent true null hypotheses in online scenarios, the procedure (3.16) controls the FWER at level α^*.

Proof: Let V be the number of Type I errors in the offline Benjamini-Hochberg test. Let V_j be the number of Type I errors through step j. By Benjamini and Hochberg (1995), $P(V \geq 1) \leq \alpha^*$.

We must show that for any $j \in \{1, \ldots, m\}$, $P(V_1 \geq 1 \text{ or } V_2 \geq 1 \text{ or } \ldots \text{ or } V_j \geq 1) \leq \alpha^*$. So, choose $j \in \{1, \ldots, m\}$ and notice that $V_1, V_2, \ldots, V_j$ is a monotone non-decreasing sequence. Therefore, $P(V_1 \geq 1 \text{ or } V_2 \geq 1 \text{ or } \ldots \text{ or } V_j \geq 1) = P(V_j \geq 1)$. Next, observe that

$$P(V_j \geq 1) = P\left(Q_{(1)} \leq \frac{\alpha^*}{m} \text{ or } Q_{(2)} \leq \frac{2\alpha^*}{m} \text{ or } \ldots \text{ or } Q_{(j)} \leq \frac{j\alpha^*}{m}\right). \tag{3.17}$$

Now recognize that $\mathbf{P}_{(m)} \leq \mathbf{Q}_{(j)}$ (where the "≤ " relationship is taken element-wise), so

$$P\left(Q_{(1)} \leq \frac{\alpha^*}{m} \text{ or } Q_{(2)} \leq \frac{2\alpha^*}{m} \text{ or } Q_{(3)} \leq \frac{3\alpha^*}{m} \text{ or } \ldots \text{ or } Q_{(j)} \leq \frac{j\alpha^*}{m}\right)$$

$$\leq P\left(P_{(1)} \leq \frac{\alpha^*}{m} \text{ or } P_{(2)} \leq \frac{2\alpha^*}{m} \text{ or } P_{(3)} \leq \frac{3\alpha^*}{m} \text{ or } \ldots \text{ or } P_{(j)} \leq \frac{j\alpha^*}{m} \text{ or } \ldots \text{ or } P_{(m)} \leq \alpha^*\right)$$

$$= P(V \geq 1) \leq \alpha^*. \tag{3.18}$$

Therefore, $P(V_j \geq 1) \leq \alpha^*$ for all $j \in \{1, \ldots, m\}$. ■

Figure 11 illustrates a linear adjustment of test levels created using the online extension procedure for a scenario with five hypotheses and a desired overall test level $\alpha^* = 0.05$.

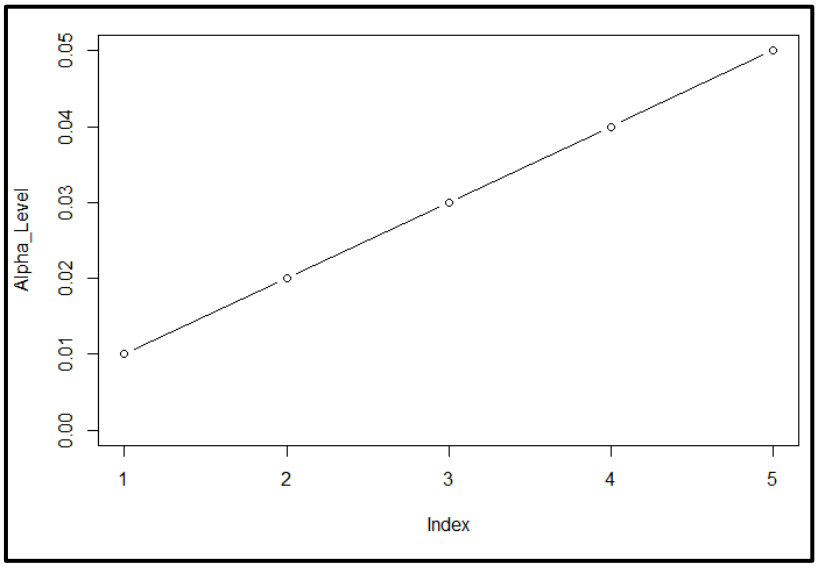

Figure 11: FWER Linear Adjustment for 5 Tests, $\alpha^* = 0.05$

4. Hypothesis Selection Procedure

With more tests being conducted due to multiple testing, more thought must be put forward in regards to the structuring of the hypotheses being tested. A faulty hypothesis structure

can rule ineffective any and all testing that is conducted. If tested subgroups are too small or skip over too much data, then detection power will suffer. Each pair of new observations added to a data set can substantially change the MNBMs computed in calculating K_N; it is not necessary to wait for large groups of observations to be acquired in between tests. So we seek to reflect that in our chosen methodology.

Our new test selects hypotheses for testing through what we call *telescope testing*. We initially wait for a small group of observations to be available for the first test which utilizes all data present and then proceed to test on all available data each time observations are added to the data set. As a result, the testing region grows in size like a telescope with each subsequent hypothesis being tested including slightly more data than the previous hypothesis. For example, consider an online change-point problem with the horizon set at 200 observations where up to m hypotheses $(H_1, H_2, \dots, H_m)$ may be tested. The initial hypothesis, H_1, is tested when there are 20 observations available and every ensuing hypothesis, H_{1+i}, is tested each time two observations are added to the data set, so hypothesis, $H_{(1+i)<m}$, is tested on $20 + 2i$ observations. The final hypothesis, H_m, is tested upon 200 observations. This selection of hypotheses creates the need for $m = 91$ hypotheses to be considered and so the linear adjustment of the FDR is constructed to handle up to and including 91 tests as seen Figure 12.

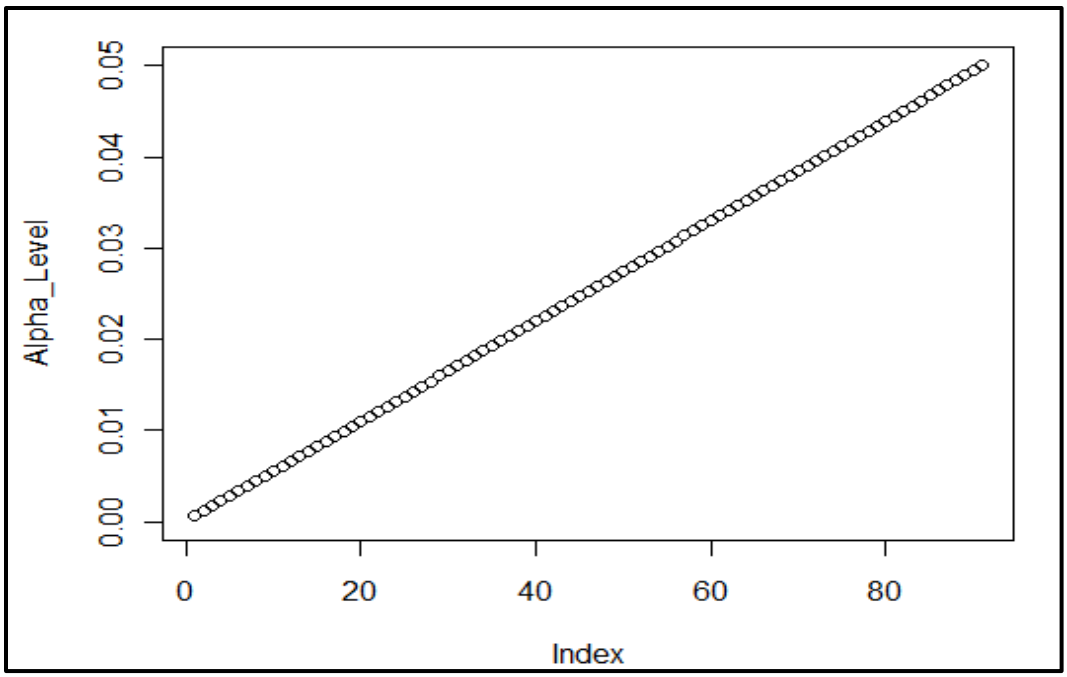

Figure 12: FDR Linear Adjustment for 91 Tests, $\alpha^* = 0.05$

With our new test fully explained, we will now proceed to demonstrate its performance through two studies. The first uses simulated data to investigate the power and advance warning capabilities of our test while varying the change-point, magnitude of change, and dimensionality. The other occurs on pseudo real-world data and allows us to see whether or not our test can potentially handle the rigors of powerfully detecting change, and detecting it early, in real-world scenarios.

IV. TEST PERFORMANCE

The new test presented in this paper is only useful if it detects change reasonably well-ahead of when it would be detected in an offline situation. In this section, we present two different investigations into the performance characteristics of our test. The first is a simulation study that examines the power of our new test and its capability to detect change ahead of the ESPM test. The second is an application of our test to simulated real-world data from the International Conference on Prognostics and Health Management Data Challenge in 2008.

The two primary performance metrics that we use in both studies are detection power, where power is defined as the probability of rejecting the null hypothesis when it is false, and advanced warning, which is the number of observations prior to test detected change. Interpoint cost is determined by the Euclidean distance function given by (2.6).

A. SIMULATION STUDY

1. Methodology

Here we simulate a variety of change-point scenarios to quantify the power of this new test and its ability to provide advance warning of change unavailable in offline testing. Each power estimate in the tables and figures of this section is the fraction of times the new test detected change under the given conditions based on 1000 simulations. Each advanced warning estimate is the average of the number of observations prior to the simulated end of life at which change is detected. If no change is detected, advanced warning is zero for the purposes of computing the mean. In every case, total sample size is $N = 200$, the test significance level is $\alpha = 0.05$, and sample space is $\mathbb{R}^5$. The choice of $N = 200$ is based on the desire to directly compare the performance of our online test to the original offline form of the ESPM test, while concurrently avoiding extremely long computation times for larger optimal nonbipartite matchings. This is equivalent to setting a horizon at 200 observations.

Observations were simulated using RStudio each observation is drawn from a standard five-variate normal distribution with density function of the form:

$$f_{MVN}(\mathbf{X};\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} e^{\frac{-1}{2}(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})} \quad \forall \ \mathbf{X} \in \mathbb{R}^5, \tag{4.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^5$ and $\Sigma \in \mathbb{R}^{5x5}$ are the distribution mean and covariance matrix respectively. The pre-change-point data for this multivariate normal case have $\boldsymbol{\mu} = 0$ and $\Sigma = I_5$, where I_5 is the 5 x 5 identity matrix. The post-change-point data have a different mean or scale as specified later in this section.

We use three different formats to select the way in which tests are conducted. In the first, we wait until there are 20 observations to first test, and then proceed to test every time two additional observations are added to the data. So, we test at 20 observations, 22, 24, 26… all the way until 200 ("20-2"). Under this format, there is a possibility for up to 91 tests to be conducted contingent on if change is detected prior to the last test. The second format again begins testing at 20 observations and then tests every time an additional 20 observations are added to the data ("20-20"). So, we test at 20, 40, 60,…, 200. Here there can be up to 10 tests. The third and final format begins testing at 40 observations and tests for every 40 additional observations after that ("40-40"). So, we test at 40, 80, 120,…, 200 and there can be a total of 5 tests. By only testing when there is an even number of observations within all three formats, we maintain a situation conducive to nonbipartite matching without the need to create dummy points.

In this study, we consider various change-points. For our purposes, change location fraction c^* is defined as $\frac{\varphi-1}{N}$, where φ is the change point as before. Notice that the value of the denominator is the sample size N, not the size of any specific tested subgroup. We examine values of 0.10, 0.25, 0.335, 0.50, 0.665, 0.75, and 0.90 for c^*, which corresponds to change at observations 21, 51, 67, 101, 133, 151, or 181.

For each of these values of c^*, we consider change magnitudes Δ of 0, 0.5, and 1.0. When $\Delta = 0$, the null hypothesis is true and the power estimate is an estimate of the test's Type I error rate. In ideal situations, the Type I error rate is equal to the chosen significance level. When $\Delta > 0$, Δ indicates the total magnitude of change. We only look at jump changes, where Δ is the magnitude of the abrupt change that occurs at the designated change point.

Our simulations only look at changes in distribution mean. Without loss of generality, this change is implemented in the first component of each observation only in the following form:

(4.2)
$$\mathbf{X}_i \sim F_{MVN}, i < \varphi$$
$$\mathbf{X}_i + (\Delta, 0, 0, 0, 0) \sim F_{MVN}, i \geq \varphi$$

due to the rotational invariance of F_{MVN}. Thus, the post-change-point data for the multivariate normal case have $\boldsymbol{\mu} = (\Delta, 0,0,0,0)'$ and $\Sigma = I_5$.

2. Performance Results

Figures 13-15 and Tables 6-7 display power estimates for our new test at a significance level of $\alpha = 0.05$. The estimated realized significance level for the test with $N = 200$ is $\alpha_{actual} = 0.04435$ giving credence to our new method's ability to control the overall test level. Critical values for the K_N^* test statistic are determined from the simulations run by Ruth (2009).

As expected, we observe a distinct improvement in power from the $\Delta = 0.5$ case to the $\Delta = 1$ case. Also it is known that ESPM test power is degraded when the change point is in the tails of the tested subgroup (away from the middle) the test suffered somewhat, since fewer pre- or post-change data (depending on the location) are present to indicate a change. Our results again generally show this to be true. For the $\Delta = 1$ case, the average powers corresponding to c^* values of 0.25, 0.335, 0.5, 0.665, and 0.75 are at or above 75% for all three testing formats whereas both tails fall below that. Comparing the two, power at the left tail where $c^* = 0.1$ exceeds that at the right tail where $c^* = 0.9$. This is likely because when $c^* = 0.1$ the change point remains in the tested subgroup for nearly all tests conducted, while when $c^* = 0.9$ post-change data does not enter the tested subgroup until the last few tests making change difficult to detect. For the 20-20 and 40-40 formats, power at $c^* = 0.1$ and $c^* = 0.9$ is higher than seen in the 20-2 format. This effect appears to result from more stringent step-wise adjustment of critical values for the 20-2 format. With the potential for up to 91 tests taking place, each step of the adjustment is smaller and so the threshold to declare significance against the first steps of this format $\left(\frac{\alpha*}{91}, \frac{2\alpha*}{91}, \frac{3\alpha*}{91}, \ldots\right)$ is more restrictive than for the first steps of the other formats (20-20: $\frac{\alpha*}{10}, \frac{2\alpha*}{10}, \frac{3\alpha*}{10}, \ldots$ and 40-40: $\frac{\alpha*}{5}, \frac{2\alpha*}{5}, \frac{3\alpha*}{5}, \ldots$). For the $\Delta = 0.5$ case, the same general relationship exists between formats and values of c^* except power estimates are much lower with highs of 40-50% depending on the format. The highest power estimates from the new online test in the $\Delta = 1$ ($\sim$95-100%) case are essentially equivalent to offline power estimates for the ESPM test ($\sim$100%), but this similarity does not hold for the $\Delta = 0.5$ case where the online test loses some power likely due to the use of step-wise adjusted significance levels (40%-50% for online test; $\sim$60% for ESPM test).

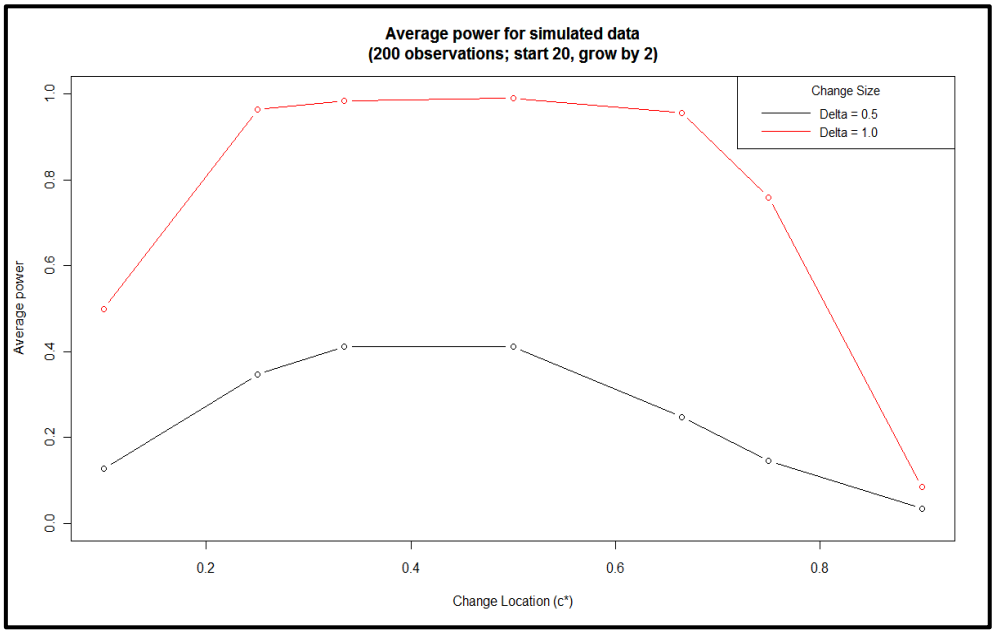

Figure 13: Average power for testing format of start 20, grow by 2, across values of c^*

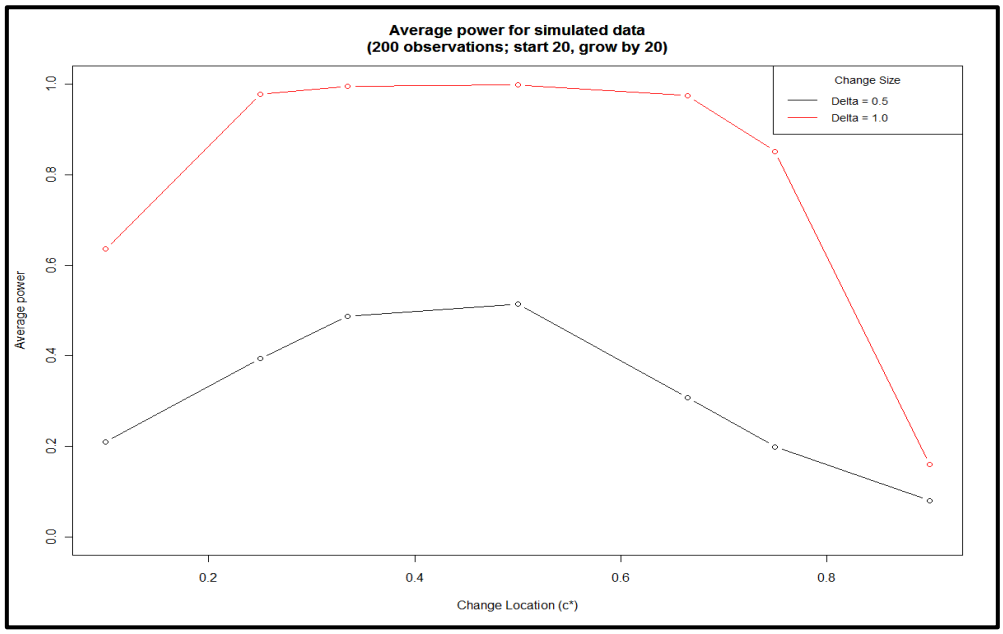

Figure 14: Average power for testing format of start 20, grow by 20, across values of c^*

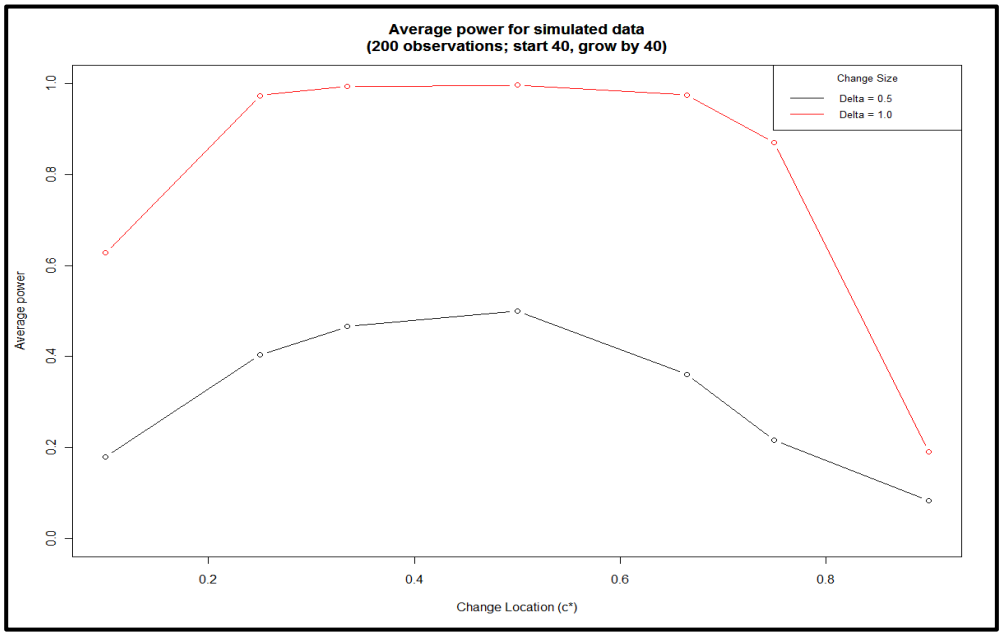

Figure 15: Average power for testing format of start 40, grow by 40, across values of c^*

Table 6: Average simulated power across different testing formats and values of c^* at change magnitude $\Delta = 0.5$

Multivariate Normal, Θ=mean, d=5, Δ = 0.5,N=H=200

		Start (Grow)		
		20 (20)	*40 (40)*	*20 (2)*
C*	*0.10*	0.209	0.179	0.127
	0.25	0.394	0.404	0.347
	0.335	0.487	0.465	0.411
	0.50	0.515	0.499	0.411
	0.665	0.307	0.360	0.248
	0.75	0.199	0.215	0.146
	0.90	0.080	0.083	0.034

***Numbers in parentheses represent telescope growth value.

Table 7: Average simulated power across different testing formats and values of c^* at change magnitude Δ = 1.0

Multivariate Normal, Θ=mean, d=5, Δ = 1.0, N=H=200

		Start (Grow)		
		20 (20)	*40 (40)*	*20 (2)*
C*	*0.10*	0.635	0.628	0.499
	0.25	0.978	0.974	0.963
	0.335	0.995	0.994	0.984
	0.50	0.998	0.996	0.990
	0.665	0.974	0.975	0.956
	0.75	0.851	0.870	0.759
	0.90	0.160	0.190	0.085

***Numbers in parentheses represent telescope growth value.

Figures 16-18 and Tables 8-9 present advance warning estimates. Each of the three formats displays similar ability to provide advanced warning. The only noticeable differences arise at the tails and are very minor other than the Δ = 1, $c^* = 0.1$ case which is a result of the reduced power of the 20-2 format in that situation. We attribute the minor differences to a combination of natural variation and dissimilarity between the step-wise growth and selection of hypotheses in each testing format. Each warning estimate is also slightly affected by the amount of warning they can provide based on the corresponding change location fraction, or c^*; early change locations are inherently able to give more warning than later ones. This appears to override the increased power found at $c^* = 0.335, 0.5,$ and 0.665. In the Δ = 1 case, the warning estimate at $c^* = 0.25$ completely overshadows the warning estimates for $c^* = 0.5$ and $c^* = 0.665$, and has a visually apparent difference over $c^* = 0.335$. For the Δ = 0.5 case, these differences are far less pronounced, but still there. The best average advanced warning times of this case are about 30 observations for all formats, while in the right tail of all formats the estimates fall below 20 observations. Upon observation, the shape of the warning estimates demonstrates resemblance to the shape of the power estimates. This is not surprising because the more often that change is detected, the greater average advanced warning should be. Despite this similarity between the shapes of the power estimates and advanced warning estimates though, the maximum advance warning estimate is not found at $c^* = 0.5$. All three formats provide the best advanced warning when the change point is near, not at, the middle of the first half of

observations, giving the advanced warning distribution a positive skew. Notice in particular that for $c^* = 0.25$, corresponding to a change point at observation 51, all three formats provide approximately 100 observations of advanced warning. In other words, the online test provides indication that a change has occurred with only 100 observations available for testing, while the corresponding offline test has to "wait" for 200 observations to be available for testing. This example highlights the value added by our online test.

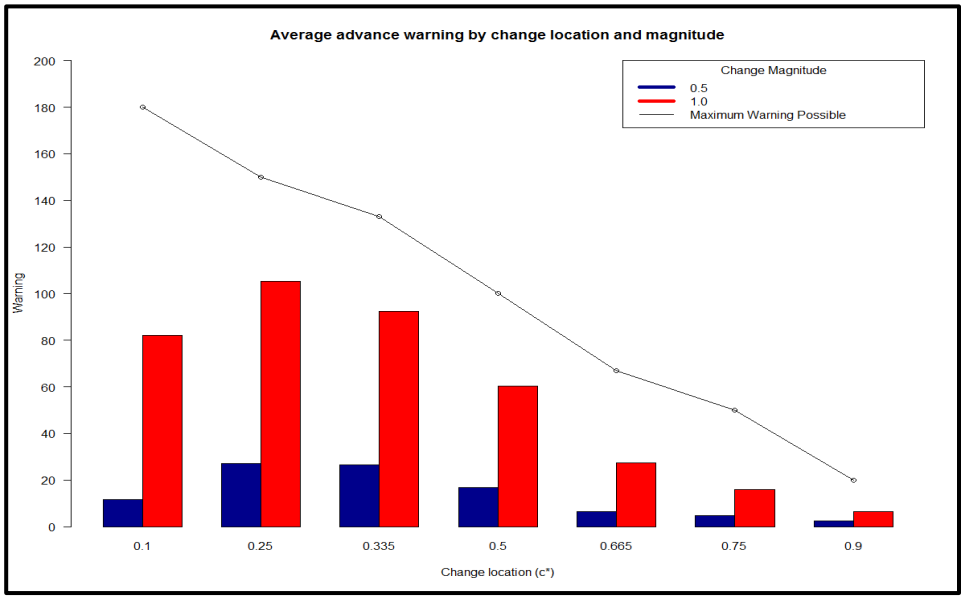

Figure 16: Average advanced warning for testing format of start 20, grow by 2, across values of c^* and Δ

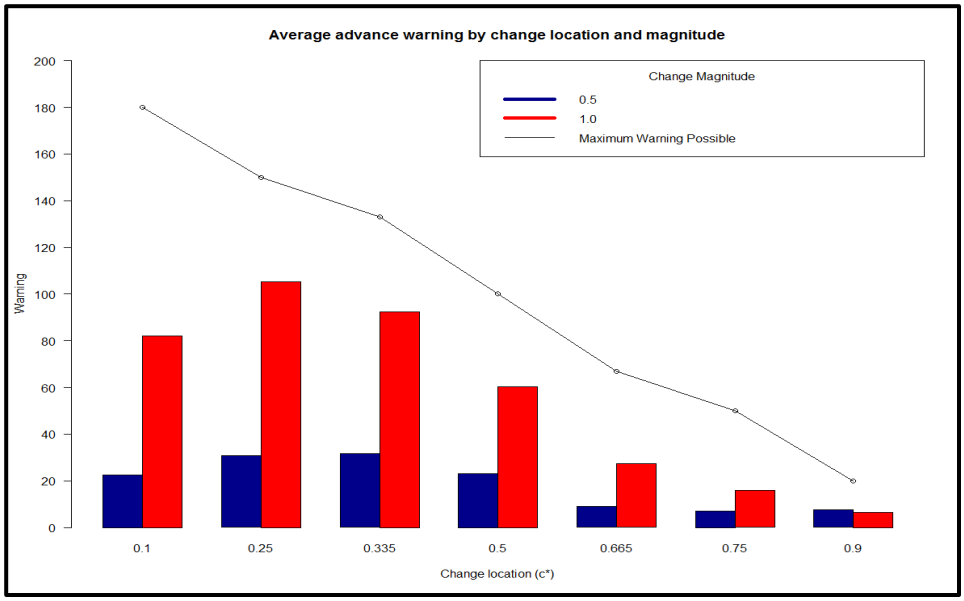

Figure 17: Average advanced warning for testing format of start 20, grow by 20, across values of c^* and Δ

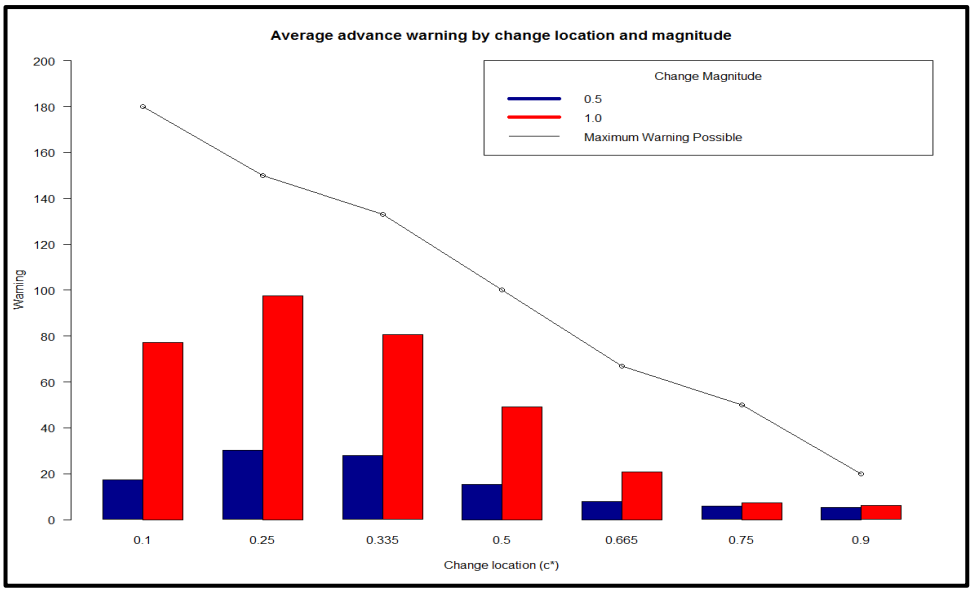

Figure 18: Average advanced warning for testing format of start 40, grow by 40, across values of c^* and Δ

Table 8: Average advanced warning across different testing formats and values of c^* at change magnitude $\Delta = 0.5$

Multivariate Normal, Θ=mean, d=5, Δ = 0.5,N=H=200

		Start (Grow)		
		20 (20)	*40 (40)*	*20 (2)*
C*	*0.10*	22.5	17.3	11.6
	0.25	30.8	30.1	27.0
	0.335	31.6	27.9	26.5
	0.50	23.2	15.4	16.8
	0.665	9.0	8.0	6.5
	0.75	7.1	5.9	4.6
	0.90	7.5	5.5	2.3

***Numbers in parentheses represent telescope growth value.

Table 9: Average advanced warning across different testing formats and values of c^* at change magnitude Δ = 1.0

Multivariate Normal, Θ=mean, d=5, Δ = 1.0, N=H=200

		Start (Grow)		
		20 (20)	*40 (40)*	*20 (2)*
C*	*0.10*	82.2	77.3	57.6
	0.25	105.4	97.6	101.5
	0.335	92.5	80.5	92.4
	0.50	60.3	49.2	62.1
	0.665	27.2	20.9	28.5
	0.75	15.9	7.5	14.3
	0.90	6.5	6.1	3.5

***Numbers in parentheses represent telescope growth value.

B. PHM DATA

1. Methodology

Having identified some of the strengths and weaknesses of our new test through simulation, we now turn our attention to performance in real-world scenarios. In such cases, data are usually not independent they rarely come from well-understood distributions, and change points are not inserted by hand at predetermined locations. To apply our test to real-world data, we use data from the International Conference on Prognostics and Health Management Data Challenge in 2008. The data set consists of 218 separate multivariate time series each from a different instance of the same complex engineering system, referred to as a "unit" (e.g., the data could be from a fleet of ships of the same type). Every unit starts with different degrees of initial wear and manufacturing variation which are unknown to the user. The wear and variation are considered typical, not part of a fault condition. There are three operational settings which have a substantial effect on unit performance. For every observation of each unit, there are three operational settings and 21 sensor measurements. As seen in the real-world, the data is contaminated with sensor noise.

At the beginning of each time series, the unit is operating normally and then develops a fault at some point during the series. This fault grows in magnitude until system failure occurs at the last observation in the series. Our goal is to detect that a change has occurred and do so prior

to the last observation available for that unit. Some units have as few as 128 observations and some have as many as 357 observations. Each "power fraction" in the tables and figures of this section is the number of units for which change was detected prior to system failure. Advanced warning estimates are the average number of observations prior to the last one in the series when change is detected. Although some units have more observations than others and can provide more warning, we treat all equally when computing the mean advanced warning.

Due to the real-world nature of this data, certain challenges arise when trying to apply our test. Control variables like the ones present in the PHM data cause multiple levels of normal operation as seen in Figure 19. The large differences between these levels can cause

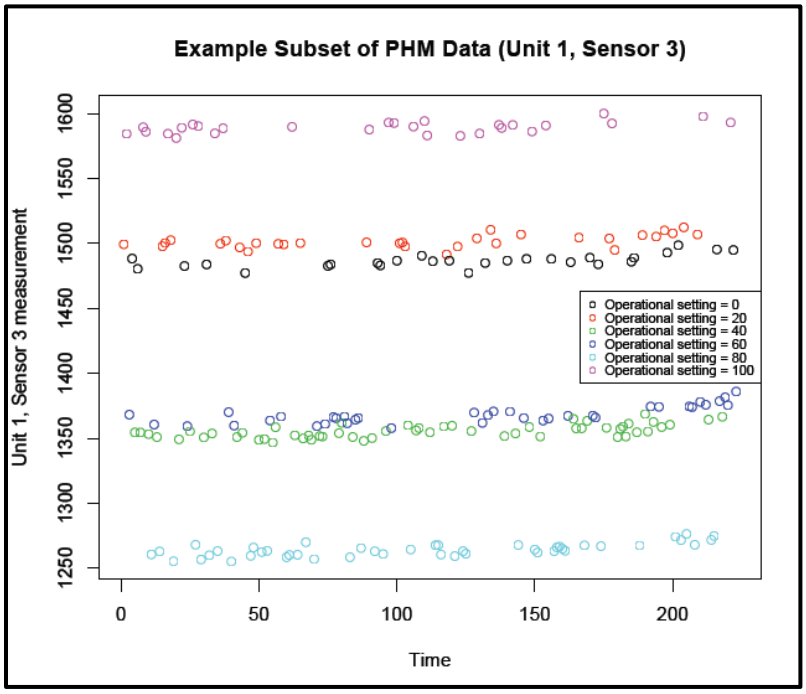

Figure 19: Subset of PHM data displaying various levels of normal operation corresponding to different control settings

graph-theoretic tests and their respective cost functions to signal change caused by changes in operational setting: clearly it is undesirable to signal unwanted change in response to such expected changes in response variables. To overcome this challenge, one might may choose to center and/or scale the data based on control variable information. For our study, we choose to center, but not scale, the PHM data. We averaged the first ten observations at each setting of the first control variable within every unit as a typical baseline for normal operation. Then, the

baseline for each setting was subtracted from every observation with the corresponding operational setting.

Another potential problem is autocorrelation. In the independent and identically distributed case, the value of a current observation does not depend at all on previous observations. However, the real-world often fails to operate in this way. Observations made on the same system close in time to each other are naturally predisposed to be close to each other, and therefore in our test be paired with each other. For example, consider an area whose geographic location inclines it to experience many slowly developing and moving low pressure systems in the atmosphere, and these systems cause persistence to daily rainfall. The daily weather in this area is a result of the past behavior of these systems and foreshadows the next day's weather as well.

In this study, we use only the 20-2 format to conduct tests as it is the format that most immediately processes newly available data. The 20-20 and 40-40 formats demonstrated slightly better performance characteristics than the 20-2 format in computer simulations, if the real-world application of interest can accommodate those lower resolution formats, then they should be considered for use as well. Because each unit has a different number of observations, each one has a different number of tests that can be conducted on it. As such, we set the horizon at a variety of levels consistent with possible machine lifespans in order to determine the proper step-wise adjustment.

In contrast to our first study, we do not know the type of change being implemented upon the system, or even when it occurs.

2. Performance Results

Figure 20 displays the power achieved when the overall test level, α^*, is fixed at 0.05 with varied horizon settings and is representative of the graph seen for other values of α^*. The lowest power appears in the left tail where $H = 125$. From there, the graph sharply rises to a power fraction close to 1.0. This demonstrates that if the selected horizon is sufficiently long and within the typical operating profile of the machine, that is to say not overly long, then detection of change prior to system failure is virtually certain. If the selected horizon is too short, then change detection power is adversely affected. Figure 21 provides another perspective on the power fraction, where its values are shown for different horizon settings graphed and different

test levels. The perspective reinforces the property seen in Figure 20 that if the horizon setting is sufficiently long and within the typical operating profile of the machine, then the power fraction will remain near or at its maximum for all reasonable values of α^*.

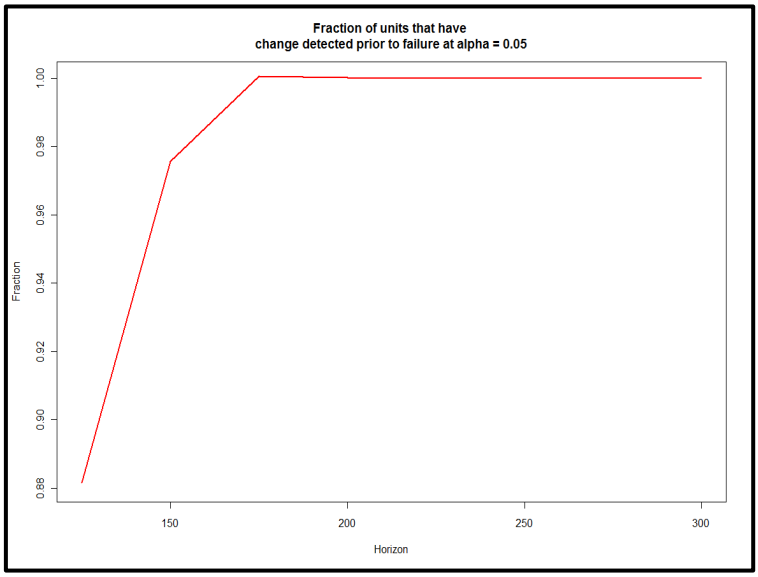

Figure 20: Fraction of units for which change is detected before failure by new test on PHM data for varying horizon settings

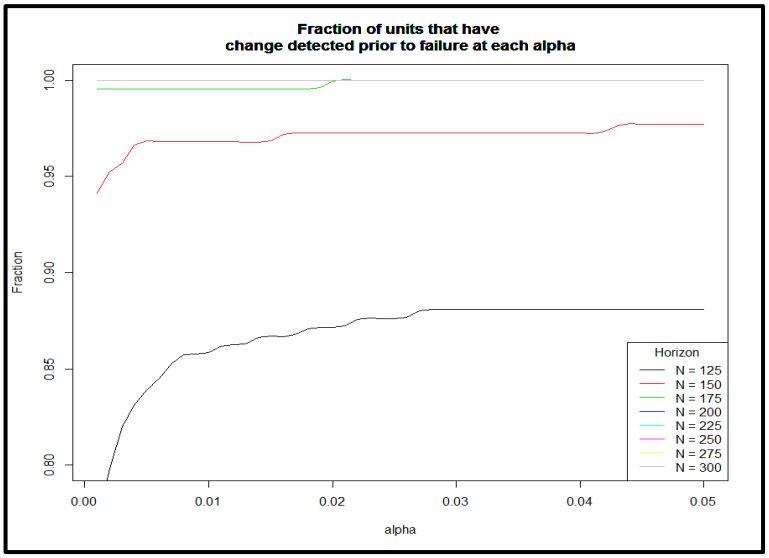

Figure 21: Average power fraction given by new test on PHM data when varying the alpha level for different horizon settings

Figure 22 presents a three-dimensional look at the advanced warning provided when both horizon settings and the overall test level is varied. It demonstrates that choosing a horizon setting that is too short will not only reduce power, but also severely cut down on advance warning regardless of the overall test level. Similar to graphs of the power fraction, a plateau appears as the horizon setting is increased, which further suggests that it is better to use an excessively long horizon within the operating profile rather than a short one. Within this plateau though, there is a distinct ridge which sits above all other horizon settings as α^* is varied. This suggests an optimal horizon setting for these machines and that similar analysis might be performed for other scenarios to determine appropriate optimal horizon settings for such cases.

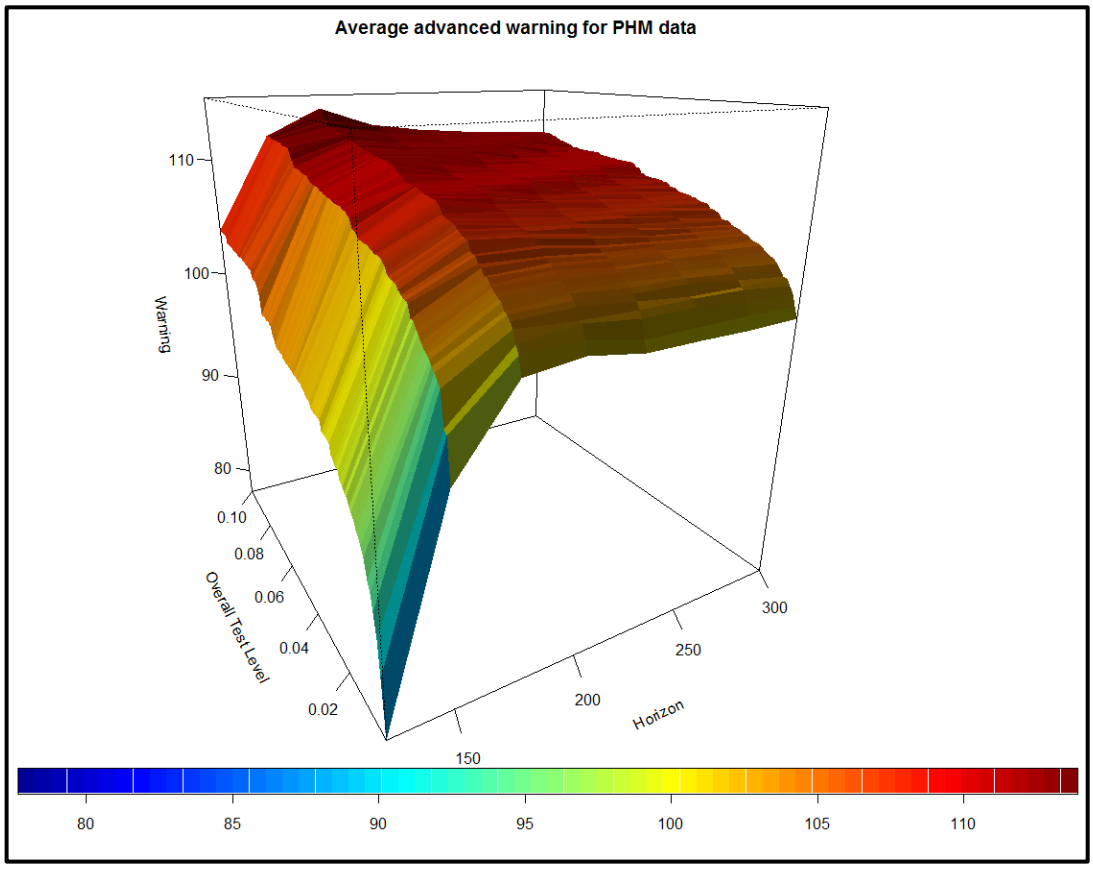

Figure 22: Average advance warning given by new test on PHM data when varying both the horizon setting and overall test level

As a whole, with the power fraction near or at 100% for optimal horizon settings regardless of α^* and advanced warning of at least 75 observations for the worst and 110 for the

best combination of α^* and horizon setting, this online extension shows great potential for successful application in real-world scenarios.

V. CONCLUSIONS & OPPORTUNITIES FOR FUTURE WORK

In this paper, we build upon the offline ESPM change-detection test of Ruth and Koyak (2011) and introduce a new multiple testing-based approach to online multidimensional change-point problems. This approach results in an effective online change detection procedure and portends great promise for future real-world applications. Our review of the field of change detection shows this to be an active area of research with many potential applications and that multivariate nonparametric approaches, let alone online approaches, are few. Most existing approaches require restrictive distributional assumptions which often limit real-world applicability.

The online extension we propose satisfies two primary requirements of multiple testing: 1) maintaining overall test level across many true null hypotheses and 2) achieving reasonable power against false null hypotheses. Assuming that the dependency conditions of Benjamini and Yekutieli (2001) are met, our test meets the first requirement by making use of the Benjamini-Hochberg Procedure (1995) to set a step-wise adjustment of test levels based on the number of hypotheses being tested and the desired overall test level. This procedure is designed to control false discovery rate, although in the setting of our interest it also controls family-wise error rate. To address the second requirement, we employ an existing, powerful, offline test – the ESPM test – and introduce a method of ingesting and testing data incrementally which we refer to as 'telescope' testing, where the testing region begins with an initial sequence of observations and then grows in size as new observations are added to the data set and incorporated into the testing region. This approach demonstrates that our proposed test maintains the desired overall test level while achieving impressive power and useful advanced warning times in many scenarios. Furthermore, our method of extending offline tests to online is not limited to extensions of the ESPM test only, and so we believe this has promise for adapting other powerful offline techniques to online scenarios.

With this project coming to its conclusion, the body of work invites several possibilities for further exploration and extension. One opportunity consists of finding an exact null distribution for the ESPM test statistic, K_N. While our simulation study and PHM data study demonstrate the efficacy of this test, the absence of a known null distribution limits the widespread use of this statistic. Two possible acceptable alternatives to finding an exact

distribution would be 1) finding a result which bounds tail distribution probabilities for K_N, or 2) finding permutation-test type results that could be applied to the ESPM test (and thus directly extended to our test).

Another opportunity lies in theoretical work in finding faster solutions to the optimal non-bipartite matching problem, which is an active area of optimization research. For an ensemble test that requires $N/2$ MNBMs to be computed, the problem of speed is very real. While MNBMs may be found in polynomial time, they are found far less efficiently than other minimum-weight subgraphs mentioned in this paper such MSTs and bipartite matchings. Ruth (2009) reports run times for MNBM ensembles created using Derigs' (1998) algorithm on the order of N^4; our experiences confirm this to be true. This is problematic for an online test like ours in cases where N is very large and data are sampled at a very high rate. More efficient, possibly even suboptimal, matching techniques would alleviate this issue; therefore, it would be useful to study possible benefits and costs of using computationally cheaper graph-theoretic approaches.

Also, our review of change detection methods briefly mentions the use of different cost functions. Often the sample space of interest naturally suggests some appropriate dissimilarity metric and analysis proceeds. The ESPM test makes use of Euclidean distance as a measure of dissimilarity, but a different cost function will affect detection power against alternate hypotheses when using ensemble methods. For real-world applications, it is worth examining which cost functions lead to the most desirable power and advanced warning characteristics for the case on hand.

Finally, since our real-world-type scenario used PHM data from an international contest for statisticians and engineers in 2008, our results regarding choice of window size are only strictly applicable to units of the type in that contest (although broad observations from that study likely transfer to other scenarios). It would be interesting and useful to apply our approach to other real-world scenarios. Areas such as image analysis, machine health diagnosis and prognosis, biosurveillance, and quality control have readily available data and provide excellent opportunities for future work.

GLOSSARY

Acyclic Graph – a graph that has no cycles.

Adjacent Edges – two distinct edges that share a vertex such as $\{v_1, v_2\}$ and $\{v_2, v_3\}$.

Adjacent Vertices – two distinct vertices, v_1 and v_2, that are joined by an edge $\{v_1, v_2\}$.

Bipartite Matching – a matching where the vertices are split into two unique subsets, S_1 and S_2, and each edge included must contain a vertex from each subset.

Circuit – a closed trail that includes at least three distinct vertices.

Closed Walk – a sequence of vertices in graph G beginning with u and ending with v such that consecutive vertices within the sequence are adjacent and $u = v$.

Complete Graph – a graph G in which all vertices are adjacent.

Connected Graph – a graph G in which there is a $u - v$ walk for every pair of vertices.

Cycle – a circuit that repeats no vertex except for the first one equaling the last.

Degree – the number of edges incident with vertex v_1.

Directed Graph – a graph G in which edges have a direction associated with them making edge $\{v_1, v_2\}$ distinct from edge $\{v_2, v_1\}$.

Graph – an ordered pair $G = (V, E)$ consisting of a finite nonempty set of vertices V connected by edges e, which are two-element unordered subsets of V.

Graph Weight – the sum of all edge weights in a weighted graph G.

Incident – a vertex and an edge that meet such as vertex v_1 and edge $\{v_1, v_2\}$.

Independent Edges – a subset of edges $E_1 \in E$ where no two edges are adjacent.

Matching – an independent set of edges in a graph G.

Maximum Matching – a matching that has at least as many edges as any other potential matching in G.

Minimum Spanning Tree – the spanning tree of the weighted graph G whose weight is the least among all possible spanning trees.

Nonbipartite Matching – a matching where each edge included does not depend on any previous partitioning of the vertices (a vertex can be paired with any vertex other than itself).

Perfect Matching – a matching that includes every vertex in G.

Subgraph – a graph $G_1 = (V_1, E_1)$ of $G = (V, E)$ if $V_1 \in V$ and $E_1 \in E$.

Spanning subgraph – a graph $G_1 = (V_1, E_1)$ of $G = (V, E)$ if $V_1 = V$ and $E_1 \in E$.

Spanning Tree – a spanning subgraph of a graph G that is also a tree.

Trail – a walk in which no edge is used more than once.

Tree – a graph G which is both acyclic and connected.

Undirected Graph – a graph in which the edges have no orientation- that is, edge $\{v_1, v_2\}$ is identical to $\{v_2, v_1\}$

Walk – a sequence of vertices in graph G beginning with u and ending with v such that consecutive vertices within the sequence are adjacent.

Weighted Graph – a graph G where there is a real number expressing some form of interpoint cost assigned to each edge in G.

LIST OF REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B,* 57 (1), 289–300.

Benjamini, Y., and Yekutieli, D. (2001), "The Control of False Discovery Rate in Multiple Testing under Dependency," *The Annals of Statistics*, 29 (4), 1165–1188.

Derigs, U. (1988), "Solving Non-Bipartite Matching Problems via Shortest Path Techniques," *Annals of Operations Research*, 13, 225–261.

Devore, J. L. (2004), "A Confidence Interval for a Population Proportion," *Probability and Statistics for Engineering and the Sciences*, 6th ed., 294-296.

Friedman, J., and Rafsky, L. (1979), "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests," *The Annals of Statistics,* 7 (4), 697–717.

Henze, N., and Penrose, M. (1999), "On the Multivariate Runs Test," *The Annals of Statistics,* 27 (1), 290–298.

Lu, B. and Rosenbaum, P. R. (2004), "Optimal Pair MatchingWith Two Control Groups," *Journal of Computational and Graphical Statistics*, 13 (2), 422– 434.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. (2001), "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse," *Journal of the American Statistical Association,* 96, 1245–1253.

PHM Data Challenge Competition (2008). http://www.phmconf.orgjOCS/index.php/phm/2008/challenge [last accessed 20 May 2009]

Rosenbaum, P. (2005), "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency," *Journal of the Royal Statistical Society, Ser. B,* 67 (4), 515–530.

Ruth (2009), "Applications of Assignment Algorithms to Nonparametric Tests for Homogeneity," unpublished Ph.D. thesis, Naval Postgraduate School, Dept. of Operations Research.

Ruth, D., and Koyak, R. (2011), "Nonparametric Tests for Homogeneity Based on Non-Bipartite Matching," *Journal of the American Statistical Association*, 106 (496), 1615 - 1625.