Sequential Change-point Detection for High-dimensional and non-Euclidean Data

Lynna Chu and Hao Chen

Abstract—**In many applications, it is often of practical and scientific interest to detect anomaly events in a streaming sequence of high-dimensional or non-Euclidean observations. We study a non-parametric framework that utilizes nearest neighbor information among the observations to detect changes in an online setting. It can be applied to data in arbitrary dimension and non-Euclidean data as long as a similarity measure on the sample space can be defined. We consider new test statistics under this framework that can detect anomaly events more effectively than the existing test while keeping the false discovery rate controlled at a fixed level. Analytic formulas approximating the average run lengths of the new approaches are derived to make them fast applicable to modern datasets. Simulation studies are provided to support theoretical results. The proposed approach is illustrated with an analysis of the NYC taxi dataset.**

Index Terms—**Anomaly detection; Online change-point; Streaming data; Graph-based tests; Non-parametric.**

I. Introduction

SEQUENTIAL change-point detection aims to detect abrupt anomalies, observations that deviate from regular behavior, in a streaming sequence of observations as quickly as possible, while controlling the number of false alarms. In many modern applications, the sequence of observations may consist of high-dimensional observations (where the dimensional of each observation is larger than the sample size) or non-Euclidean objects (for example, sequences of networks or images). Examples include fraud detection involving large amounts of customer transactions [1]; disease surveillance or medical monitoring using sequences of images or multiple diagnostic measures [2], [3]; and consumer-based data streams such as wearable and smart device monitoring, internet of things sensors, network security systems, cybersecurity, and other web applications [4]–[6]. In all these examples, the goal is to detect a change (if a change is present) as quickly as possible once it occurs, while also limiting the risk of false discovery.

The sequential change-point setting can be formulated as follows: let the observation at t be denoted as $\mathbf{Y}_t,\ t = 1, 2, \dots, n, \dots$. Here, t could be the time index or some other meaningful indices, $\mathbf{Y}_t$ could be a vector, image, or network, and n is the index for the observation currently being observed. When there is no change-point, $\mathbf{Y}_t$'s are identically distributed from an unknown distribution, denoted as F_0. If there is a change-point at τ, the observation after τ are from a different (unknown) distribution:

$$\mathbf{Y}_t \sim F_0,\ t = 1, \dots, \tau - 1,$$

$$\mathbf{Y}_t \sim F_1,\ t = \tau, \tau + 1, \dots,$$

where F_0 and F_1 are two different probability measures.

This formulation is very general. F_0 and F_1 are unknown and not specified. We do not impose any constraints on how the change happens here. For example, if the observation is a high-dimensional vector, the change may occur in a subset of (unknown) data streams and the subset may be of size one.

Our aim is to construct a stopping rule (denoted as T) that will alert us to a change as quickly as possible, once it occurs, but keep the number of false discoveries at a fixed level. Moreover, in keeping with the spirit of the problem formulation, the stopping rule should be able to handle modern data types and should be relatively fast and easy to implement. To be precise, this means we would like to define T such that the detection delay, $E_\tau(T - \tau | T > \tau)$, is small, subject to a fixed average run length (formally $E_\infty(T) \geq c$, where c is a pre-specified large value) without making assumptions on the underlying sequence of observations. Here E_τ denotes the expectation under the hypothesis that the true change-point happens at τ and E_∞ denotes the expectation under the hypothesis of no change.

A. Related Works

When the data is univariate (or scalar), the sequential change-point detection has been studied extensively (see [7] and [8] for a review). For low-dimensional data, likelihood based methods have been explored which require knowledge or parameter estimates of the probability density functions (see for example: [9]–[12]). For high-dimensional data, the available methods are somewhat limited and may impose strict assumptions on the data. As an example, many existing methods have the assumption that the different data streams are independent [13]–[17]. Other works allow for more flexible application; for example kernel-based methods [18] and a modified sliding window algorithm [19]. Computationally efficient methods have also been proposed which combine summary statistics based on geometric entropy minimization (GEM) with the cumulative sum (CUSUM) algorithm [20], [21]. For network data, [22] proposed a sequential approach that acts by embedding each graph into a vector domain, where a conventional multivariate change-point detection procedure can be then applied. In general, for non-parametric methods

The work was supported in part by the NSF awards DMS-1513653 and DMS-1848579.

Lynna Chu is with the Department of Statistics, Iowa State University (email: lchu@iastate.edu).

Hao Chen is with the Department of Statistics, University of California, Davis (email: hxchen@ucdavis.edu).

applicable to high-dimensional and non-Euclidean data, theoretical analysis establishing false discovery control is very difficult to carry out.

Recently, [23] proposed a new non-parametric framework that utilizes nearest neighbor information to detect changes in an online setting. They also provided a general, analytical formula for false discovery control. This method can be applied to data in arbitrary dimensions (with no assumption that the different data streams are independent) and to non-Euclidean data. The author proposed to use the following stopping rule:

$$T_Z(b_Z) = \inf\left\{n - N_0 : \max_{n-n_1\le t\le n-n_0} Z_{L|\mathbf{y}}(t,n) > b_Z\right\}, \tag{1}$$

where n_0, n_1, and L are pre-specified values, N_0 is the number of historical observations with no change-point, $n > N_0$, and $Z_{L|\mathbf{y}}(t,n)$ is a two-sample test statistic that tests whether $\{\mathbf{Y}_{n-L+1},\ldots,\mathbf{Y}_t\}$ and $\{\mathbf{Y}_{t+1},\ldots,\mathbf{Y}_n\}$ are from the same distribution. We refer to $Z_{L|\mathbf{y}}(t,n)$ as the *edge-count two-sample test based on k-nearest neighbor* (k-NN) in the following for simplicity. For more details of this test, please see Section II. The author also provided an analytic formula to compute b_Z such that the average run length $\mathsf{E}_\infty(T_Z(b_Z))$ is controlled at a pre-determined value. Simulation studies show that this method beats likelihood-based methods when the dimension is high.

Despite these nice properties, we find that the edge-count two-sample test on k-NN can have low power for some common types of changes when dimension is moderate to high, causing the stopping rule (1) to behave unexpectedly. To illustrate, consider a simple scenario where data are from a d-dimensional Gaussian distribution and there is a change at $\tau = 201$:

$$\mathbf{Y}_1,\ldots,\mathbf{Y}_{200} \overset{iid}{\sim} \mathcal{N}_d(\mathbf{0},\Sigma), \quad \mathbf{Y}_{201},\ldots \overset{iid}{\sim} \mathcal{N}_d(\mu,\sigma^2\Sigma)$$

with $\Sigma(i,j) = 0.3^{|i-j|}$. We consider two types of changes:

- Scenario 1 (only mean differs): $||\mu||_2 = \Delta$.
- Scenario 2 (both mean and variance differ): $||\mu||_2 = \Delta$ and σ.

Tables I presents the performance of $T_Z(b_Z)$ for both scenarios based on 1,000 simulation runs. Here $k = 5$, $L = 200$, $n_0 = 25$, and $n_1 = 175$. The table reports the fraction of trials (out of 1,000) to successfully detect the change within 30 (or 50) observations after the change occurs. The average detection delay (EDD) is estimated as the average elapsed time between when the change occurs ($\tau = 201$) and when $T_Z(b_Z)$ detects a change. False alarms are not counted here. In each scenario, the threshold b_Z is computed by formulas given in [23] such that $\mathsf{E}_\infty(T_Z(b_Z)) = 2000$.

In Table I, we see that the method performs worse in Scenario 2 than Scenario 1 under two different comparison criteria: the fraction of trials that can be detected given a fixed time is smaller under Scenario 2 and the average detection delay (EDD) is larger under Scenario 2. However, common sense tells us that the additional change in variance should make the two distributions more different and the change in Scenario 2 easier to detect. This phenomenon is due to the curse-of-dimensionality (a more detailed explanation can be found in Section II-B) and results in the existing method having diminished power to detect general changes.

TABLE I
PERFORMANCE OF $T_Z(b_Z)$ FOR SCENARIOS 1 AND 2, $\Delta = 2.6$, $\sigma = 0.78$, $d = 100$

	Scenario 1	Scenario 2
< 30	0.15	0.06
< 50	0.67	0.48
EDD	44.29 ± 14.17	50.21 ± 13.96

B. Our contributions

To address the problem of the stopping rule $T_Z(b_Z)$, we propose three new stopping rules:

$$T_S(b_S) = \inf\left\{n - N_0 : \max_{n-n_1\le t\le n-n_0} S_{L|\mathbf{y}}(t,n) > b_S\right\}, \tag{2}$$

$$T_W(b_W) = \inf\left\{n - N_0 : \max_{n-n_1\le t\le n-n_0} W_{L|\mathbf{y}}(t,n) > b_W\right\}, \tag{3}$$

$$T_M(b_M) = \inf\left\{n - N_0 : \max_{n-n_1\le t\le n-n_0} M_{L|\mathbf{y}}(t,n) > b_M\right\}. \tag{4}$$

The definitions of $S_{L|\mathbf{y}}(t,n)$, $W_{L|\mathbf{y}}(t,n)$, and $M_{L|\mathbf{y}}(t,n)$ are provided in Sections II-C, II-D and II-E, respectively. Under the same setup detailed in Table I, Table II shows that these new stopping rules are more successful in detecting the change quickly after it has occurred. They also have shorter detection delays than $T_Z(b_Z)$ under both above scenarios and all have shorter detection delays in Scenario 2 than that in Scenario 1. Further comparisons between the stopping rules can be found in Section IV; these demonstrate that the new stopping rule have improved power and detection delay compared to existing methods over a range of general scenarios.

TABLE II
PERFORMANCE OF NEW STOPPING RULES UNDER SCENARIO 1 (TOP) AND SCENARIO 2 (BOTTOM), $\Delta = 2.6$, $\sigma = 0.78$, $d = 100$

	Z	W	S	M
< 30	0.15	0.51	0.41	0.49
< 50	0.67	0.91	0.86	0.89
EDD	44.29 ±14.17	32.27 ±12.90	35.14 ±15.48	32.61 ±13.62

	Z	W	S	M
< 30	0.06	0.85	0.81	0.82
< 50	0.48	0.98	0.98	0.98
EDD	50.21 ± 13.96	23.21 ± 8.01	24.16 ± 8.48	23.74 ± 7.82

To construct these new stopping rules, we propose new two-sample tests on k-NN. Specifically, we extend the generalized (S) /weighted (W) /max-type (M) edge-count test defined on an undirected similarity graph [24], [25] to the directed k-NN graph. The generalized edge-count and max-type tests on k-NN are well defined except for a particular construction of a

k-NN graph (see Theorem 1). The detailed definitions of these stopping rules are given in Section II.

To make the new stopping rules useful for real-data applications, we provide analytical formulas to compute the stopping thresholds, b_S, b_W, and b_M, such that the average run length (ARL) for each new stopping rule is controlled at a pre-determined value. This involves studying how the k-NN graph updates and obtaining expressions that explicitly characterize these dynamics, which lead to the development of new theoretical treatments. Specifically, for all the stopping rules, more accurate expressions of the graph updates are derived and the techniques used are an improvement over the approach utilized in [23]. We demonstrate that the analytical formulas for the stopping thresholds are reasonable to use and we further improve upon their accuracy for finite sample sizes by implementing a skewness correction technique on the thresholds.

In general, each test statistic has its own niche where it dominates. When interested in general change (for example, both mean and variance change), the stopping rules $T_S(b_S)$ and $T_M(b_M)$ are recommended. The stopping rule $T_M(b_M)$ has an advantage over $T_S(b_S)$ in that we can obtain more accurate analytical expression for the ARL for false discovery control. If the change of interest is in mean only, the stopping rule based on $T_W(b_W)$ is recommended. See Section IV for a comparison of their performance. In this paper, the types of changes we explore are confined to mean and/or variance change. However, the approach can be used to detect other changes in distribution, such as changes in covariance. For illustration, Table III shows the performance of the stopping rules under covariance change only. The data are again generated from d-dimensional Gaussian with $d = 100$, $k = 5$, $L = 200$, $n_0 = 25$, $n_1 = 175$, and there is change at $\tau = 201$:

$$\mathbf{Y}_1, \dots, \mathbf{Y}_{200} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_0), \quad \mathbf{Y}_{201}, \dots \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_1),$$

with $\Sigma_0(i,j) = 0.3^{|i-j|}$ and $\Sigma_1(i,j) = 0.68^{|i-j|}$.

TABLE III
PERFORMANCE OF STOPPING RULES UNDER COVARIANCE CHANGE.

	Z	W	S	M
< 30	0.026	0.263	0.336	0.334
< 50	0.069	0.603	0.677	0.663
EDD	92.17 ±27.69	65.65 ±30.77	43.90 ±25.85	45.27 ±26.50

We see here again that the new stopping rules based on W, S, and M perform better than the stopping rule based on Z.

These new approaches are implemented in an `R` package `gStream`.

The organization of the rest of the paper is as follows. Section II discusses the new stopping rules in details. Section III studies the asymptotic properties of the proposed stopping rules and analytic ways to determine the thresholds. The performance of the new methods are further explored in Section IV and the new methods are illustrated on a real data set in Section V.

II. New Tests

The test statistics in the stopping rules (2) - (4) trace from the offline version of the problem studied in [25] where the statistics were defined on an undirected similarity graph. For online detection, observations keep arriving and the similarity graph updates as a new observation arrives. Therefore one needs to understand the dynamics of the series of similarity graphs. [23] studied the directed nearest neighbor graphs as the dynamics of nearest neighbor graphs can be well understood. In this work, we continue to use the directed nearest neighbor graphs to construct improved statistics. The extension to other types of graphs is saved for future work.

A. Notation

First we define a random variable which indicates whether or not an observation $\mathbf{Y}_i$ is among the rth nearest neighbor to another observation $\mathbf{Y}_j$ among the observations in n_L. Specifically, for any $n > N_0$ and $i, j \in n_L \triangleq \{n-L+1, \dots, n\}$, we let $A^{(r)}_{n_L,ij} = \mathrm{I}(\mathbf{Y}_j \text{ is the } r\text{th NN of } \mathbf{Y}_i \text{ among } \mathbf{Y}_{n-L+1}, \dots, \mathbf{Y}_n)$, where $\mathrm{I}(\cdot)$ is the indicator function that takes value 1 if the event is true and 0 otherwise. In terms of graph construction, each observation points to its k nearest neighbors. For example, if $A^{(r)}_{n_L,ij} = 1$, then $\mathbf{Y}_j$ is the rth nearest neighbor of $\mathbf{Y}_i$ and there is a directed edge from $\mathbf{Y}_i$ pointing to $\mathbf{Y}_j$ (if $r \le k$). We define $A^{+}_{n_L,ij} = \sum_{r=1}^{k} A^{(r)}_{n_L,ij}$ to be the indicator function that $\mathbf{Y}_j$ is one of the first k NNs of $\mathbf{Y}_i$ among the observations in n_L. We use $\mathbf{y}_i$'s to denote the realizations of $\mathbf{Y}_i$'s and let $a^{+}_{n_L,ij} = \sum_{r=1}^{k} a^{(r)}_{n_L,ij}$ with $a^{(r)}_{n_L,ij} = \mathrm{I}(\mathbf{y}_j \text{ is the } r\text{th NN of } \mathbf{y}_i \text{ among } \mathbf{y}_{n-L+1}, \dots, \mathbf{y}_n)$. For any n, each $t \in \{n-L+1, \dots, n\}$ divides the data sequence into two groups: one group being the observations before $t : \{\mathbf{Y}_{n-L+1}, \dots, \mathbf{Y}_t\}$ (Group 1) and the other group being the observations after $t : \{\mathbf{Y}_{t+1}, \dots, \mathbf{Y}_n\}$ (Group 2). Define,

$$\begin{aligned} b_{0,ij}(t, n_L) =& \mathrm{I}(n-L+1 \le i \le t \text{ and } t < j \le n \text{ or} \\ & t < i \le n \text{ and } n-L+1 \le j \le t), \\ b_{1,ij}(t, n_L) =& \mathrm{I}(n-L+1 \le i \le t \text{ and } n-L+1 \le j \le t), \\ b_{2,ij}(t, n_L) =& \mathrm{I}(t < i \le n \text{ and } t < j \le n). \end{aligned}$$

Then $b_{0,ij}$ is the indicator function that $\mathbf{Y}_i$ and $\mathbf{Y}_j$ belong to different groups, $b_{1,ij}$ is the indicator function that $\mathbf{Y}_i$ and $\mathbf{Y}_j$ both belong to Group 1, and $b_{2,ij}$ is the indicator function that $\mathbf{Y}_i$ and $\mathbf{Y}_j$ both belong to Group 2.

We define our test statistics as follows:

$$R_{0,L}(t,n) = \sum_{i=n-L+1}^{n} \sum_{j=n-L+1}^{n} (A^{+}_{n_L,ij} + A^{+}_{n_L,ji}) B_{0,ij}(t, n_L),$$

$$R_{1,L}(t,n) = \sum_{i=n-L+1}^{n} \sum_{j=n-L=1}^{n} (A^{+}_{n_L,ij} + A^{+}_{n_L,ji}) B_{1,ij}(t, n_L),$$

$$R_{2,L}(t,n) = \sum_{i=n-L+1}^{n} \sum_{j=n-L=1}^{n} (A^{+}_{n_L,ij} + A^{+}_{n_L,ji}) B_{2,ij}(t, n_L).$$

$B_{0,ij}(t, n_L)$, $B_{1,ij}(t, n_L)$, and $B_{2,ij}(t, n_L)$ are the random variable versions of $b_{0,ij}(t, n_L)$, $b_{1,ij}(t, n_L)$, and $b_{2,ij}(t, n_L)$ such that the distribution of these random variables is defined to be the permutation distribution. The permutation distribution is the distribution induced by all $L!$ possible permutations of observation *indices* among the observations in n_L. The null hypothesis and independence assumption imply that the observation indices are exchangeable, and therefore under the permutation distribution, the true (unknown) distribution of the observations remains unchanged when the null hypothesis is true.

It is clear that $R_{0,L}(t,n)$ is twice the number of edges in the k-NN graph connecting observations before t and after t, $R_{1,L}(t,n)$ is twice the number of edges connecting observations prior to t, and $R_{2,L}(t,n)$ is twice the number of edges that connect observations after t. The notation of the graph-based test quantities emphasizes their dependency on the graph which is constructed on the L most recent observations, with the most recent observation indexed at n. Figure 1 illustrates these test statistics constructed for different times t and n. In the top row of Figure 1, $n = 30$ and we construct the graph on the $L = 20$ most recent observations: $y_{11}, y_{12}, \dots, y_{30}$. The edge-counts are calculated for different values of $t \in \{11, \dots, 30\}$. $R_{0,L}(t,n)$ is twice the number of directed black edges, $R_{1,L}(t,n)$ is twice the number of directed red edges, and $R_{2,L}(t,n)$ is twice the number of directed blue edges. It is clear for a fixed n, the graph does not change: each graph in the first row of Figure 1 is the same. But as new observations continually arrive, the graph itself will update. For example when $n = 40$, we now construct the graph on observations $y_{21}, y_{22}, \dots, y_{40}$ (see the second row of Figure 1). For $n = 40$, the edge-counts are then calculated for different values of $t \in \{21, \dots, 40\}$.

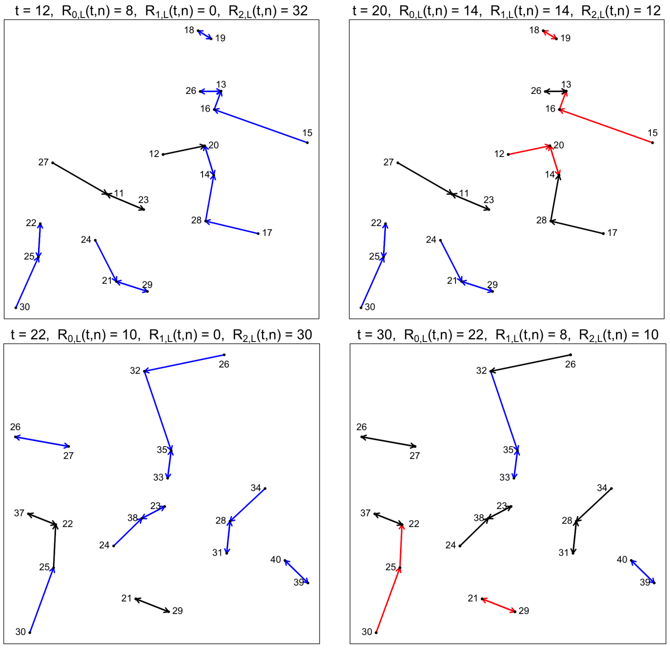
t = 12, R0,L(t,n) = 8, R1,L(t,n) = 0, R2,L(t,n) = 32
t = 20, R0,L(t,n) = 14, R1,L(t,n) = 14, R2,L(t,n) = 12
t = 22, R0,L(t,n) = 10, R1,L(t,n) = 0, R2,L(t,n) = 30
t = 30, R0,L(t,n) = 22, R1,L(t,n) = 8, R2,L(t,n) = 10

Fig. 1. The construction of graph-based test quantities $R_{0,L}(t,n)$ (black edges), $R_{1,L}(t,n)$ (red edges), and $R_{2,L}(t,n)$ (blue edges) for different values of t on a directed k-NN graph. Each observation points to its k nearest neighbors. Here $k = 1$. In the first row, $n = 30$ and the k-NN graph is constructed on the $L = 20$ most recent observations: $y_{11}, y_{12}, \dots, y_{30}$. In the second row, $n = 40$ and the k-NN graph is constructed on the $L = 20$ most recent observations: $y_{21}, y_{22}, \dots, y_{40}$.

B. Limitations of the method based on the edge-count test (Z)

The stopping rule $T_Z(b_Z)$ (1) is based on the edge-count two-sample test statistic $Z_{L|y}(t,n)$ [23]. To obtain $Z_{L|y}(t,n)$, first a directed k-NN graph is constructed based on a similarity measure (for example, Euclidean distance). Then the number of edges in the k-NN graph that connect observations before t ($\mathbf{Y}_{n-L+1}, \dots, \mathbf{Y}_t$) and after t ($\mathbf{Y}_{t+1}, \dots, \mathbf{Y}_n$) is counted (we refer to this as the between-sample edge-count). Please see Figure 1 for an illustration of how $R_{0,L}(t,n)$ is computed. To make $R_{0,L}(t,n)$ comparable across different t, we define its standardized version as $Z_{L|y}(t,n) = -\frac{R_{0,L}(t,n) - \mathsf{E}(R_{0,L}(t,n))}{\sqrt{\text{Var}(R_{0,L}(t,n))}}$. Analytical expressions of $\mathsf{E}(R_{0,L}(t,n))$ and $\text{Var}(R_{0,L}(t,n))$ can be obtained under the permutation distribution, which is defined as all $L!$ possible rearrangements of the observation indices among n_L, and are omitted here for brevity. A relatively low between-sample edge-count (or a large $Z_{L|y}(t,n)$) indicates the observations before and after t are less mixed and this is evidence against the null hypothesis of no change. The intuition is that observations before t tend to find their nearest neighbor among other observations before t (and similarly for observations after t), which implies a distributional difference between the two groups of observations.

The rationale of a relatively small between sample edge-count ($R_{0,L}(t,n)$)) holds particularly well under the scenario of mean change only and/or low-dimensional data. However, when the dimension of the data is moderate/high and the change in distribution is not only in mean, for example an additional variance change is present, this rationale breaks down. This is because in the high-dimensional setting observations that are similar (i.e. from the same distribution) are not necessarily close in distance. Consider two distributions that differ in both mean and variance and the dimension of the data is $d = 100$. The observations would be separated into two layers: observations with the smaller variance in the inner layer and observations with the larger variance in the outer layer. Since the volume of a d-dimensional space increases exponentially in d, for practical sample sizes it is not uncommon for observations in the outer layer to find themselves to be closer to observations in the inner layer compared to other observations in the outer layer. In this scenario the between-sample edge-count would be relatively large under the alternative, rendering $Z_{L|y}(t,n)$ ineffective. The new two-sample test statistics $S_{L|y}(t,n)$ and $M_{L|y}(t,n)$ address this curse-of-dimensionality problem.

Moreover, even for mean change only the stopping rule based on $Z_{L|y}(t,n)$ can still suffer from increased detection delay. This limitation of $Z_{L|y}(t,n)$ is due to a variance boosting problem under unequal sample sizes in the two-sample test setting. For more details on the variance boosting issue in the two-sample test setting see [26]. In the sequential setting, this means that using $Z_{L|y}(t,n)$ can lead to increased detection delay since it may only be able to detect the change when it is near the middle of the sequence. To resolve this, we propose the weighted edge-count two-sample test $W_{L|y}(t,n)$.

C. The method based on the generalized edge-count test (S)

The generalized edge-count two-sample test at $t \in \{n-L+1, \dots, n\}$ under k-NN can be defined as

$$S_{L|\mathbf{y}}(t,n) = \begin{pmatrix} \bar{R}_{1,L}(t,n)) \\ \bar{R}_{2,L}(t,n)) \end{pmatrix}^T (\mathbf{\Sigma}_{\mathbf{y}}(t,n))^{-1} \begin{pmatrix} \bar{R}_{1,L}(t,n) \\ \bar{R}_{2,L}(t,n) \end{pmatrix}, \tag{5}$$

where $\bar{R}_{1,L}(t,n) = R_{1,L}(t,n) - \mathsf{E}(R_{1,L}(t,n))$, $\bar{R}_{2,L}(t,n) = R_{2,L}(t,n) - \mathsf{E}(R_{2,L}(t,n))$, and $\mathbf{\Sigma}_{\mathbf{y}}(t,n) = \text{Var}((R_{1,L}(t,n), R_{2,L}(t,n))^T|\mathbf{y})$ such that $\mathsf{E}(\cdot)$ and $\text{Var}(\cdot)$ denote the expectation and variance taken under the permutation distribution.

If a change-point $\tau > N_0$ occurs in the sequence, we would expect $S_{L|\mathbf{y}}(t,n)$ to be large when $n > \tau$ and t close to τ. The test statistic is defined in this way so that either direction of deviations in the number of within-sample edges from its null expectation would contribute to the test statistic. For example, under the location alternatives, we would expect both $R_{1,L}(t,n)$ and $R_{2,L}(t,n)$ to be larger than their null expectations, which would lead to a large $S_{L|\mathbf{y}}(t,n)$. Under the scale alternatives, the group with the smaller variance would have a within-edge count larger than its null expectation and the group with a larger variance would have a within-edge count smaller than its null expectation (due to the curse-of-dimensionality), which would also lead to a large $S_{L|\mathbf{y}}(t,n)$. Therefore, this test is powerful for both location and scale alternatives.

Under the permutation distribution, the analytical expressions for $\mathsf{E}(R_{1,L}(t,n)|\mathbf{y})$, $\mathsf{E}(R_{2,L}(t,n)|\mathbf{y})$, and $\mathbf{\Sigma}_{\mathbf{y}}(t,n) = (\Sigma_{i,j}(t,n)|\mathbf{y})_{i,j=1,2}$ can be calculated through combinatorial analysis. Note that $\mathsf{E}(R_{1,L}(t,n)|\mathbf{y}) = \mathsf{E}(R_{1,L}(t,n))$ and $\mathsf{E}(R_{2,L}(x,n)|\mathbf{y}) = \mathsf{E}(R_{2,L}(t,n))$. Let $x = t-(n-L)$.

$$\begin{aligned}
&\mathsf{E}(R_{1,L}(x,n)) = \tfrac{2kx(x-1)}{(L-1)}, \\
&\mathsf{E}(R_{2,L}(t,n)) = \tfrac{2k(L-x)(L-x-1)}{(L-1)}, \\
&\Sigma_{\mathbf{y},11}(x,n) = \tfrac{4x(x-1)(L-x)}{(L-1)(L-2)(L-3)}((L-x-1)\times \\
&(k + \tfrac{1}{L}\sum_{i,j\in n_L} a^+_{ij}a^+_{ji}) + (x-2)\tfrac{1}{L}\sum_{i,j\in n_L} a^+_{ji}a^+_{li} - \tfrac{k^2x(L-3)}{(L-1)}), \\
&\Sigma_{\mathbf{y},22}(x,n) = \tfrac{4x(L-x)(L-x-1)}{(L-1)(L-2)(L-3)}((x-1)(k + \tfrac{1}{L}\sum_{i,j\in n_L} a^+_{ij}a^+_{ji}) \\
&\qquad + (L-x-2)\tfrac{1}{L}\sum_{i,j,l\in n_L} a^+_{ji}a^+_{li} - \tfrac{k^2(L-x)(L-3)}{(L-1)}), \\
&\Sigma_{\mathbf{y},12}(x,n) = \Sigma_{\mathbf{y},21}(x,n) = \tfrac{4x(x-1)(L-x)(L-x-1)}{(L-1)(L-2)(L-3)}(\tfrac{k^2(L-3)}{L-1} \\
&\qquad + k + \tfrac{1}{L}\sum_{i,j\in n_L} a^+_{ij}a^+_{ji} - \tfrac{1}{L}\sum_{i,j\in n_L} a^+_{ji}a^+_{li}).
\end{aligned}$$

The generalized edge-count two-sample test statistic is well defined when $\Sigma_{\mathbf{y}}(t,n)$ is invertible.

The following theorem ensures that the statistic is well defined except in very rare scenarios. These scenarios can be checked by calculating the node in-degree of each observation in the k-NN graph. A node's in-degree is the number of other observations that find that node to be its nearest neighbors.

Theorem 1. *For $L \geq 5$, the generalized edge-count two-sample test statistic under k-NN is well defined except for when all nodes have an in-degree of exactly k, i.e. $d_i = k\ \forall i$, where $d_i = \sum_{j\in n_L} a^+_{ji}$.*

Based on Theorem 1, as long as each node in the graph does not have an in-degree of k and the graph is constructed on at least 5 observations, then $\Sigma_{\mathbf{y}}(t,n)$ is invertible and $S_{L|y}(t,n)$ is well-defined. A proof of the above theorem is in Supplement A.

This leads us to the stopping rule based on the generalized edge-count test under k-NN (2):

$$T_S(b_S) = \inf\left\{ n - N_0 : \max_{n-n_1\leq t\leq n-n_0} S_{L|\mathbf{y}}(t,n) > b_S \right\}.$$

D. The method based on the weighted edge-count test (W)

Following the same notations in Section II-C, for each $t \in \{n-L+1, \dots, n\}$, the weighted edge-count two-sample test statistic under k-NN can be defined as

$$R_{w,L}(t,n) = q(t,n)R_{1,L}(t,n) + p(t,n)R_{2,L}(t,n),$$

where $p(t,n) = \frac{t-(n-L)-1}{L-2}$ and $q(t,n) = 1 - p(t,n)$. Since it is more difficult for the sample with a smaller sample size to form an edge within the same sample, $R_{1,L}(t,n)$ and $R_{2,L}(t,n)$ are weighted by the inverse of their corresponding sample sizes. The test statistic defined in this way resolves the variance boosting problem described in Section II-B. Relatively large values of $R_{w,L}(t,n)$ are evidence against the null hypothesis of no change. Let

$$W_{L|\mathbf{y}}(t,n) = \frac{R_{w,L}(t,n) - \mathsf{E}(R_{w,L}(t,n))}{\sqrt{\text{Var}(R_{w,L}(t,n)|\mathbf{y})}}. \tag{6}$$

Under the permutation distribution, analytical formulas for $\mathsf{E}(R_{w,L}(t,n))$ and $\text{var}(R_{w,L}(t,n)|\mathbf{y})$ can be calculated based on $\mathsf{E}(R_{1,L}(t,n))$, $\mathsf{E}(R_{2,L}(t,n))$, and $\mathbf{\Sigma}_{\mathbf{y}}(t,n)$ provided in Section II-C:

$$\begin{aligned}
&\mathsf{E}(R_{w,L}(t,n)) = \tfrac{2kL(L-n+t-1)(n-t-1)}{(L-1)(L-2)}, \\
&\text{Var}(R_{w,L}(t,n)|\mathbf{y}) \\
&= \tfrac{4(L-n+t)(L-n+t-1)(n-t)(n-t-1)}{(L-1)(L-2)(L-3)}\times \\
&(k + \tfrac{\sum_{i,j\in n_L} a^+_{ij}a^+_{ji}}{L} - \tfrac{\sum_{i,j,l\in n_L} a^+_{ji}a^+_{li}}{L(L-2)} - \tfrac{k^2(L-3)}{(L-1)(L-2)}).
\end{aligned}$$

The variance of $R_{W|L}(t,n)$ is well defined if the inequality (7) holds. Since $\sum_{i,j\in n_L} a^+_{ij}a^+_{ji} \geq 0$ by definition and $\frac{1}{L}\sum_{i,j,l\in n_L} a^+_{ji}a^+_{li} \leq k(L-1)^2 + k^2$, we need:

$$\sum_{i,j\in n_L} a^+_{ij}a^+_{ji} > \tfrac{k(L-2k+Lk-1)}{(L-1)(L-2)} = \tfrac{k}{(L-2)} + \tfrac{k^2}{(L-1)}. \tag{7}$$

The stopping rule based on the weighted edge-count test under k-NN is (3):

$$T_W(b_W) = \inf\left\{ n - N_0 : \max_{n-n_1\leq t\leq n-n_0} W_{L|\mathbf{y}}(t,n) > b_W \right\}.$$

E. The method based on the max-type edge-count test (M)

We can define the max-type test statistic under k-NN based on the following lemma:

Lemma 1. *The generalized edge-count two-sample test under k-NN can be expressed as*

$$S_{L|\mathbf{y}}(t,n) = W^2_{L|\mathbf{y}}(t,n) + D^2_{L|\mathbf{y}}(t,n),$$

where $W_{L|\mathbf{y}}(t,n)$ is defined in (6), and

$$D_{L|\mathbf{y}}(t,n) = \frac{R_{\mathit{diff},L}(t,n) - \mathsf{E}(R_{\mathit{diff},L}(t,n))}{\sqrt{\mathit{Var}(R_{\mathit{diff},L}(t,n)|\mathbf{y})}} \tag{8}$$

with $R_{\mathit{diff},L}(t,n) = R_{1,L}(t,n) - R_{2,L}(t,n)$.

The proof of this lemma is in Supplement B. The analytical expressions for the expectation and variance of $R_{\text{diff},L}(t,n)$ under the permutation null are:

$$\mathsf{E}(R_{\text{diff},L}(t,n)) = 2k(L-2n+2t),$$
$$\text{Var}(R_{\text{diff},L}(t,n)|\mathbf{y}) = \tfrac{4(L-n+t)(n-t)}{(L-1)}\left(\tfrac{1}{L}\sum_{ij\in n_L} a^+_{ji}a^+_{li} - k^2\right).$$

The variance of $R_{\text{diff},L}(t,n)|\mathbf{y}$ is well-defined as long as $d_i \neq k\,\forall\, i$, in other words as long as each node does not have an in-degree of k.

From the above lemma, $S_{L|\mathbf{y}}(t,n)$ is the sum of squares of two uncorrelated quantities (these two quantities are further asymptotically independent; details given in Section III). Here, $W_{L|\mathbf{y}}(t,n)$ is sensitive to location changes: when the change is in mean, $W_{L|\mathbf{y}}(t,n)$ tends to be large. On the other hand, $D_{L|\mathbf{y}}(t,n)$ is more sensitive to scale changes: when the change is in variance, $|D_{L|\mathbf{y}}(t,n)|$ tends to be large. The sign of $D_{L|\mathbf{y}}(t,n)$ depends on whether the distribution after the change has a larger spread or not. This leads to the following max-type edge-count two-sample test statistic under k-NN:

$$M_{L|\mathbf{y}}(t,n) = \max(|D_{L|\mathbf{y}}(t,n)|, W_{L|\mathbf{y}}(t,n)). \tag{9}$$

When there is a change in location and/or scale, depending on the signal of interest, it is useful to consider an extended version of the max-type edge-count two-sample test:

$$M_{\xi,L|\mathbf{y}}(t,n) = \max(|D_{L|\mathbf{y}}(t,n)|, \xi W_{L|\mathbf{y}}(t,n)), \tag{10}$$

where $\xi \geq 0$. Different choices of ξ lead to different focuses of the alternatives. For example, if we are more interested in locational changes, we could choose a large ξ. On the other hand, setting ξ to be small would favor detecting scale changes. When $\xi = 1$, the test reduces to the plain max-type edge-count test. For more detailed discussion on how to select ξ, see Supplement H in [25] (under the offline change-point detection setting, but similar arguments apply to the online setting).

The stopping rule based on the max-type edge-count test under k-NN is as follows :

$$T_{M_\xi}(b_{M_\xi}) = \inf\left\{ n - N_0 : \max_{n_1'\leq t\leq n_0'} M_{\xi,L|\mathbf{y}}(t,n) > b_{M_\xi} \right\}, \tag{11}$$

where $n_1' = n - n_1$ and $n_0' = n - n_0'$. This reduces to (4) when $\xi = 1$.

III. Average run length

Given the new stopping rules presented in Section II, we would like to determine the thresholds b_S, b_W, and b_{M_ξ} in an analytic way such that the false discovery rate is controlled at a pre-specified value. A common way to measure the false discovery rate under the online change-point detection is the average run length, i.e., the expected time to stop when there is no change-point, which we denote as $\mathsf{E}_\infty(T_S(b_S))$, $\mathsf{E}_\infty(T_W(b_W))$, and $\mathsf{E}_\infty(T_{M_\xi}(b_{M_\xi}))$.

In the comparisons in Section I-B (Table II), the thresholds were chosen such that the average run lengths are 2,000 based on simulation runs. This is doable when the underlying distribution of the sequence is known. However, in many applications, the distribution of the sequence is unknown. Furthermore, since new observations keep arriving, resampling based methods, such as permutation and bootstrap, are not appropriate here and even if they were, directly resampling could be very time consuming. Therefore, to make the method fast applicable, we seek to derive analytical expressions for the average run lengths. Given the non-parametric nature of the proposed method, we would not be able to get exact analytic formulas for the average run lengths under finite L. In the following, we first approach the problem asymptotically (Section III-A), and then make adjustments for finite samples (Section III-B).

A. Asymptotic results

To derive analytical expressions for the average run length, we must study the asymptotic distribution of the stopping rules. Since these stopping rules are composed of the random fields $\{S_{L|\mathbf{y}}(t,n_L)\}$, $\{W_{L|\mathbf{y}}(t,n)\}$, and $\{M_{\xi,L|\mathbf{y}}(t,n)\}$, we study their asymptotic properties. To obtain the limiting distribution of these random fields, we only need to focus on $\{D_{L|\mathbf{y}}(t,n)\}$ and $\{W_{L|\mathbf{y}}(t,n)\}$. We show that the limiting distribution of the random fields converge to independent two-dimensional Gaussian random fields (Theorem 2). This proof utilizes Stein's method [27]. To fully specify the Gaussian processes, we must derive the covariance functions of the new processes (Theorem 3). Since the k-NN graph updates each time new observations arrive, we must study the dynamics of the k-NN series for the new test statistics. While [23] laid a framework for graph-based sequential detection, the techniques developed in [23] were not directly applicable to the stopping rules proposed in this manuscript and needed to be adapted. Integration of techniques from [25] were useful in this development. However, since the new stopping rules can cover more types of change, we found that a direct extension of these previous works were not sufficient in developing accurate analytical formulas approximating ARL. To push forward the theory, we carefully studied the dynamics of directed k-NN graphs and considered additional ways the graph can update. A more in-depth understanding of the graph updates led us to derive analytical expressions for these updates and incorporate them into our analytical formulas for the thresholds. A comparison of the improvement over a direct extension of previous works can be found in Supplement F. Putting together Theorems 2 and 3 with results from [28], allow us to obtain analytical expressions for the average run lengths of the new stopping rules.

In the following, we provide a sketch of the key steps to obtain each result. We defer readers to the Supplement C for a more technical treatment. The subsequent results are derived under the following condition:

Condition 1. *There is a positive constant $\mathbb{C}$, $1 \le \mathbb{C} < \infty$, depending only on k, such that*

$$\sup_{1 \le j \le n} \left(\sum_{i=1}^{n} A_{n,ij}^{+} \right) \le \mathbb{C}, \quad n \in \mathbb{N}.$$

In k-NN, each observation points to its first k NNs, so the out-degree of each observation is k. However the in-degree of each observation can vary. This condition states that the in-degree of each observation is bounded. It is satisfied almost surely for multivariate data [29], [30].

Let

$$D_L(t,n) = \frac{R_{\text{diff},L}(t,n) - \mathsf{E}(R_{\text{diff},L}(t,n))}{\sqrt{\text{Var}(R_{\text{diff},L}(t,n))}},$$

$$W_L(t,n) = \frac{R_{w,L}(t,n) - \mathsf{E}(R_{w,L}(t,n))}{\sqrt{\text{Var}(R_{w,L}(t,n))}},$$

which replaces the conditional variances of the test statistics with the unconditional variances. See Supplement C for more details.

Our proof that the limiting distributions converge to independent two-dimensional random fields depends on utilizing Stein's method [27]. The general idea of Stein's method is to show for a random variable W and a standard normal random variable Z, that for some family of functions $h \in Lip(1)$, the following bound holds:

$$\sup_{h \in Lip(1)} |\mathsf{E}h(W) - \mathsf{E}h(Z)| \le \delta,$$

where δ depends on the structure of W. The specific form of Stein's method used here requires $A_{n_L,ij}^{+}$ to be locally dependent. However, even for different i, j, l, r, $A_{n_L,ij}^{+}$ and $A_{n_L,lr}^{+}$ are dependent due to the constraint that $\sum_{j \in n_L} A_{n_L,ij}^{+} = k$ for all $i \in n_L$. Following [23], we relax these dependencies by considering a similar set of Bernoulli random variables $\{\tilde{A}_{n_L,ij}^{+}\}_{i,j \in n_L}$. We keep the following probabilities unchanged:

$$\begin{aligned}
\mathsf{P}(\tilde{A}_{n_L,ij}^{+} = 1) &= \mathsf{P}(A_{n_L,ij}^{+} = 1), \\
\mathsf{P}(\tilde{A}_{n_L,ij}^{+} = 1, \tilde{A}_{n_L,ji}^{+} = 1) &= \mathsf{P}(A_{n_L,ij}^{+} = 1, A_{n_L,ji}^{+} = 1), \\
\mathsf{P}(\tilde{A}_{n_L,ji}^{+} = 1, \tilde{A}_{n_L,li}^{+} = 1) &= \mathsf{P}(A_{n_L,ji}^{+} = 1, A_{n_L,li}^{+} = 1),
\end{aligned}$$

but relax the other dependencies such that $\tilde{A}_{n_L,ij}^{+}$ is independent of $\{\tilde{A}_{n_L,il}^{+}, \tilde{A}_{n_L,li}^{+}\}_{l \ne j}$, and $\tilde{A}_{n_L,ij}^{+}$ and $\tilde{A}_{n_L,lr}^{+}$ are independent when i, j, l, r are all different. Then $\tilde{A}_{n_L,ij}^{+}$ are only locally dependent and can be analyzed through the Stein's method [27].

We are now ready to present the main results.

Theorem 2. *Under Condition 1, as $L \to \infty$, the finite dimensional distributions of $\{D_L([uL],[vL]) : 0 < v-1 < u < v < \infty\}$ and $\{W_L([uL],[vL]) : 0 < v-1 < u < v < \infty\}$ converge to independent two-dimensional Gaussian random fields, which we denote as $\{D^\star(u,v) : 0 < v-1 < u < v < \infty\}$ and $\{W^\star(u,v) : 0 < v-1 < u < v < \infty\}$, respectively. Here $[x]$ denotes the largest integer smaller than or equal to x for any real number x.*

The detailed proof for Theorem 2 is in Supplement C.

Based on Theorem 2, we can approximate $\mathsf{E}_\infty(T_S(b_S))$, $\mathsf{E}_\infty(T_W(b_W))$, and $\mathsf{E}_\infty(T_{M_\xi}(b_{M_\xi}))$ by examining the asymptotic behavior of our stopping rules:

$$T_S^\star(b_S) = \inf\left\{ n - N_0 : \max_{n_1' \le t \le n_0'} S^\star(\tfrac{t}{L}, \tfrac{n}{L}) > b_S \right\}, \tag{12}$$

$$T_w^\star(b_W) = \inf\left\{ n - N_0 : \max_{n_1' \le t \le n_0'} W^\star(\tfrac{t}{L}, \tfrac{n}{L}) > b_W \right\}, \tag{13}$$

$$T_{M_\xi}^\star(b_{M_\xi}) = \inf\left\{ n - N_0 : \max_{n_1' \le t \le n_0'} M_\xi^\star(\tfrac{t}{L}, \tfrac{n}{L}) > b_{M_\xi} \right\}, \tag{14}$$

where $n_1' = n - n_1$ and $n_0' = n - n_0'$. Our approximations involve the function $\nu(x)$ defined as

$$\nu(x) = 2x^{-2} \exp\{-2 \sum_{m=1}^{\infty} m^{-1} \Phi\left(-\tfrac{1}{2} x m^{1/2}\right)\}, x > 0.$$

This function is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A simple approximation given in [31] is sufficient for numerical purposes:

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2) - 0.5)}{(x/2)\Phi(x/2) + \phi(x/2)},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and $\phi(\cdot)$ the density function of the standard normal distribution.

According to [28], $T_S^\star(b_S)$, $T_W^\star(b_W)$, and $T_D^\star(b_D)$ are asymptotically exponentially distributed $T_S^\star(b_S) \sim \exp(\lambda_S)$, $T_W^\star(b_W) \sim \exp(\lambda_W)$, $T_D^\star(b_D) \sim \exp(\lambda_D)$, with means:

$$\mathsf{E}_\infty(T_S^\star(b_S)) = \frac{1}{\lambda_S} \approx \frac{\pi \exp(b_S/2)}{c_0 b_S H(c_0, h_1, h_2)} \tag{15}$$

$$\mathsf{E}_\infty(T_W^\star(b_W)) = \frac{1}{\lambda_S} \approx \frac{\sqrt{2\pi} \exp(b_W^2/2)}{c_1^2 b_W G(c_2, g_{W,1}, g_{W,2})} \tag{16}$$

$$\mathsf{E}_\infty(T_D^\star(b_D)) = \frac{1}{\lambda_D} \approx \frac{\sqrt{2\pi} \exp(b_D^2/2)}{2c_2^2 b_D G(c_2, g_{D,1}, g_{D,2})} \tag{17}$$

$$\mathsf{E}_\infty(T_{M_\xi}^\star(b_{M_\xi})) \approx \tag{18}$$

$$\begin{cases} \dfrac{\mathsf{E}_\infty(T_D^\star(b_{M_\xi}))\mathsf{E}_\infty(T_W^\star(b_{M_\xi}/\xi))}{\mathsf{E}_\infty(T_D^\star(b_{M_\xi})) + \mathsf{E}_\infty(T_W^\star(b_{M_\xi}/\xi))} & \text{when } \xi > 0, \\ \xi \mathsf{E}_\infty(T_D^\star(b_{M_\xi})) & \text{when } \xi = 0, \end{cases}$$

with

$$\begin{aligned}
H(c, h_1, h_2) = \int_0^{2\pi} \int_{x_0}^{x_1} &\{h_1(x,\omega) h_2(x,\omega) \times \\
&\nu(\sqrt{2c\, h_1(x,\omega)}) \nu(\sqrt{2c\, h_2(x,\omega)})\} \, dx d\omega,
\end{aligned}$$

$$G(c, g_1, g_2) = \int_{x_0}^{x_1} g_1(x) g_2(x) \nu(c\sqrt{2g_1(x)}) \nu(c\sqrt{2g_2(x)}) dx,$$

$$
\begin{aligned}
g_{W,1}(x) &= \tfrac{\partial_{-}\rho^\star_W(\delta_1,0)}{\partial\delta_1}\Big|_{\delta_1=0} \equiv -\tfrac{\partial_{+}\rho^\star_W(\delta_1,0)}{\partial\delta_1}\Big|_{\delta_1=0},\\
g_{D,1}(x) &= \tfrac{\partial_{-}\rho^\star_D(\delta_1,0)}{\partial\delta_1}\Big|_{\delta_1=0} \equiv -\tfrac{\partial_{+}\rho^\star_D(\delta_1,0)}{\partial\delta_1}\Big|_{\delta_1=0},\\
g_{W,2}(x) &= -\tfrac{\partial_{+}\rho^\star_W(\delta_2,0)}{\partial\delta_2}\Big|_{\delta_1=0},\\
g_{D,2}(x) &= -\tfrac{\partial_{+}\rho^\star_D(\delta_2,0)}{\partial\delta_2}\Big|_{\delta_1=0},\\
h_1(x,\omega) &= g_{W,1}(x)\sin^2(\omega) + g_{D,1}(x)\cos^2(\omega),\\
h_2(x,\omega) &= g_{W,2}(x)\sin^2(\omega) + g_{D,2}(x)\cos^2(\omega).
\end{aligned}
$$

We now have approximations of the average run lengths for the new stopping rules. The only remaining unspecified quantities are the directional partial derivatives of the covariance functions of the Gaussian random fields. Their analytical expressions are derived in the following theorem.

Theorem 3. *For two-dimensional fields $\{W^\star(u,v): 0 < v-1 < u < v < \infty\}$ and $\{D^\star(u,v): 0 < v-1 < u < v < \infty\}$, the directional partial derivatives are*

$$
\begin{aligned}
g_{W,1}(x) &= \frac{1}{2x(1-x)},\\
g_{W,2}(x) &= \frac{x^2-x+1}{x(1-x)} - \frac{2k\,p^{(k)}_{k+1,\infty}}{k+p_{k,\infty}},\\
g_{D,1}(x) &= \frac{1}{x(1-x)},\\
g_{D,2}(x) &= \frac{10q_{k,\infty} - 4kq^{(k)}_{k+1,\infty} - (6k^2-10k)}{2(q_{k,\infty}-k^2+k)} - \frac{1}{2x(1-x)},
\end{aligned}
$$

where

$$
p^{(k)}_{k+1,\infty} = \sum_{r=1}^{k} p_\infty(r,k+1), \quad q^{(k)}_{k+1,\infty} = \sum_{r=1}^{k} q_\infty(r,k+1).
$$

Here, $p_{k,\infty}$ is the limiting expected number of mutual NNs a node has in k-NN, $q_{k,\infty}$ is the limiting expected number of nodes that share a NN with another node in k-NN, $p_\infty(r,s)$ is the limiting expected number of mutual NNs shared between the rth and sth NNs, and similarly $q_\infty(r,s)$ is the limiting expected number of nodes shared between the rth and sth NNs. Explicitly, $p_{k,\infty} = \sum_{r=1}^{k}\sum_{s=1}^{k} p_\infty(r,s)$, and $q_{k,\infty} = \sum_{r=1}^{k}\sum_{s=1}^{k} q_\infty(r,s)$, with $p_\infty(r,s) = \lim_{L\to\infty}\frac{1}{L}\sum_{i,j\in n_L} A^{(r)}_{n_L,ij}A^{(s)}_{n_L,ji}$ and $q_\infty(r,s) = \lim_{L\to\infty}\frac{1}{L}\sum_{i,j,l\in n_L, j\neq l} A^{(r)}_{n_L,ji}A^{(s)}_{n_L,li}$.

To derive these partial derivatives, we studied the dynamics of the k-NN series as new observations are added. It turns out that a few key quantities are enough to characterize the dynamics in the asymptotic domain. The proof of this theorem is in Supplement D.

B. Finite L

We now consider the practical scenario where L is finite. Based on results in Section III-A, $\mathsf{E}_\infty(T_S(b_S))$, $\mathsf{E}_\infty(T_W(b_W))$, and $\mathsf{E}_\infty(T_{M_\xi}(b_M))$ can be approximated by

$$
\mathsf{E}_\infty(T_S(b_S)) \approx \frac{L\pi\exp(b_S/2)}{b_S^2 H_L(b_S,h_1,h_2)} \tag{19}
$$

$$
\mathsf{E}_\infty(T_W(b_W)) \approx \frac{L\sqrt{2\pi}\exp(b_W^2/2)}{b_W^3 G_L(b_W,g_{W,1},g_{W,2})}, \tag{20}
$$

$$
\begin{aligned}
&\mathsf{E}_\infty(T_{M_\xi}(b_{M_\xi}))\\
&\approx \begin{cases} \dfrac{\mathsf{E}_\infty(T_D(b_{M_\xi}))\mathsf{E}_\infty(T_W(\frac{b_{M_\xi}}{\xi}))}{\mathsf{E}_\infty(T_D(b_{M_\xi})) + \mathsf{E}_\infty(T_W(\frac{b_{M_\xi}}{\xi}))} & \text{when } \xi > 0,\\ \mathsf{E}_\infty(T_D(b_{M_\xi})) & \text{when } \xi = 0, \end{cases}
\end{aligned}
$$

$$
\text{where } \mathsf{E}_\infty(T_D(b_D)) \approx \frac{L\sqrt{2\pi}\exp(b_D^2/2)}{2b_D^3 G_L(b_D,g_{D,1},g_{D,2})}, \tag{21}
$$

with $H_L()$ and $G_L()$ are finite sample versions of $H()$ and $G()$, respectively. In practice, when L is finite we use $g_{W,2}(L,x)$ and $g_{D,2}(L,x)$ in place of $g_{W,2}(x)$ and $g_{D,2}(x)$ in the above formulas, respectively, where

$$
\begin{aligned}
g_{W,2}(L,x) &= \tfrac{x^2-x+1}{x(1-x)} - \tfrac{2k\,p^{(k)}_{k+1,L}}{k+p_{k,L}},\\
g_{D,2}(L,x) &= \tfrac{10q_{k,\infty}-4kq^{(k)}_{k+1,L}-(6k^2-10k)}{2(q_{k,L}-k^2+k)} - \tfrac{1}{2x(1-x)}.
\end{aligned}
$$

Here, $p_{k,L}$, $p^{(k)}_{k+1,L}$, $q_{k,L}$, and $q^{(k)}_{k+1,L}$ are the finite sample versions of $p_{k,\infty}$, $p^{(k)}_{k+1,\infty}$, $q_{k,\infty}$, and $q^{(k)}_{k+1,\infty}$ and can be estimated in a data-driven way.

C. Skewness correction

Analytical approximations provided in Section III-B become less precise for finite L when n_0 is relatively small. This is mainly because the convergence of $W_L(t,n)$ and $D_L(t,n)$ to normal is slow if $(n-t)/L$ is close to 0 or 1. This problem becomes more severe when dimension is high. To improve upon the analytic approximations for finite sample sizes, we perform skewness correction. We adopt a skewness correction approach discussed in [32] that does the correction up to different extents based on the amount of skewness at each value of t. In particular, we provide better approximations to the marginal probabilities $\mathsf{P}(W^\star(u-x,w) \in b+du)$ and $\mathsf{P}(D^\star(u-x,w) \in b+du)$. Following the method based on cumulant-generating functions and change of measure (details refer to [32]), we can approximate the marginal probability by

$$
\frac{1}{\sqrt{2\pi(1+\gamma\theta_b)}}\exp(-\theta_b - u\theta_b/b + \theta_b^2(1+\gamma\theta_b/3)/2),
$$

where θ_b is chosen such that $\dot{\psi}(\theta_b) = b$. By a third Taylor approximation, we get $\theta_b \approx (-1+\sqrt{1+2\gamma_L(t,n)b})/\gamma_L(t,n)$, where $\gamma_L(t,n) := \mathsf{E}_P(Z_L(t,n)^3)$ and explicit expressions are derived using combinatorial analysis.

Skewness corrected thresholds are only obtained for $W_{L|\mathbf{y}}(t,n)$ and $D_{L|\mathbf{y}}(t,n)$, but not $S_{L|\mathbf{y}}(t,n)$. This is because for $S_{L|\mathbf{y}}(t,n)$ the integrand can easily be non-finite and the approach depends heavily on extrapolation. Therefore, the stopping rule based on $M_{L|\mathbf{y}}(t,n)$ is often recommended because it can detect general changes but we can obtain more accurate stopping thresholds.

D. Checking accuracy of analytic formulas for the average run lengths

Here, we check the accuracy of the analytic formulas for the average run lengths. For all three new tests, we have analytic formulas based on asymptotic results (19), (20) and

(21), and for the tests based on the weighted/max-type edge-count tests, we have analytic formulas after skewness correction, (Supplement E-7), (Supplement E-8). We compare the empirical ARL obtained from these analytic formulas (with skewness correction with applicable) to those obtained from $1,000$ Monte Carlo simulations. The analytical thresholds are obtained so that the average run length is $2,000$. We generated data from three different settings: multivariate normal with $d = 10$ (denoted by $C1$), multivariate t_5 with $d = 100$ (denoted by $C2$), and multivariate log-normal with $d = 1000$ (denoted by $C3$).

Results for different choices of n_0 are shown in Tables IV - VIII. We set $n_1 = L - n_0$ and $k = 5$. The asymptotic analytic results are denoted by 'A1' and the skewness corrected approximations are denoted by 'A2'. We see in general that the empirical ARLs obtained from the asymptotic approximations ('A1') are not very close, illustrating the need for skewness correction here. After skewness correction, the empirical ARLs are much closer to $2,000$. It is clear that the accuracy of the skewness corrected approximations depends on n_0: in general, when $n_0 = 40$, the skewness corrected approximations do well across most dimensions.

We also investigate how the analytical threshold approximations perform compare to the Monte Carlo thresholds in the presence of change. In Tables V, VII, and IX we report the power, defined as the fraction of trials able to detect the change within 50 observations after the change occurs, and the average detection delay (reported in parenthesis) obtained using analytical formulas ('A1'), analytical formulas with skewness correction ('A2') and Monte Carlo thresholds ('MC') . The thresholds are obtained so that the ARL is set to be $2,000$. The amount of signal used correspond to the power setup described for Tables X, XII, and XI in Section IV, respectively. It is clear that in the presence of change, the analytical formulas (with skewness correction) and Monte Carlo thresholds all lead to similar results in power and detection delay.

TABLE IV
EMPIRICAL ARL OBTAINED FROM ANALYTICAL b_S SUCH THAT $\mathsf{E}_\infty(T_S(b_S)) = 2,000$, $L = 200$.

	$n_0 = 35$	$n_0 = 40$
$C1$	2286.56	2314.81
$C2$	2330.53	2517.39
$C3$	1855.41	2396.45

TABLE V
POWER AND DETECTION DELAY (REPORTED IN PARENTHESIS) FOR $T_S(b_S)$ OBTAINED USING ANALYTICAL THRESHOLDS ('A1') AND MONTE CARLO THRESHOLDS ('MC').

	$n_0 = 35$		$n_0 = 40$	
	A1	MC	A1	MC
$C1$	0.22 (67.72 ±33.03)	0.23 (67.46 ±32.92)	0.24 (66.71 ±33.17)	0.22 (70.72) ±34.88)
$C2$	0.50 (48.62 ±19.01)	0.51 (48.29 ±19.00)	0.51 (48.29 ±18.99)	0.54 (47.82 ±19.35)
$C3$	0.65 (45.84 ±24.71)	0.64 (46.64 ±25.06)	0.66 (45.58 ±24.61)	0.67 (45.16 ±24.76)

TABLE VI
EMPIRICAL ARL OBTAINED FROM ANALYTICAL b_W SUCH THAT $\mathsf{E}_\infty(T_W(b_W)) = 2,000$, $L = 200$.

	$n_0 = 35$		$n_0 = 40$	
	A1	A2	A1	A2
$C1$	1081.55	2216.99	1247.97	2363.42
$C2$	1430.04	2075.81	1619.72	2079.41
$C3$	1380.30	2086.30	1679.95	1904.12

IV. POWER ASSESSMENT

A. Multivariate data

To examine the performance of the three new test statistics, we compare them to the existing approach in [23] $(\max_t Z_{L|\mathbf{y}}(t, n))$ and two parametric likelihood-based approaches: Hotelling's T^2 test when there is change in mean and the generalized likelihood ratio test when there is variance change (both these two-sample tests are adapted to the scan statistic setting). The simulation setup is as follows: there are $N_0 = 200$ historical observations and a change occurs at $t = 400$ (200 new observations after the start of the test). The observations are independent and follow a d-dimensional distribution. When there is a change in mean, the observations are shifted from 0 by amount Δ in Euclidean distance. When the covariance matrix changes, to make the change less significant, only the first $d/5$ of the diagonal elements change with a multiple of σ, and the rest are unchanged. The amount of change is chosen so that the tests have moderate power to be comparable. For fair comparison, we use Monte Carlo simulations to determine the threshold for each of the test so that their average run lengths are all 2,000. Power is reported as the fraction of trials for which the change-point is detected within 50 observations after the change occurred. In the following, we use 'HT' to refer to the scan statistic over the Hotelling's T^2 statistic and use 'GLR' to refer to the scan statistic over the generalized likelihood ratio statistic. For $d > 100$, in order for HT and GLR to be applicable in higher dimensions, we treat each data stream as if it is independent so that the covariance matrix's inverse and determinant are well-defined.

When there is both mean and variance change, Table X shows the results under the Gaussian setting. When $d \leq 500$, the GLR dominates in power. However, when the dimension increases, GLR is no longer able to retain competitive power compared to S and M. On the other hand, this setting is not well-suited for W which is meant to capture mean change only and its performance is the worst here.

To consider other distributions, we also compared the tests for multivariate log-normal data and multivariate t_5 data. The results for the log-normal data are shown in Table XI. Here there is a change in mean parameter only and Δ is chosen such that the location change dominates. In this setting, W's performance dominates. We see that all the new tests outperform Z and the parametric tests for $d > 10$.

The result for the multivariate t_5 data are shown in Table XII. When there is a change in both mean and variance, Z is unable to outperform the new test statistics. Among the

TABLE VII
POWER AND DETECTION DELAY FOR $T_W(b_W)$ OBTAINED USING ANALYTICAL THRESHOLDS ('A1'), SKEWNESS CORRECTED THRESHOLDS ('A2'), AND MONTE CARLO THRESHOLDS ('MC').

	$n_0 = 35$			$n_0 = 40$		
	A1	A2	MC	A1	A2	MC
$C1$	0.21 (70.68 ±39.21)	0.13 (71.90 ±35.87)	0.13 (72.46 ±36.03)	0.21 (69.97 ±39.51)	0.13 (74.06 ±36.77)	0.12 (74.20 ±38.05)
$C2$	0.48 (46.25 ±19.06)	0.46 (47.18 ±18.92)	0.46 (47.85 ±19.41)	0.49 (45.77 ±18.69)	0.47 (47.01 ±19.10)	0.46 (47.13 ±18.98)
$C3$	0.77 (39.47 ±22.05)	0.75 (40.53 ±22.63)	0.74 (41.53 ±22.76)	0.78 (38.80 ±21.44)	0.77 (39.63 ±22.23)	0.76 (40.61 ±22.61)

TABLE VIII
EMPIRICAL ARL OBTAINED FROM ANALYTICAL b_{M_ξ} SUCH THAT $\mathsf{E}_\infty(T_{M_\xi}(b_{M_\xi})) = 2,000$, $L = 200$.

	$n_0 = 35$		$n_0 = 40$	
	A1	A2	A1	A2
$C1$	1050.28	1980.73	1189.45	1997.47
$C2$	1390.37	1859.99	1507.60	2110.34
$C3$	1245.26	2053.76	1633.61	1935.36

new test statistics, their performance depends on whether the mean or variance signal dominates. When the mean change is stronger (for example, when $d = 100$), W performs comparably well. However, when the variance change is stronger (for example, when $d \geq 1000$), M and S dominate.

Based on the results of these tables, we see that the new graph-based methods perform well under various scenarios and have improved detection delay over the existing method in [23]. In general, if one is certain that the change is locational, the test based on W is recommended; while for more general changes, the tests based on S and M are recommended.

B. Network data

Non-Euclidean data, also referred to as object data, is simply data that does not lie in the Euclidean space. Examples include networks, images, shapes, and trees. Fundamental statistical tools involving vector space analysis are no longer applicable for object data, making it challenging to directly analyze these data types. To demonstrate the new test statistics' power on non-Euclidean data, we evaluate the proposed methods on a sequence of networks. We generate random networks using the configuration model, which is a specific method that allows us to generate random networks with a specified degree sequence (please see [33], [34] for an overview). For every node with degree k_i, we create k half-edges (referred to as 'stubs'). The network is created by iteratively selecting two stubs uniformly at random and connecting them to form an edge. This is done until no stubs remain and will result in a network with a predefined degree sequence. For the graph-based results, we show the results for graphs constructed using two different similarity measures. Specifically, for a network at time t, we encode the network using an adjacency matrix A_t with 1 for element (i, j) if node i and j are connected, and 0 otherwise. The similarity measures are:

1) Similarity 1: $||A_t - A_s||_F$,
2) Similarity 2: $\frac{||A_t - A_s||_F}{\sqrt{||A_t||_F * ||A_s||_F}}$.

TABLE IX
POWER AND DETECTION DELAY FOR $T_{M_\xi}(b_{M_\xi})$ OBTAINED USING ANALYTICAL THRESHOLDS ('A1'), SKEWNESS CORRECTED THRESHOLDS ('A2'), AND MONTE CARLO THRESHOLDS ('MC').

	$n_0 = 35$			$n_0 = 40$		
	A1	A2	MC	A1	A2	MC
$C1$	0.28 (66.55 ±36.01)	0.19 (72.48 ±36.08)	0.17 (72.51 ±34.72)	0.26 (68.67 ±36.26)	0.21 (70.89 ±35.61)	0.22 (70.04 ±35.31)
$C2$	0.61 (46.49 ±19.29)	0.55 (47.68 ±19.67)	0.50 (48.27 ±19.41)	0.63 (46.32 ±19.25)	0.58 (46.83 ±19.40)	0.56 (47.54 ±19.67)
$C3$	0.75 (40.76 ±22.67)	0.73 (41.83 ±22.87)	0.71 (43.11 ±23.10)	0.75 (40.46 ±22.54)	0.75 (41.41 ±23.01)	0.74 (41.86 ±22.86)

TABLE X
MULTIVARIATE GAUSSIAN DATA, MEAN AND VARIANCE CHANGE.

	Power				
d	10	100	500	1000	2000
Δ	0.35	0.5	0.9	1	0.85
σ	0.55	0.65	0.8	0.9	0.9
HT	0.02 (125.76) (±45.92)	0 - -	0.003 (128.87) -	0.008 (119.04) -	0.005 (120.64) -
GLR	**0.34** **(59.81** **±23.77)**	**1** **(26.66** **±3.79)**	**1** **(29.99** **±4.23)**	0.17 (73.21 ±30.34)	0.42 (57.13 ±23.29)
Z	0.034 (101.70) (±32.32)	0.09 (89.82) (±30.25)	0.12 (81.23) (±28.08)	0.07 (92.61) (±33.36)	0.10 (83.69) (±31.31)
W	0.13 (72.46) (±36.03)	0.17 (60.34) (±40.25)	0.14 (57.20) (±36.58)	0.08 (65.67) (±43.13)	0.08 (62.81) (±42.16)
S	0.23 (67.46) (±32.92)	0.83 (34.75) (±20.00)	0.98 (22.95) (±9.82)	0.78 (35.44) (±19.88)	**0.98** **(22.22)** **(±9.53)**
M	0.17 (72.51) (±34.72)	0.81 (35.07) (±20.14)	0.98 (22.25) (±10.23)	**0.79** **(35.22)** **(±19.97)**	**0.98** **(21.25)** **(±9.31)**

Under this setting, we compare the graph-based approach to another method designed specifically to detect changes in a stream of networks/graphs [22]. The approach proposed in [22] is quite general: it does not make explicit assumptions on the network/graph and can be applied to a stream of graphs of varying sizes. Their approach consists of two steps: (1) each graph G_t is mapped to a vector y_t through a prototype-based embedding, and (2) a change is then detected in the stream of vectors y_t using any conventional multivariate change-detection procedure. Specifically, embedding is carried out by assessing the dissimilarity between graphs and a selection of prototypes. In the paper, they rely on the graph edit distance (GED), which count and weights the edit operations that are needed in order to make two input graphs equal and is applicable to graphs where the nodes are unidentified. To carry out their approach, we follow the implementation proposed in [22]; specific details can be found in Section IV of [22].

Since computing the GED is quite computationally expensive, we allow each network to have only 6 nodes. We set a change at $\tau = 101$. Before and after the change, the total node degree remains the same, but the sequence of node degree

TABLE XI
MULTIVARIATE LOG-NORMAL DATA, DIFFER IN THE MEAN PARAMETER.

	Power				
d Δ	10 0.95	100 1.6	500 1.9	1000 2	2000 2.1
HT	**0.82** **(33.17** **±24.15**)	0.32 (50.79 ±35.58)	0.24 (40.22 ±17.60)	0.13 (45.92 ±29.19)	0.10 (60.28 ±40.27)
GLR	0.06 (93.59 ±53.90)	0.09 (60.75 ±43.58)	0.07 (71.51 ±56.15)	0.05 (78.72 ±55.72)	0.03 (68.55) ±57.57)
Z	0.43 (55.66 ±29.73)	0.37 (55.32 ±19.77)	0.25 (60.60 ±19.96)	0.21 (64.99 ±22.31)	0.17 (68.96 ±22.85)
W	0.43 (56.46) ±31.97)	**0.86** **(35.29)** **±15.97**)	**0.84** **36.67** **±16.64**)	**0.74** **(41.53)** **±22.76**)	**0.62** **(45.87)** **±25.19**)
S	0.39 (57.97 ±32.76)	0.81 (38.96 ±18.14)	0.74 (41.63 ±19.78)	0.64 (46.64 ±25.06)	0.51 (50.88 ±28.22)
M	0.41 (57.34) ±32.36)	0.85 (36.10) ±16.49)	0.81 (38.01) ±17.32)	0.71 (43.11) ±23.10)	0.59 (47.75) ±27.25)

TABLE XII
MULTIVARIATE t DATA WITH 5 DEGREES OF FREEDOM, MEAN AND VARIANCE DIFFERENCE.

	Power				
d Δ σ	10 0.20 0.30	100 1.9 0.65	500 2.2 0.68	1000 1.6 0.7	2000 3.3 0.78
HT	0.10 (67.63) ±19.59)	0 - -	0.23 (68.95) ±26.05)	0.005 (123.67) -	0.33 (58.32) ±21.57)
GLR	0.16 (71.30 ±21.79)	0.027 (87.67 ±30.92)	0.087 (74.37 ±22.30)	0.091 (71.52 ±21.22)	0.10 (76.21 ±21.07)
Z	0.07 (73.23) ±22.97)	0.17 (73.05) ±17.69)	0.06 (74.67) ±19.97)	0.16 (70.31) ±23.39)	0.19 (69.95) ±20.65)
W	0.10 (55.32) ±22.91)	0.46 **(47.85)** **±19.41**)	0.36 (44.94) ±18.79)	0.13 (44.41) ±22.01)	0.34 (44.17) ±17.97)
S	**0.13** **(58.31** **±23.06**)	**0.51** (48.29 ±19.00)	**0.53** **(41.01** **±18.96**)	0.64 (34.29 ±16.74)	**0.67** **(36.79** **±15.80**)
M	0.09 (58.32 ±22.55)	0.50 (48.27 ±19.41)	0.47 (42.97 ±19.99)	**0.67** **(33.49** **±16.26**)	0.59 (38.30 ±17.47)

changes. Before the change, each network is constructed such that nodes 1 and 4 have node degree of 1 and the remaining nodes have node degree 2. After the change, nodes 3 and 6 have node degree of 1 and the remaining nodes have node degree 2. Observe that in this setting Similarity 1 and 2 are equivalent and so we only report results for Similarity 1. The first 50 observations are treated as training observations for [22]. We set $L = 50$ for the graph-based approach. For fair comparison with the method in [22], we set the window size to be 50 and we implement their method using both GED and the Forbenius norm (Similarity 1) as the similarity measure between networks. The ARL is set to be 2,000. Power is defined as the number of trials (out of 100) where the change is detected within 50 observations. Among those trials where the change is detected, we also report the expected detection delay (EDD) and its standard deviation when applicable. We note that the approach in [22] depends on the choice of embedding and its hyper-parameters, which should be chosen carefully. A thorough discussion is provided in [22]. We carry out simulations under a variety of settings for the number of dimensions to embed (d) and the number of prototypes to select (nproto).

Table XIII show the power performance of the graph-based methods compared to the approach in [22]. We can see that all of the graph-based statistics do better in terms of power and detection delay compared to [22].

TABLE XIII
POWER COMPARISON FOR CONFIGURATION NETWORK MODEL.

	Power	EDD
Zambon et. al [22] (d=5, nproto = 5)		
GED	2	51.00 ± NA
Similarity 1	4	26 ± 25
Zambon et. al [22] (d=10, nproto = 10)		
GED	12	17.67 ± 23.57
Similarity 1	4	26 ± 25
Zambon et. al [22] (d=15, nproto = 15)		
GED	26	24.08 ± 24.93
Similarity 1	20	46 ± 15
Z		
Similarity 1	92	20.12 ± 5.03
W		
Similarity 1	92	21.45 ± 5.28
S		
Similarity 1	82	25.89 ± 7.16
M		
Similarity 1	82	21.09 ± 7.42

For further exploration of the graph-based approach, we generate a sequence of networks such that each network has 20 nodes with a pre-specified degree sequence (see below for details). A change in the sequence of networks happens at $\tau = 101$. The first 50 networks in the sequence are treated as historical observations. Power is reported as the fraction of trials for which the change-point is detected within 50 observations after the change occurred. Among those trials where the change is detected within the first 50 observations, the expected detection delay and its standard deviation is also reported (EDD).

The networks are generated under two different settings:

1) A fixed degree change in the network: before the change, half of the nodes have out-degree and in-degree 1 and half the nodes have out-degree and in-degree 3. After the change, 5 nodes have out-degree and in-degree k_1 and 5 nodes have out-degree and in-degree k_2. The remaining nodes remain unchanged.
2) A random degree change in the network: before the change, half of the nodes have out-degree and in-degree 1 and half the nodes have out-degree and in-degree 3. After the change, half of the nodes have out-degree and in-degree randomly selected from k_3 to k_4 and the remaining half of the nodes have out-degree and in-degree 3.

Tables XIV and XV report the power and expected detection delay of the graph-based test statistics for similarity measure 1 and 2, respectively. For similarity measure 1, the performance of S and M dominate in almost all settings, with the exception of when $k_1 = 2, k_2 = 2$; in general S and M do well with respect to both power and expected detection delay. For similarity measure 2, the performance of W and Z improve substantially, while the performance of S and M remain stable.

TABLE XIV
POWER COMPARISON OF GRAPH-BASED METHODS FOR CONFIGURATION NETWORK MODEL WITH 20 NODES UNDER SIMILARITY MEASURE 1.

	Fixed degree			
	$k_1 = 2,\ k_2 = 2$		$k_1 = 4,\ k_2 = 5$	
	Power	EDD	Power	EDD
Z	74	16.32 ± 4.40	14	23.71 ± 8.00
W	80	25.41 ± 6.58	57	36.42 ± 4.32
S	72	10.65 ± 3.42	71	8.97 ± 2.50
M	60	9.08 ± 4.06	55	7.94 ± 2.38

	Random degree			
	$k_3 = 1,\ k_4 = 3$		$k_3 = 1,\ k_4 = 7$	
	Power	EDD	Power	EDD
Z	0	NA	41	18.26 ± 5.83
W	1	$45 \pm -$	55	31.41 ± 4.40
S	78	13.42 ± 3.12	65	9.32 ± 2.72
M	59	12.05 ± 2.98	48	8.83 ± 3.03

TABLE XV
POWER COMPARISON OF GRAPH-BASED METHODS FOR CONFIGURATION NETWORK MODEL WITH 20 NODES UNDER SIMILARITY MEASURE 2.

	Fixed degree			
	$k_1 = 2,\ k_2 = 2$		$k_1 = 4,\ k_2 = 5$	
	Power	EDD	Power	EDD
Z	74	16.32 ± 4.40	2	$42.5 \pm -$
W	80	25.41 ± 6.58	4	$25.75 \pm -$
S	72	10.65 ± 3.42	65	7.66 ± 1.06
M	60	9.08 ± 4.06	51	7.98 ± 2.45

	Random degree			
	$k_3 = 1,\ k_4 = 3$		$k_3 = 1,\ k_4 = 7$	
	Power	EDD	Power	EDD
Z	23	22 ± 8.10	1	$46 \pm -$
W	64	13.38 ± 3.39	12	13.30 ± 5.69
S	65	10.45 ± 2.07	63	9.15 ± 2.75
M	58	12.03 ± 2.73	48	12.17 ± 2.39

V. A REAL DATA APPLICATION

We compare the new approaches to the method in [23] using the yellow taxi trip records data. The data set is publicly available on the NYC Taxi & Limousine Commission (TLC) website. It provides information on the taxi pickup and drop-off date/times, longitude and latitude coordinates of pickup and drop-off locations, trip distances, fares, rate types, payments types, and driver-reported passenger counts.

Based on this data set, a natural question to ask is: Can we detect a change in traffic patterns during peak travel seasons? Here, we focus on those trips that began at John F. Kennedy International Airport and we look at two different time periods: the months of June through August and November through December in 2015. The dataset has been completely collected at the time of analysis. However, we treat it as if the data were being observed in order to illustrate how the proposed method works. For simplicity, the boundary of JFK airport was set to be 40.63 to 40.66 latitude and −73.80 to −73.77 longitude.

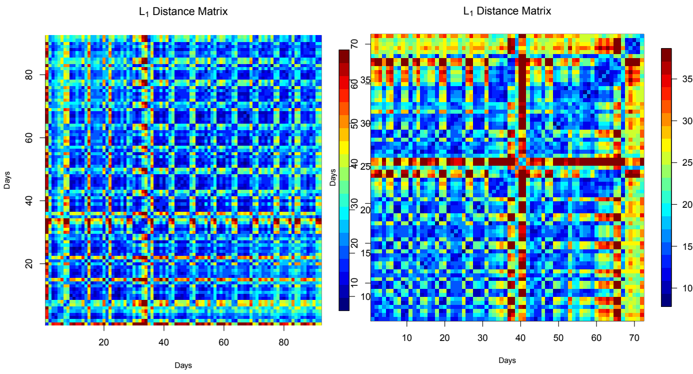

Fig. 2. Left panel: Heatmap of L_1 norm distance matrix of vector v_i for $i = 1, \dots 93$, corresponding to dates June 1, 2015 - Aug. 31, 2015. Right panel: Heatmap of L_1 norm distance matrix of vector v_i for $i = 1, \dots 72$, corresponding to dates Oct. 21, 2015 - Dec. 31, 2015.

For those trips that began with a pickup at JFK, we extract information on their longitude and latitude drop-off coordinates. Using longitude/latitude coordinates, we create a 30 by 30 grid of New York City and count the number of taxi drop-offs that fall within each cell, where each cell represents a longitude, latitude coordinate range. Then for each day, we have a 30 by 30 matrix such that each element represents the number of taxi drop-offs in each location.

TABLE XVI
DETECTED STOPPING TIMES FOR NYC TAXI PICKUPS FROM JFK FOR JUNE 1, 2015 - AUGUST 31, 2015.

	Reported stopping times	Estimated change-point
Z	—	—
W	07/03 - 07/04	06/29 (Day 30)
M	07/03 - 07/04	06/29 (Day 30)
S	07/03 - 07/05	06/29 (Day 30)

We apply the new approaches, as well as Z, to detect changes in the months of June through August 2015. We use data from the month of May as historical data. Applying the offline change-point detection method in [32] and [25] on the observations in May, we find there is no change-point in the first 30 days, so we set $L = 30$, $n_0 = 5$, and $n_1 = L - n_0$.

TABLE XVII
DETECTED STOPPING TIMES FOR NYC TAXI PICKUPS FROM JFK FOR OCTOBER 21, 2015 - DECEMBER 31, 2015.

	Reported stopping times	Estimated change-point
Z	11/27 - 11/31 12/23 - 12/25 12/30 - 12/31	11/21 (Day 32) 12/10 (Day 51) 12/26 (Day 67)
W	11/28 12/23 - 12/26 12/29 - 12/31	11/21 (Day 32) 12/19 (Day 60) 12/26 (Day 67)
M	11/28 12/23 - 12/26 12/29 - 12/31	11/21 (Day 32) 12/19 (Day 60) 12/26 (Day 67)
S	11/27 - 11/30 12/23 - 12/26 12/29 - 12/31	11/21 (Day 32), 11/23 (Day 34) 12/19 (Day 60) 12/26 (Day 67)

We denote A_i to be the 30 by 30 matrix on day i and v_i to be the vector form of A_i, which is now 900 by 1. The L_1 norm is used to construct the k-NN graph representing similarity between days. Here, the new test statistics (W, M, and S) all report a stopping time of July 3 and July 4 whereas Z is unable to detect any anomaly event (Table XVI). The change-point triggering these stopping times is estimated to be June 29. To perform a sanity check, we plot a heatmap of the L_1 distance matrix used to the construct the k-NN graph (see Figure 2, left panel). Based on the heatmap, we can see there is a clear signal happening around Day 30, which corresponds with the results from the new test statistics.

To detect changes in November and December 2015, we use data from the months of September and October 2015 as historical data. Applying the offline change-point detection method in [32] and [25] on the observations in September and October, we find there is no change-point in the first 50 days. Therefore, we treat the first 50 observations from Sept. 1 - Oct. 20 as historical observations and we begin the test at Oct. 21. We set $L = 50$, $n_0 = 8$, and $n_1 = L - n_0$. The stopping times based on the new test statistics report back dates that seem to be quite reasonable (see Table XVII). We see that multiple stopping times are caused by the same anomaly event. When the signal is large enough, the new test statistics (W, M, and S) and Z perform similarly: all are able to detect a change in travel pattern close to Thanksgiving and the Christmas holidays. Again to check our results, we plot a heatmap of the L_1 distance matrix used to the construct the k-NN graph. We can see that there is a clear signal starting roughly around Day 30 and again around Day 60 and Day 67, which matches the results reported from the test statistics. In comparison with the heatmap from the months of June through August, the signal from the summer months is much weaker and in that case Z is unable to detect any anomaly event.

VI. Conclusion

We propose new graph-based test statistics under k-NN for detecting change-points sequentially as data are generated. We study the asymptotic properties of the stopping rules based on the new test statistics, and derived the analytic formulas to approximate the average run lengths of the new stopping rules. To accommodate finite samples, skewness corrected approximations were also derived for the weighted and max-type edge-count statistic under k-NN. The skewness-corrected versions give much more accurate approximations to the average run lengths and can be used reliably in practice. The performance of the proposed test statistics are examined under various common scenarios. Simulation studies reveal that the new test statistics have shorter detection delays for a wider range of alternatives and exhibit power gains for scale change when compared to parametric tests and the test statistic proposed in [23]. Specifically, simulation results show that the weighted-edge count statistic (W) is useful at quickly detecting mean changes. When a change in variance is also of interest, the generalized edge-count statistic (S) and max-type edge-count statistic (M) are more effective in detecting changes and obtain faster detection. Together with the fact that skewness corrected average run length approximations can be obtained for the max-type edge-count statistic, the stopping rule T_{M_ξ} is recommended for sequential detection of general changes.

Acknowledgments

Hao Chen was supported in part by NSF Grants DMS-1513653 and 1848579.

References

[1] R. J. Bolton, D. J. Hand *et al.*, “Unsupervised profiling methods for fraud detection,” *Credit scoring and credit control VII*, pp. 235–255, 2001.

[2] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann, “Inferring change points in the covid-19 spreading reveals the effectiveness of interventions,” *medRxiv*, 2020.

[3] F. Pervaiz, M. Pervaiz, N. A. Rehman, and U. Saif, “Flubreaks: early epidemic detection from google flu trends,” *Journal of medical Internet research*, vol. 14, no. 5, p. e125, 2012.

[4] M. Zhang, A. Raghunathan, and N. K. Jha, “Medmon: Securing medical devices through wireless monitoring and anomaly detection,” *IEEE Transactions on Biomedical circuits and Systems*, vol. 7, no. 6, pp. 871–881, 2013.

[5] M. Xie, S. Han, B. Tian, and S. Parvin, “Anomaly detection in wireless sensor networks: A survey,” *Journal of Network and computer Applications*, vol. 34, no. 4, pp. 1302–1325, 2011.

[6] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blazek, and H. Kim, “A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods,” *IEEE transactions on signal processing*, vol. 54, no. 9, pp. 3372–3382, 2006.

[7] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer Science & Business Media, 1985.

[8] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC, 2014.

[9] E. S. Page, “Continuous inspection schemes,” *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[10] G. Lorden, “Procedures for reacting to a change in distribution,” *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[11] M. Pollak and D. Siegmund, “Sequential detection of a change in a normal mean when the initial value is unknown,” *The Annals of Statistics*, vol. 19, no. 1, pp. 394–416, 1991.

[12] T. L. Lai, “Sequential changepoint detection in quality control and dynamical systems,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 613–658, 1995.

[13] A. G. Tartakovsky and V. V. Veeravalli, “Asymptotically optimal quickest change detection in distributed sensor systems,” *Sequential Analysis*, vol. 27, no. 4, pp. 441–475, 2008.

[14] Y. Mei, “Efficient scalable schemes for monitoring a large number of data streams,” *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[15] Y. Xie and D. Siegmund, “Sequential multi-sensor change-point detection,” *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[16] Y. Wang and Y. Mei, “Large-scale multi-stream quickest change detection via shrinkage post-change estimation,” *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6926–6938, 2015.

[17] H. P. Chan, “Optimal sequential detection in multi-stream data,” *The Annals of Statistics*, vol. 45, no. 6, pp. 2736–2763, 2017.

[18] F. Desobry, M. Davy, and C. Doncarli, “An online kernel change detection algorithm,” *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 2961–2974, 2005.

[19] N. Keriven, D. Garreau, and I. Poli, “Newma: a new method for scalable model-free online change-point detection,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 3515–3528, 2020.

[20] Y. Yilmaz, “Online nonparametric anomaly detection based on geometric entropy minimization,” in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 3010–3014.

[21] M. N. Kurt, Y. Yilmaz, and X. Wang, “Real-time nonparametric anomaly detection in high-dimensional settings,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[22] D. Zambon, C. Alippi, and L. Livi, “Concept drift and anomaly detection in graph streams,” *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5592–5605, 2018.

[23] H. Chen, “Sequential change-point detection based on nearest neighbors,” *The Annals of Statistics*, vol. 47, no. 3, pp. 1381–1407, 2019.

[24] H. Chen and J. H. Friedman, “A new graph-based two-sample test for multivariate and object data,” *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 397–409, 2017.

[25] L. Chu and H. Chen, “Asymptotic distribution-free change-point detection for multivariate and non-euclidean data,” *The Annals of Statistics*, vol. 47, no. 1, pp. 382–414, 2019.

[26] H. Chen, X. Chen, and Y. Su, “A weighted edge-count two-sample test for multivariate and object data,” *Journal of the American Statistical Association*, pp. 1–10, 2018.

[27] L. H. Chen and Q.-M. Shao, “Stein’s method for normal approximation,” *An introduction to Stein’s method*, vol. 4, pp. 1–59, 2005.

[28] D. Siegmund and E. Venkatraman, “Using the generalized likelihood ratio statistic for sequential detection of a change-point,” *The Annals of Statistics*, pp. 255–271, 1995.

[29] P. J. Bickel and L. Breiman, “Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test,” *The Annals of Probability*, pp. 185–214, 1983.

[30] N. Henze, “A multivariate two-sample test based on the number of nearest neighbor type coincidences,” *The Annals of Statistics*, pp. 772–783, 1988.

[31] D. Siegmund and B. Yakir, *The statistics of gene mapping*. Springer Science & Business Media, 2007.

[32] H. Chen and N. Zhang, “Graph-based change-point detection,” *The Annals of Statistics*, vol. 43, no. 1, pp. 139–176, 2015.

[33] M. Molloy and B. Reed, “A critical point for random graphs with a given degree sequence,” *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[34] M. Newman, *Networks*. Oxford university press, 2018.