ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Graph-Based Change-Point Analysis

Hao Chen[1] and Lynna Chu[2]

[1]Department of Statistics, University of California-Davis, Davis, California, USA; email: hxchen@ucdavis.edu

[2]Department of Statistics, Iowa State University, Ames, Iowa, USA; email: lchu@iastate.edu

Annu. Rev. Stat. Appl. 2023. 10:475–99

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-122121-033817

Keywords

nonparametrics, graph-based tests, scan statistic, tail probability, high-dimensional data, network data, non-Euclidean data

Abstract

Recent technological advances allow for the collection of massive data in the study of complex phenomena over time and/or space in various fields. Many of these data involve sequences of high-dimensional or non-Euclidean measurements, where change-point analysis is a crucial early step in understanding the data. Segmentation, or offline change-point analysis, divides data into homogeneous temporal or spatial segments, making subsequent analysis easier; its online counterpart detects changes in sequentially observed data, allowing for real-time anomaly detection. This article reviews a nonparametric change-point analysis framework that utilizes graphs representing the similarity between observations. This framework can be applied to data as long as a reasonable dissimilarity distance among the observations can be defined. Thus, this framework can be applied to a wide range of applications, from high-dimensional data to non-Euclidean data, such as imaging data or network data. In addition, analytic formulas can be derived to control the false discoveries, making them easy off-the-shelf data analysis tools.

1. INTRODUCTION

Given recent technological advancements, scientists in many fields are collecting massive data for studying complex phenomena over time and/or space. Such data often involve sequences of high-dimensional or non-Euclidean measurements that cannot be analyzed through traditional approaches. Insights on such data often come from segmentation, or offline change-point analysis, which divides a completely observed sequence into homogeneous temporal or spatial segments, or its online counterpart, which detects changes in sequentially observed data. They are crucial early steps in understanding the data and in detecting anomalous events.

Change-point analysis has been extensively studied for univariate and low-dimensional data (for various aspects of classic change-point analysis, see Siegmund 1985, Basseville & Nikiforov 1993, Brodsky & Darkhovsky 1993, Carlstein et al. 1994, Csörgö & Horváth 1997, Chen & Gupta 2000). However, many applications involve moderate- to high-dimensional data or even non-Euclidean data, including the following:

- Network evolution: Data on networks have become increasingly common. For example, emails, phone or online chat records, and records of communications within scientific collaborations can be used to construct networks of social interactions among individuals (Barabâsi et al. 2002, Kossinets & Watts 2006, Eagle et al. 2009). High-throughput biological experiments have led to the ubiquitous study of protein- or gene-interaction networks (Wagner 2001, Pastor-Satorras et al. 2003, Huang et al. 2009). A large part of these studies is characterizing how the network evolves over time, for example, whether there is an abrupt shift in network connectivity at any given time. Here, the observation at each time point is a graphical encoding of the network.
- Image analysis: Image data collected over time appear in diverse applications, from neuroscience (Cabeza & Nyberg 2000) to video surveillance (Collins et al. 2000) to climatology (Long et al. 2001). The detection of abrupt events, such as regional brain activation/deactivation, security breaches, or storms, can be formulated as a change-point problem (Radke et al. 2005, Tian et al. 2005). Temporal video segmentation is also common in indexing, annotating, and retrieving digital materials (Koprinska & Carrato 2001, Li et al. 2002, Guimarães et al. 2003). In all these applications, the data at each time point consist of the digital encoding of an image.
- Sequence or text analysis: In genomic sequence analysis, it is often of interest to find regions of the genome with different DNA-word compositions, such as regions from external sources caused by horizontal gene transfer (Tsirigos & Rigoutsos 2005). Similar problems arise in text analysis. For example, many classic works in both western and eastern literature have ongoing authorship debates (Guia & Wittlin 1999, Riba & Ginebra 2006, Hu et al. 2014). A data-driven approach is to statistically test for abrupt changes in writing style, which can be reflected by word usage. In both settings, each observation in the sequence is a vector of word counts over a large dictionary of words.
- Multiple sensor detection: In a sensor network, hundreds or thousands of sensors are deployed to detect events of interest. For example, hundreds of monitors are placed worldwide to detect solar flares, which are large energy releases from the Sun and can affect Earth's ionosphere and disrupt long-range radio communications (Kappenman 2012, Qu et al. 2005). Many times, the structure of the sensor network can be used to boost the power of the detection. Here, the observation can be viewed as a structured vector or a semistructured vector.
- Transportation data: Volumes of transportation data are collected over time, including ride-sharing and taxi data. For example, the New York City Taxi & Limousine Commission

provides public information on taxi pickup and drop-off dates/times, longitude and latitude coordinates of pickup and drop-off locations, trip distances, and driver-reported passenger counts. These datasets can be analyzed for disruptions or changes in traffic patterns. Here, the observation could be a grid of longitude-latitude points, with each cell representing the frequency of rides in that area for a particular unit of time.

- Neuropixels data: Electrophysiological recording techniques have become more sophisticated, incorporating simultaneous spiking data from increasing numbers of neurons across multiple brain regions. In particular, hair-thin probes densely packed with hundreds of recording sites, called Neuropixels, can record spiking activity from hundreds or even thousands of cells (Jun et al. 2017, Stringer et al. 2019). The combined high temporal resolution and broad spatial coverage of these probes offer a new picture of the coordinated activity in the brain. Here, the observation is a high-dimensional vector, with the coordinates of the vector dependent in an unknown way.

The field of change-point analysis is thriving, given the challenges arising in various fields. Now, the ability to deal with high-throughput data and data with complicated structures is becoming a necessity. Recent developments include the introduction of faster algorithms (Killick et al. 2012, Niu & Zhang 2012, Celisse et al. 2018), effective ways of detecting multiple change-points (Fryzlewicz 2014, 2020; Frick et al. 2014; Zou et al. 2020; Chen et al. 2021), simultaneously detecting change-points in multiple sequences (Zhang et al. 2010, Xie & Siegmund 2013, Chan & Walther 2015, Wang & Samworth 2018), and nonparametric approaches with mild conditions on the data (Desobry et al. 2005, Harchaoui et al. 2009, Lung-Yut-Fong et al. 2011, Matteson & James 2014).

However, methods for multivariate data are limited in many ways, and there is little research for non-Euclidean data. For change-point analysis on multivariate data, most work is based on parametric methods. For example, the problem of detecting common mean shifts in a sequence of independent multivariate observations has been studied under the assumption of multivariate Gaussian distribution with identity covariance (Zhang et al. 2010, Siegmund et al. 2011, Xie & Siegmund 2013) and with general covariance (Srivastava & Worsley 1986, James et al. 1992). For sequences of networks, parametric methods have been proposed that make specific assumptions about the underlying network structure (Wang et al. 2014). In general, parametric change-point tests for multivariate/non-Euclidean data only work under stringent assumptions and are not robust to the violation of these assumptions. Also, existing parametric methods cannot be applied to very high dimensions unless strong assumptions are made to avoid the estimation of a large number of nuisance parameters. In the nonparametric context, methods based on kernels (Desobry et al. 2005, Harchaoui et al. 2009, Garreau & Arlot 2018, Arlot et al. 2019, Chang et al. 2019), marginal rank statistics (Lung-Yut-Fong et al. 2011), and *U*-statistics (Matteson & James 2014) were proposed. These are more broadly applicable than parametric methods. However, these nonparametric tests did not offer a fast analytical formula for false positive control.

In many modern data applications, due to the large volume of the data sequences, one would like to have fast ways, such as analytic *p*-value approximations, of controlling type I errors in change detection to make the method useful for large datasets. This was equipped in many parametric approaches, such as those of Zhang et al. (2010), Niu & Zhang (2012), and Xie & Siegmund (2013). However, the parametric approaches with fast type I error control either target univariate/low-dimensional data, or analyze high-dimensional data but require the coordinates of the high-dimensional vector to be independent, which is far from an adequate way to analyze many modern data sequences. For nonparametric methods applicable for high-dimensional data and beyond, it is usually difficult to provide a generic analytic way to control type I error.

Chen & Zhang (2015) proposed a new nonparametric approach that utilizes similarity information among observations—a similarity graph is constructed on the observations, and the test statistic is defined on the graph. The authors worked out a way to derive analytic formulas to approximate the permutation p-values, making the method straightforward to apply to large datasets. Later, Chu & Chen (2019) proposed new test statistics that could detect more general changes. Both Chen & Zhang (2015) and Chu & Chen (2019) focused on offline change-point detection. In some applications, it is important to detect changes on the fly. Chen (2019b) solved this issue by proposing an online approach based on k-nearest neighbors (k-NNs), and Chu & Chen (2022) proposed additional stopping rules to improve the detection power. Since all these methods utilize a similarity graph constructed on the observations, we refer to them as graph-based change-point methods. In the following, these methods are discussed in detail. The methods described below are implemented in R packages `gSeg` (for offline detection) and `gStream` (for online detection).

2. OFFLINE GRAPH-BASED CHANGE-POINT FRAMEWORK

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ be the data sequence, where $\mathbf{y}_t$ could lie in a high-dimensional or a non-Euclidean space. Here, we focus on the single change-point alternative to illustrate the idea—that is, there possibly exists a time τ such that $\mathbf{y}_t$ has one distribution for $t \leq \tau$ and another distribution for $t > \tau$. The changed interval alternative, where there exists a time interval $(\tau_1, \tau_2]$ such that $\mathbf{y}_t$ has one distribution for $t \in (\tau_1, \tau_2]$ and possibly another distribution for $t \notin (\tau_1, \tau_2]$, has similar fundamental ideas (for details, see Chen & Zhang 2015, Chu & Chen 2019).

2.1. Graph-Based Test Statistics

The building blocks of the graph-based change-point detection framework are graph-based two-sample tests, which are tests based on a similarity graph constructed on all observations, with each observation a node in the graph. The similarity graph can be a given graph that reflects the similarity between observations (Chen & Zhang 2013). More generally, it can be constructed based on a similarity measure through a certain criterion, such as a minimum spanning tree (MST) (Friedman & Rafsky 1979), which is a tree connecting all observations with the total distance across edges minimized, a minimum distance pairing (Rosenbaum 2005), a nearest neighbor graph (Henze 1988), or a graph constructed from domain knowledge. Let G be the similarity graph on all observations in the sequence. Four statistics are considered by Chen & Zhang (2015) and Chu & Chen (2019), and they are all based on three quantities computed from the graph. Let $g_i(t) = \mathrm{I}(i > t)$, where $\mathrm{I}(\cdot)$ is the indicator function that takes value 1 if event A is true and 0 otherwise. The three quantities are as follows:

$$R_0(t) = \sum_{(i,j)\in G} \mathrm{I}(g_i(t) \neq g_j(t)),$$

$$R_1(t) = \sum_{(i,j)\in G} \mathrm{I}(g_i(t) = g_j(t) = 0), \quad \text{and}$$

$$R_2(t) = \sum_{(i,j)\in G} \mathrm{I}(g_i(t) = g_j(t) = 1).$$

Since each t divides the observations into two samples, these three quantities are the number of edges connecting observations between the two samples ($R_0(t)$) or within each sample ($R_1(t), R_2(t)$). **Figure 1** illustrates the computation of $R_i(t)$, $i = 0, 1, 2$ on a small artificial dataset.

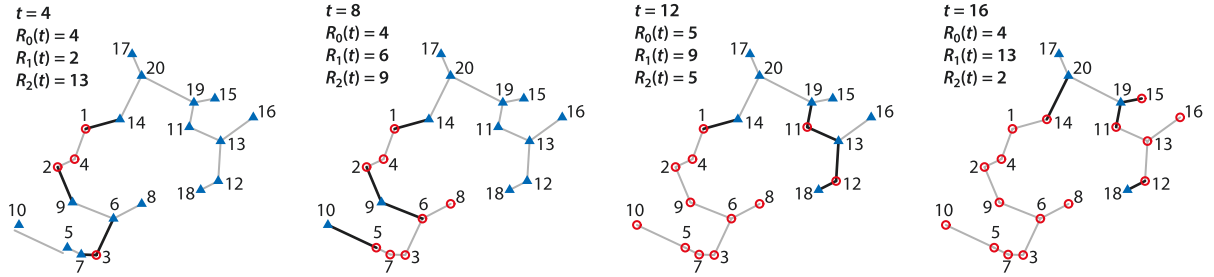

Figure 1

The computation of $R_0(t)$, $R_1(t)$, and $R_2(t)$ for four different times t on a small artificial dataset of length $n = 20$, with G the minimum spanning tree on the Euclidean distance. Each point corresponds to one observation labeled by its index. The first 10 points are randomly drawn from $\mathcal{N}(\mathbf{0}, I_2)$, and the next 10 points are randomly drawn from $\mathcal{N}((2,2)^T, I_2)$, where $\mathcal{N}$ is the normal distribution. Each t divides the observations into two groups, one group for observations before and at t (*red circles*) and the other group for observations after t (*blue triangles*). We see that G, the graph in each panel, does not change as t changes, but the group identities of some observations change, causing $R_0(t)$ (*bold edges*), $R_1(t)$ (*edges connecting the circles*), and $R_2(t)$ (*edges connecting the triangles*) to change.

The four statistics considered are listed below:

$$Z_0(t) = -\frac{R_0(t) - \mathrm{E}(R_0(t))}{\sqrt{\mathrm{Var}(R_0(t))}}, \tag{1.}$$

$$Z_{\mathrm{w}}(t) = \frac{R_{\mathrm{w}}(t) - \mathrm{E}(R_{\mathrm{w}}(t))}{\sqrt{\mathrm{Var}(R_{\mathrm{w}}(t))}}, \quad R_{\mathrm{w}}(t) = \frac{n-t-1}{n-2}R_1(t) + \frac{t-1}{n-2}R_2(t), \tag{2.}$$

$$S(t) = \begin{pmatrix} R_1(t) - \mathrm{E}(R_1(t)) \\ R_2(t) - \mathrm{E}(R_2(t)) \end{pmatrix}^T \Sigma_R^{-1} \begin{pmatrix} R_1(t) - \mathrm{E}(R_1(t)) \\ R_2(t) - \mathrm{E}(R_2(t)) \end{pmatrix}, \quad \Sigma_R = \mathrm{Var}\left(\begin{pmatrix} R_1(t) \\ R_2(t) \end{pmatrix} \right), \text{ and} \tag{3.}$$

$$M(t) = \max(Z_{\mathrm{w}}(t), |Z_{\mathrm{diff}}(t)|), \quad Z_{\mathrm{diff}}(t) = \frac{R_d(t) - \mathrm{E}(R_d(t))}{\sqrt{\mathrm{Var}(R_d(t))}}, \quad R_d(t) = R_1(t) - R_2(t), \tag{4.}$$

where the expectation and variance are defined under the permutation null distribution, i.e., $1/n!$ probability is placed on each of the $n!$ permutations of $\{\mathbf{y}_t : t = 1, \ldots, n\}$. These four statistics are referred to as original, weighted, generalized edge-count, and max-type edge-count statistics, respectively. They each have certain advantages under different scenarios. The original and weighted edge-count statistics, $Z_0(t)$ and $Z_{\mathrm{w}}(t)$, aim to detect mean shifts. The original edge-count statistic, $Z_0(t)$, tends to do well when the change in mean is near the center of the sequence but suffers from a variance boosting problem when the number of observations before and after t are unequal. The weighted edge-count statistic, $Z_{\mathrm{w}}(t)$, resolves this variance boosting problem by weighting $R_1(t)$ and $R_2(t)$ with the inverse of their corresponding sample sizes. Observe that $Z_0(t)$ and $Z_{\mathrm{w}}(t)$ are equivalent when the number of observations before and after t are equal.

Both $Z_0(t)$ and $Z_{\mathrm{w}}(t)$ are designed so that a relatively small $R_0(t)$ [or relatively large $R_1(t)$ and $R_2(t)$] provide evidence against the null hypothesis of no distributional difference. The intuition is straightforward: If the observations before and after t really do come from different distributions, then observations would tend to be closer to those from the same distribution, resulting in a relatively small R_0 at t. This rationale holds particularly well when the change is in mean only and/or the data are relatively low dimensional. However, when the dimension of the data is moderate or high and the change in distribution is not only in mean—for example, a change in variance is also present—this rationale breaks down. This is a result of the curse

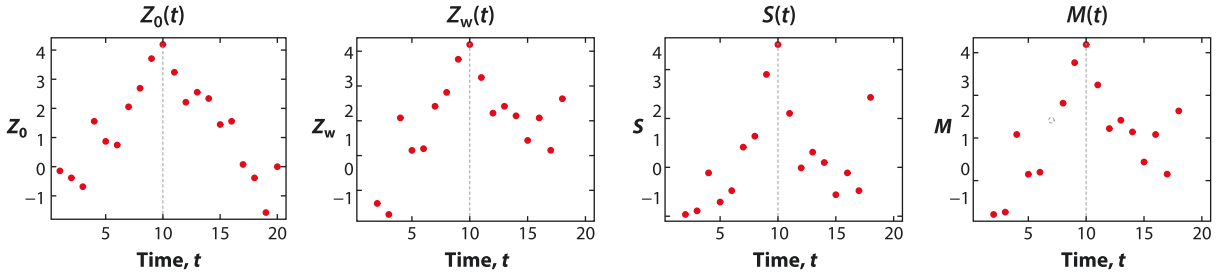

Figure 2

The profiles of original, weighted, generalized, and max-type edge-count statistics (Z_0, Z_w, S, and M, respectively) over time t for the small artificial dataset in **Figure 1**.

of dimensionality; in high-dimensional space, observations that are similar (i.e., from the same distribution) are not necessarily close in distance. To resolve this, the test statistics $S(t)$ and $M(t)$ were proposed to target more general changes. The generalized edge-count statistic $S(t)$ is defined to be more robust to the curse of dimensionality: It allows either direction of $R_1(t) - \mathrm{E}(R_1(t))$ and $R_2(t) - \mathrm{E}(R_2(t))$ to contribute to the test statistic. Interestingly, $S(t)$ can be decomposed into two uncorrelated quantities, $Z_w(t)$ and $Z_{\mathrm{diff}}(t)$, where $Z_w(t)$ is sensitive to changes in location and $Z_{\mathrm{diff}}(t)$ is sensitive to changes in scale (see Chu & Chen 2019, lemma 3.1). This inspired the max-type edge-count statistic $M(t)$, which takes the maximum of $Z_w(t)$ and $|Z_{\mathrm{diff}}(t)|$. Both S and M are recommended for general changes, and their power performance tends to be similar. For more detailed reasoning and comparisons of the test statistics, readers are directed to Friedman & Rafsky (1979), Chen & Zhang (2015), Chen & Friedman (2017), Chen et al. (2018), and Chu & Chen (2019).

The profiles of the four scan statistics on the artificial dataset in **Figure 1** are shown in **Figure 2**. It is clear that the scan statistics all achieve a maximum at $\tau = 10$:

$$\max_{n_0 \le t \le n_1} Z_0(t), \quad \max_{n_0 \le t \le n_1} Z_w(t), \quad \max_{n_0 \le t \le n_1} S(t), \quad \text{and} \quad \max_{n_0 \le t \le n_1} M(t) \quad (n_0, n_1 \text{ prespecified}).$$

2.2. Analytical *p*-Value Approximations

The null hypothesis of no change-point is rejected when the maximum scan statistic is greater than some threshold. When n is small, this threshold could be determined by performing random permutations directly. However, when n is large, this becomes computationally prohibitive, and Chen & Zhang (2015) and Chu & Chen (2019) provided accurate analytic formulas to approximate the permutation p-value for each of the scan statistics, allowing fast application of the methods.

To obtain these analytical p-value approximations, the asymptotic properties of the stochastic processes $Z_0(t)$, $Z_w(t)$, $S(t)$, and $M(t)$ were studied. Since $S(t)$ can be decomposed into the sum of squares of $Z_w(t)$ and $Z_{\mathrm{diff}}(t)$, and due to the way $M(t)$ is defined, this boils down to studying the basic processes $Z_w(t)$, and $Z_{\mathrm{diff}}(t)$. Under certain conditions on the graph, the limiting distributions of $\{Z_0([nu])\colon 0 < u < 1\}$, $\{Z_{\mathrm{diff}}([nu])\colon 0 < u < 1\}$, and $\{Z_w([nu])\colon 0 < u < 1\}$ converge to Gaussian processes in finite-dimensional distributions (Chen & Zhang 2015, theorem 3.1, and Chu & Chen 2019, theorem 4.1). The proofs for these theorems utilize Stein's method (Chen & Shao 2005). Explicit expressions for the covariance functions of the Gaussian processes are also derived through combinatorial analysis (Chen & Zhang 2015, lemma 3.3, and Chu & Chen 2019, theorem 4.3).

Using these results, the asymptotic approximations of the tail probabilities are derived:

$$\mathrm{P}\left(\max_{n_0\le t\le n_1} Z_0(t) > b\right) \approx b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_0^*(x)\nu(b\sqrt{2h_0^*(x)/n})\mathrm{d}x, \quad 5.$$

$$\mathrm{P}\left(\max_{n_0\le t\le n_1} Z_{\mathrm{w}}(t) > b\right) \approx b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_{\mathrm{w}}^*(x)\nu(b\sqrt{2h_{\mathrm{w}}^*(x)/n})\mathrm{d}x, \quad 6.$$

$$\mathrm{P}\left(\max_{n_0\le t\le n_1} S(t) > b\right) \approx \frac{b\,e^{-b/2}}{2\pi}\int_0^{2\pi}\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} u^*(x,\omega)\nu(\sqrt{2b\,u^*(x,\omega)/n})\mathrm{d}x\mathrm{d}\omega, \quad \text{and} \quad 7.$$

$$\mathrm{P}\left(\max_{n_0\le t\le n_1} M(t) > b\right) = 1 - \mathrm{P}\left(\max_{n_0\le t\le n_1} |Z_{\mathrm{diff}}(t)| < b\right)\mathrm{P}\left(\max_{n_0\le t\le n_1} Z_{\mathrm{w}}(t) < b\right), \quad 8.$$

where $\mathrm{P}\left(\max_{n_0\le t\le n_1}|Z_{\mathrm{diff}}(t)| < b\right) \approx 1 - 2b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_{\mathrm{diff}}^*(x)\nu(b\sqrt{2h_{\mathrm{diff}}^*(x)/n})\mathrm{d}x$.

The functions $h_0^*(x)$, $h_{\mathrm{w}}^*(x)$, and $h_{\mathrm{diff}}^*(x)$ capture the autocorrelation of the processes $Z_0(t)$, $Z_{\mathrm{w}}(t)$, and $Z_{\mathrm{diff}}(t)$, respectively, and are defined as follows:

$$h_0^*(x) = \frac{1}{2x(1-x)} + \frac{2}{4x(1-x)+(1-2x)^2(r_1-4r_0)},$$

$$h_{\mathrm{w}}^*(x) = \frac{1}{x(1-x)}, \quad \text{and}$$

$$h_{\mathrm{diff}}^*(x) = \frac{1}{2x(1-x)},$$

with $r_0 := \lim_{n\to\infty}|G|/n, r_1 := \lim_{n\to\infty}\sum_{i=1}^n |G_i|^2/|G|$, and $u^*(x,\omega) = h_{\mathrm{w}}^*(x)\sin^2(\omega) + h_{\mathrm{diff}}^*(x)\cos^2(\omega)$. Observe that the p-value approximations for $Z_{\mathrm{w}}(t)$, $S(t)$, and $M(t)$ are distribution-free and do not depend on the underlying similarity graph G at all.

The approximations also require the function $\nu(x)$. This function is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A simple approximation given by Siegmund et al. (2007) is sufficient for numerical purposes:

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2)-0.5)}{(x/2)\Phi(x/2)+\phi(x/2)}, \quad 9.$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative density function and standard normal density function, respectively.

For finite sample sizes, convergence of the limiting distributions to normal can be slow, especially near the boundaries of the sequence. This problem can become more severe when the dimension is high. **Figure 3** plots the skewness of $Z_0(t)$, $Z_{\mathrm{w}}(t)$, and $Z_{\mathrm{diff}}(t)$. It is clear that when t moves away from the center of the sequence, the statistic $Z_{\mathrm{w}}(t)$ is right skewed and $Z_0(t)$ is left skewed. On the other hand, $Z_{\mathrm{diff}}(t)$ is right skewed for small values of t and left skewed for large values of t. To make the analytical p-value approximations practical for finite sample sizes, adjustments are made in the form of skewness correction. Since the extent of the skewness depends on t, we adopt a skewness correction approach that does the correction up to different extents based on the amount of skewness at each value of t. The approach aims to provide a better approximation to the marginal distributions $\mathrm{P}(Z_0(t) > b)$, $\mathrm{P}(Z_{\mathrm{w}}(t) > b)$, and $\mathrm{P}(Z_{\mathrm{diff}}(t) > b)$. This skewness correction involves computing the third moment of the test statistics, which can be done using combinatorial analysis. By incorporating the skewness correction, approximations to the tail probabilities are improved for the test statistics $Z_0(t)$, $Z_{\mathrm{w}}(t)$, and $M(t)$. Although skewness correction can be done for

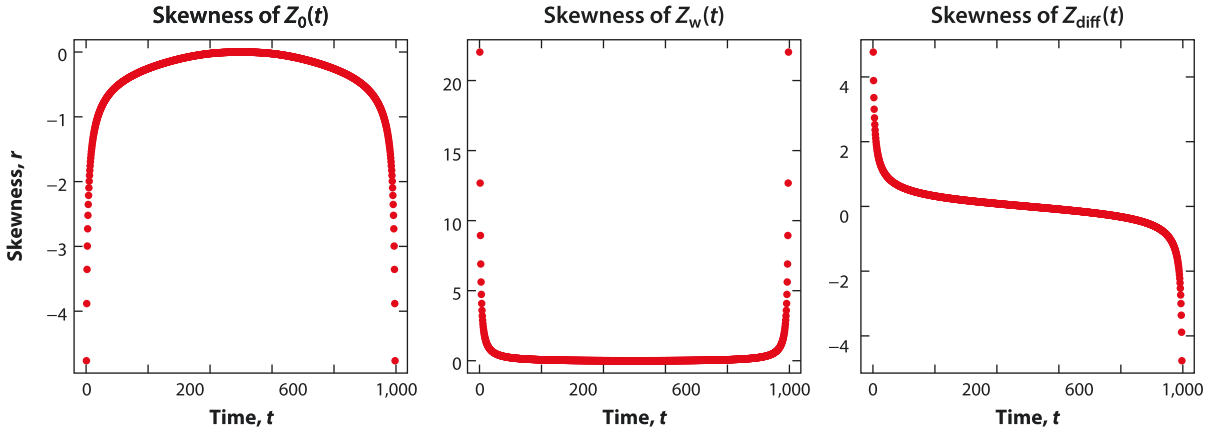

Figure 3

Plots of skewness of $Z_0(t)$, $Z_w(t)$, and $Z_{diff}(t)$ against t for a sequence of 1,000 observations randomly generated from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{100})$. The graph is the minimum spanning tree constructed on the L_2 distance.

$S(t)$, the approach relies heavily on extrapolation. Therefore, while both $S(t)$ and $M(t)$ can be used to detect general changes, $M(t)$ is often preferred since a more accurate p-value approximation can be obtained by incorporating skewness correction.

3. ONLINE GRAPH-BASED CHANGE-POINT FRAMEWORK

For online detection, observations are continually arriving, and the aim is to detect changes on the fly. Examples include fraud detection (Bolton & Hand 2001, Chandola et al. 2009), disease surveillance and medical monitoring (Wong et al. 2003, Pervaiz et al. 2012, Malladi et al. 2013, Zhang et al. 2013, Dehning et al. 2020), and network intrusions (Tartakovsky et al. 2006, Xie et al. 2011). The online detection setting can be formulated as follows: The observations $\mathbf{Y}_t, t = 1, 2, \ldots, n, \ldots$ are identically distributed from an unknown distribution F_0. If there is a change-point at τ, the observations after τ are from a different unknown distribution F_1. No constraints on how the change happens are imposed. 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.

The setup is as follows: We assume that there are N_0 historical observations, and we begin the online testing from observation $N_0 + 1$ until the test declares a change happened. Since the similarity graph updates as new observations arrive, the main challenge in extending the offline framework to online is to understand the dynamics of the series of similarity graphs. Chen (2019b) considered similarity structure represented by nearest neighbors (NNs). It turns out that the dynamics of NN graphs can be characterized by a small number of events: the updates of mutual NNs, shared NNs, and all three-way interactions among the NN relations. Given the data and a similarity measure, these events can be directly analyzed and analytical expressions for these events can be derived. Concrete stopping rules that incorporate these quantities are proposed for Z_0 by Chen (2019b) and versions for Z_w, S, and M are extended by Chu & Chen (2022). For all the stopping rules, analytical formulas for false discovery control are also derived.

3.1. Comparison of Stopping Rules

We first discuss the rationale for how the stopping rules are constructed. Let n be the index of the observation we are currently observing, and let $Z_{|y}(t, n)$ be the online version of $Z_0(t)$, which is the

standardized between-sample edge-count. Chen (2019b) considered three stopping rules:

$$T_1(b_1) = \inf\left\{n - N_0 : \max_{n_0 \le t \le n-n_0} Z_{|\mathbf{y}}(t,n) > b_1, n \ge N_0\right\}, \tag{10.}$$

$$T_2(b_2) = \inf\left\{n - N_0 : \max_{n-n_1 \le t \le n-n_0} Z_{|\mathbf{y}}(t,n) > b_2, n \ge N_0\right\}, \quad \text{and} \tag{11.}$$

$$T_3(b_3) = \inf\left\{n - N_0 : \max_{n-n_1 \le t \le n-n_0} Z_{L|\mathbf{y}}(t,n) > b_3, n \ge N_0\right\}, \tag{12.}$$

with n_0, n_1, and L prespecified. Here, b_1, b_2, and b_3 are chosen so that the false discovery rate for each stopping rule is controlled at a prespecified level. The stopping rule T_1 is fairly straightforward: n_0 is chosen so as to avoid fluctuations in the test statistic at the boundaries while still retaining fast detection. However, T_1 keeps early observations as candidate change-points, which could result in efficiency loss when n is large. Therefore, T_2 only keeps those more recent observations as candidate change-points. For both T_1 and T_2, at time n, the k-NNs are considered from among all n observations. Another modification is to construct the directed k-NN graphs from only the L most recent observations $Y_{n-L+1}, \ldots, Y_n$; this gives rise to T_3, with $Z_{L|\mathbf{y}}(t,n)$ denoting the between-sample edge-count two-sample test constructed from only the L most recent observations.

The performances of stopping rules are conventionally evaluated by their detection delay, defined as the time elapsed between when the change occurs and its detection, while controlling the false discovery rate at a prespecified level. Chen (2019b) compared the average detection delay of the three stopping rules and found that T_3 has some advantages. Since T_3 is only constructed from the L most recent observations, it is not affected by where τ is located. In contrast, the detection delay of T_1 and T_2 can be affected by the location of the change-point in the sequence. Moreover, T_3 retains a computational advantage since only the L most recent observations need to be stored. In what follows, the new stopping rules refer to T_3 such that graph-based test statistics are constructed on the L most recent observations.

3.2. Sequential Detection Based on *k*-Nearest Neighbors

For online detection, the test statistics are analogous to their offline counterparts. We denote $Z_{L|\mathbf{y}}$, $S_{L|\mathbf{y}}$, $W_{L|\mathbf{y}}$, and $M_{L|\mathbf{y}}$ to be the online versions of Z_0, S, Z_w, and M, respectively. Key differences are that the test statistics are defined explicitly for directed k-NN graphs, and the directed k-NN graph is constructed on only the most recent L observations: $\mathbf{Y}_{n-L+1}, \ldots, \mathbf{Y}_n$, where n denotes the current observation we are observing. Specifically, for any $n > N_0$ and $i, j \in n_L \triangleq \{n - L + 1, \ldots, n\}$, we let

$$A^{(r)}_{n_L,ij} = \mathrm{I} \quad (\mathbf{Y}_j \text{ is the } r\text{th-NN of } \mathbf{Y}_i \text{ among } \mathbf{Y}_{n-L+1}, \ldots, \mathbf{Y}_n).$$

In terms of graph construction, each observation points to its k NNs. For example, if $A^{(r)}_{n_L,ij} = 1$, then Y_j is the rth NN of Y_i and there is a directed edge from Y_i pointing to 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 $a^{(r)}_{n_L,ij} = \mathrm{I}\,(\mathbf{y}_j \text{ is the } r\text{th NN of } \mathbf{y}_i \text{ among } \mathbf{y}_{n-L+1}, \ldots, \mathbf{y}_n)$.

For any n, each $t \in \{n - L + 1, \ldots, n\}$ divides the data sequence into two groups: those observations between $n - L + 1$ and t (sample 1) and those observations after t (sample 2). Let $B_{0,ij}(t, n_L)$, $B_{1,ij}(t, n_L)$, and $B_{2,ij}(t, n_L)$ be the indicator random variables that denote whether the observations $\mathbf{Y}_i$ and $\mathbf{Y}_j$ belong to different samples, both belong to sample 1, or both belong to

sample 2, respectively. The graph-based quantities are defined as

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

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

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

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 before t, and $R_{2,L}(t,n)$ is twice the number of edges connecting observations after t. Then $Z_{L|\mathbf{y}}$, $S_{L|\mathbf{y}}$, $W_{L|\mathbf{y}}$, and $M_{L|\mathbf{y}}$ can be constructed from these quantities in a similar way as their offline versions. Under the permutation distribution, the analytical expressions for the expectation and variance can be derived, and explicit expressions are provided by Chen (2019b) and Chu & Chen (2022).

The stopping rules based on the graph-based test statistics under k-NN are defined as follows:

$$T_Z(b_Z) = \inf\left\{n - N_0 : \max_{n-n_1 \le t \le n-n_0} Z_{L|\mathbf{y}}(t,n) > b_Z\right\}, \tag{13.}$$

$$T_S(b_S) = \inf\left\{n - N_0 : \max_{n-n_1 \le t \le n-n_0} S_{L|\mathbf{y}}(t,n) > b_S\right\}, \tag{14.}$$

$$T_W(b_W) = \inf\left\{n - N_0 : \max_{n-n_1 \le t \le n-n_0} W_{L|\mathbf{y}}(t,n) > b_W\right\}, \quad \text{and} \tag{15.}$$

$$T_M(b_M) = \inf\left\{n - N_0 : \max_{n-n_1 \le t \le n-n_0} M_{L|\mathbf{y}}(t,n) > b_M\right\}. \tag{16.}$$

The rationale of the test statistics is the same as in the offline setting, and each stopping rule has a niche where it dominates. For mean changes, we recommend using the stopping rule based on $W_{L|\mathbf{y}}$ since it has shorter detection delay and higher power compared with $Z_{L|\mathbf{y}}$. For general changes, the stopping rules based on $S_{L|\mathbf{y}}$ and $M_{L|\mathbf{y}}$ can be used.

3.3. Average Run Length

We would like to determine the thresholds b_Z, b_S, b_W, and b_M in a way that controls the false discovery rate. In the online setting, a common way to measure the false discovery rate is the average run length, i.e., the expected time to stop when there is no change-point. Therefore, we would like to choose the thresholds so that each of $\mathrm{E}_\infty(T_Z(b_Z))$, $\mathrm{E}_\infty(T_S(b_S))$, $\mathrm{E}_\infty(T_W(b_W))$, and $\mathrm{E}_\infty(T_M(b_M))$ is a prespecified value, for example, 10,000.

When the underlying distribution of the sequence is known, the thresholds can be obtained via Monte Carlo simulations. However, in many applications, the distribution of the sequence is unknown. Furthermore, since new observations keep arriving, resampling-based methods, such as permutation or bootstrap, are not applicable. Therefore, to make the method useful for real applications, analytical expressions for the average run lengths were derived. To obtain these expressions, the limiting distribution of the basic random fields $\{Z_{L|\mathbf{y}}(t,n)\}$, $\{D_{L|\mathbf{y}}(t,n)\}$, and $\{W_{L|\mathbf{y}}(t,n)\}$ were studied, where $\{D_{L|\mathbf{y}}(t,n)\}$ is the online version of Z_{diff}. Under conditions on the graph, $\{Z_{L|\mathbf{y}}(t,n)\}$, $\{D_{L|\mathbf{y}}(t,n)\}$, and $\{W_{L|\mathbf{y}}(t,n)\}$ converge to two-dimensional Gaussian random fields in finite-dimensional distributions. To fully specify the Gaussian random fields, the

covariance functions of the processes are derived. This involves studying the dynamics of the k-NN series as the new observations are added. It turns out that a few key quantities are enough to characterize their dynamics, i.e.,

$$X_1 = \sum_{i,j \in m_L \cap n_L} \left(A^+_{m_L,ij} + A^+_{m_L,ji}\right)\left(A^+_{n_L,ij} + A^+_{n_L,ji}\right),$$

$$X_2 = \sum_{i \in m_L \cap n_L;\, j \in m_L;\, l \in n_L} \left(A^+_{m_L,ij} + A^+_{m_L,ji}\right)\left(A^+_{n_L,il} + A^+_{n_L,li}\right), \quad \text{and}$$

$$X_3 = \sum_{i,j \in m_L \cap n_L; l,r \in n_L} \left(A^+_{m_L,ij} + A^+_{m_L,ji}\right)\left(A^+_{n_L,lr} + A^+_{n_L,rl}\right),$$

with $m_L \triangleq \{m - L + 1, \ldots, m\}$ and $m < n$. These boil down to obtaining analytical expressions for the updates of mutual NNs ($\mathrm{E}(X_1)$), shared NNs ($\mathrm{E}(X_2)$), and three-way interactions of NNs ($\mathrm{E}(X_3)$). We refer readers to Chen (2019b) and Chu & Chen (2022) for a more technical treatment of the asymptotic results.

The analytical expressions for the average run lengths are

$$\mathrm{E}_\infty(T_Z(b_Z)) \approx \frac{L\sqrt{2\pi}\exp(b_Z^2/2)}{b_Z^3 \int_{n_0/L}^{n_1/L} g_{Z,1}(x) g_{Z,2}(x) \nu(\sqrt{2b_Z^2 2 g_{Z,1}(x)/L}) \nu(\sqrt{2b_Z^2 g_{Z,2}(x)/L}) \mathrm{d}x}, \tag{17.}$$

$$\mathrm{E}_\infty(T_W(b_W)) \approx \frac{L\sqrt{2\pi}\exp(b_W^2/2)}{b_W^3 \int_{n_0/L}^{n_1/L} g_{W,1}(x) g_{W,2}(x) \nu(\sqrt{2b_W^2 2 g_{W,1}(x)/L}) \nu(\sqrt{2b_W^2 g_{W,2}(x)/L}) \mathrm{d}x}, \tag{18.}$$

$$\mathrm{E}_\infty(T_S(b_S)) \approx \frac{\pi \exp(b_S/2)}{b_S^2 \int_0^{2\pi} \int_{n_0/L}^{n_1/L} h_1(x,\omega) h_2(x,\omega) \nu(\sqrt{2b_S h_1(x,\omega)/L}) \nu(\sqrt{2b_S h_2(x,\omega)/L}) \mathrm{d}x \mathrm{d}\omega}, \text{ and} \tag{19.}$$

$$\mathrm{E}_\infty(T_M(b_M)) \approx \frac{L\sqrt{2\pi}\exp(b_M^2/2)}{2b_M^3 \int_{n_0/L}^{n_1/L} g_{D,1}(x) g_{D,2}(x) \nu(\sqrt{2b_D^2 g_{D,1}(x)/L}) \nu\sqrt{2b_D^2 g_{D,2}(x)/L}) \mathrm{d}x}. \tag{20.}$$

Here, $\nu(\cdot)$ is defined as in Equation 9, and

$$g_{Z,1}(x) = \frac{16x(1-x)(k + p_{k,\infty}) + 2(1-2x)^2(q_{k,\infty} - k^2 + k)}{\sigma^2(x)},$$

$$g_{Z,2}(x) = \frac{16x^2(1-x)^2(p_{k,\infty} + q_{k,\infty} + k^2 + 2p^{(k)}_{k,\infty} - 2q^{(k)}_{k,\infty}}{\sigma^2(x)} + \frac{4x(1-x)(2q^{(k)}_{k,\infty} - 3q_{k,\infty} + k^2 + k) + 2(q_{k,\infty} - k^2 + k)}{\sigma^2(x)},$$

$$g_{W,1}(x) = \frac{1}{x(1-x)},$$

$$g_{W,2}(x) = \frac{x^2 - x + 1}{x(1-x)} - \frac{2k\, p^{(k)}_{k+1,\infty}}{k + p_{k,\infty}},$$

$$g_{D,1}(x) = \frac{1}{2x(1-x)},$$

$$g_{D,2}(x) = \frac{10q_{k,\infty} - 4kq^{(k)}_{k+1,\infty} - (6k^2 - 10k)}{2(q_{k,\infty} - k^2 + k)} - \frac{1}{2x(1-x)},$$

where

$$\sigma^2(x) = 4x(1-x)(4x(1-x)(k + p_{k,\infty}) + (1-2u)^2(q_{k,\infty} - k^2 + k)),$$

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

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

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 $q_\infty(r, s)$ is the limiting expected number of nodes shared between the rth and sth NNs. In practice, these quantities can be estimated through historical data and can further be updated by new observations as long as no change-point is detected. To obtain more accurate approximations of the thresholds, skewness correction is also implemented for the stopping rules T_Z, T_W, and T_M.

4. REAL DATA APPLICATIONS

To illustrate the graph-based change-point approach, we apply the test statistics to data obtained from the New York City Taxi & Limousine Commission to see whether or not we can detect a change in travel patterns. This dataset provides information on taxi pickup and drop-off date, time, and longitude-latitude coordinates. We focus only on those taxi trips that began at John F. Kennedy International Airport, and for each trip, we count the number of taxi drop-offs that occur within various locations of New York City. Explicitly, using longitude-latitude coordinates, we create a 30 by 30–cell grid of New York City and count the number of taxi drop-offs that fall within each cell. Then for each day, we have a 30 by 30 matrix such that each matrix element represents the number of taxi drop-offs in each location.

4.1. Offline Setting

In this setting, we are interested in detecting intervals of change, rather than a single change-point. Let A_i be the 30 by 30 matrix on day i. The $L_{1,1}$ norm is used to construct the MST graph representing similarity between days. The results can be seen in **Table 1**. The edge-count statistic $Z_0(t_1, t_2)$ reports November 21–December 31, 2015 (days 52–92), as the changed interval result. The remaining graph-based test statistics all report Christmas and the preceding week, December 18–December 25, 2015 (days 79–86), as the changed interval. All these tests reject the null hypothesis of no change, with p-value < 0.001.

As there might be more than one changed interval, we further perform the tests on the period October 1–December 17, 2015. During this time period, $Z(t_1, t_2)$ selects October 27–December 17, 2015 (days 27–78), as the changed interval, while $S(t_1, t_2)$, $Z_w(t_1, t_2)$, and $M(t_1, t_2)$ all report the week including Thanksgiving, November 20–27, 2015 (days 51–58), as the changed interval. All these tests reject the null hypothesis of no change as well, with p-value < 0.001.

Table 1 Changed interval results and corresponding *p*-values (reported in parentheses) for dataset of NYC taxi pickups from JFK

Time period (2015)	Z_0	Z_w	S	M
Oct. 1–Dec. 31	Nov. 21–Dec. 31 (<0.001)	Dec. 18–Dec. 25 (<0.001)	Dec. 18–Dec. 25 (<0.001)	Dec. 18–Dec. 25 (<0.001)
Oct. 1–Dec. 17	Oct. 27–Dec. 17 (0.0011)	Nov. 20–Nov. 27 (<0.001)	Nov. 20–Nov. 27 (<0.001)	Nov. 20–Nov. 27 (<0.001)
Oct. 1–Nov. 20	Oct. 22–Nov. 19 (0.0017)	Nov. 16–Nov. 19 (0.0414)	Nov. 16–Nov. 19 (0.0109)	Nov. 16–Nov. 19 (0.0428)

Z_0, Z_w, S, and M indicate original, weighted, generalized, and max-type edge-count statistics, respectively. Abbreviations: JFK, John F. Kennedy International Airport; NYC, New York City. Table adapted with permission from Chu & Chen (2019) published in *The Annals of Statistics*.

We further continue this process by performing the test on the period October 1–November 20, 2015. The original edge-count test $Z(t_1, t_2)$ reports a changed interval from October 22–November 19, 2015 (days 22–50). It reject the null hypothesis of no change as well, with a small *p*-value (0.0017). $S(t_1, t_2)$, $Z_w(t_1, t_2)$, and $M(t_1, t_2)$ report a changed interval of November 16–19, 2015 (days 47–50) but fail to reject the null hypothesis at the 0.01 significance level.

From the reported changed intervals, the results from the three new tests are more sensible: the week including Thanksgiving, and Christmas and the preceding week. To perform an additional sanity check, we plot the distance matrix of this whole period (**Figure 4*a***). It is evident that there is some change occurring around day 60 and day 80, matching with the results from the test statistics Z_w, S, and M. On the other hand, the distance matrix for the first 51 days seems much more uniform (**Figure 4*b***).

For comparison, two recently proposed nonparametric change-point methods and one parametric approach were considered as alternative methods to detect changes in the New York City taxi dataset. Arlot et al. (2019) considered a penalized kernel least squares estimator, first proposed by Harchaoui & Cappé (2007) and made computationally more efficient by Celisse et al. (2018).

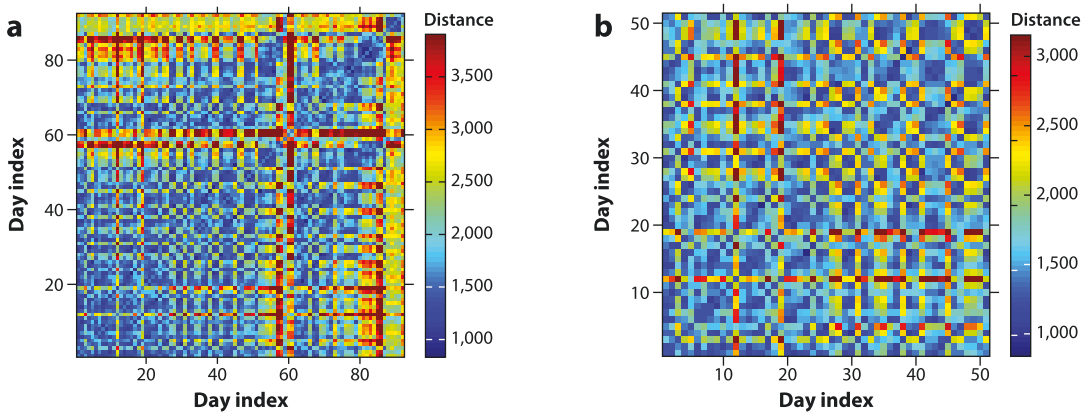

Figure 4

(*a*) Heatmap of $L_{1,1}$ distance matrix for the period of October 1–December 31, 2015 (indexed by 1, . . . ,92). Thanksgiving is day 57 and Christmas is day 86, and it is evident that some changes occur in this timeframe. (*b*) Heatmap of $L_{1,1}$ distance matrix for the period of October 1–November 20, 2015. Figure adapted with permission from Chu & Chen (2019) published in *The Annals of Statistics*.

Table 2 Change-point results for dataset of NYC taxi pickups from JFK

Method	Estimated change-points
Kernel multiple change-point detection (Harchaoui & Cappé 2007, Arlot et al. 2019)	No change-points detected.
Random forest change-point detection (Londschien et al. 2022)	Oct. 20 (day 20), Dec. 7 (day 68), and Dec. 26 (day 87)
Change-point estimation via sparse projection (Wang & Samworth 2018)	NA

The third technique was not suitable for the dataset and returned an error. Abbreviations: JFK, John F. Kennedy International Airport; NA, not applicable; NYC, New York City.

Applied to the New York City taxi dataset, for a range of penalties, the kernel multiple change-point approach either detected almost all time points as change-points or none at all. Given these results, we conclude the approach was unable to recover any meaningful change-points in the sequence. A random forest approach to change-point detection was proposed by Londschien et al. (2022). While the current implementation does not support changed-intervals, these results are somewhat comparable to the results obtained using the newer graph-based test statistics (see **Table 2**). Examining **Figure 4*a***, there does not appear to be any signal around day 20. However, there is a signal around day 60 and day 80, which may be what the Londschien et al. (2022) approach is picking up on for the latter change-points (days 68 and 87). Finally, Wang & Samworth (2018) proposed a parametric approach to high-dimensional change-point estimation via sparse projection. However since their method assumes normality and focuses on detecting sparse mean change, it is not appropriate for the New York City taxi dataset. Moreover, the implementation of their approach involves rescaling the data matrix by estimating the standard deviation through median absolute deviation. For the New York City taxi dataset, which consists of a sequence of sparse count matrices, this can result in a standard deviation of 0, leading to fatal errors in the implementation.

4.2. Online Setting

Here, we focus on trips from two different time periods: the months of June–July and November, 2015. The dataset had been completely collected at the time of analysis. However, we treat it as if the data were being sequentially observed in order to illustrate how the proposed method works.

To detect changes in the months of June–July 2015, we use data from the month of May as historical data. Applying the offline change-point detection method of Chen & Zhang (2015) and Chu & Chen (2019) 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 $n_1 = L - n_0$. We denote as A_i the 30 by 30 matrix on day i. The $L_{1,1}$ norm is used to construct the k-NN graph representing similarity between days. Here, the stopping rules T_W, T_M, and T_S all report a stopping time of July 3 and July 4, whereas T_Z is unable to detect any anomaly event (**Table 3**). The change-point triggering these stopping times is estimated to be June 29. To perform a sanity check, we plot a heatmap of the $L_{1,1}$ distance matrix used to construct the k-NN graph (**Figure 5*a***). Based on the heatmap, we can see there is a clear signal happening around day 60, which corresponds with the results from T_W, T_M, and T_S.

To detect changes in November 2015, we use data from the months of September and October 2015 as historical data. Applying the offline change-point detection method of Chen & Zhang (2015) and Chu & Chen (2019) 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

Table 3 Detected stopping times for dataset of NYC taxi pickups from JFK for May 31–July 31, 2015

	Reported stopping times	Estimated change-point
T_Z	NA	NA
T_W	Jul. 3–4	June 29 (day 60)
T_M	Jul. 3–4	June 29 (day 60)
T_S	Jul. 3–5	June 29 (day 60)

The stopping rule T_Z does not detect any anomaly event. The stopping rules T_W, T_M, and T_S all detect an anomaly event on July 3 and July 4. Abbreviations: JFK, John F. Kennedy International Airport; NA, not applicable; NYC, New York City. Table adapted with permission from Chu & Chen (2022); © 2022 IEEE.

September 1–October 20 as historical observations, and we begin the test at October 21. We set $L = 50$, $n_0 = 8$, and $n_1 = L - n_0$. The stopping times based on T_Z, T_W, T_M, and T_S report back dates that seem to be quite reasonable (see **Table 4**). We see that multiple stopping times are caused by the same anomaly event. When the signal is large enough, the graph-based test statistics perform similarly: All are able to detect a change in travel patterns close to Thanksgiving. Again, to check our results, we plot a heatmap of the $L_{1,1}$ distance matrix used to the construct the k-NN graph (**Figure 5*b***). We can see that there is a clear signal starting roughly around day 82, which matches the results reported from the test statistics.

5. RECENT ADVANCEMENTS

5.1. Dealing with Repeated Observations

The graph-based approach relies on a similarity graph constructed on the observations. When there are repeated observations in the data or ties in the pairwise distance matrix, which is common for discrete data, such as network data, the conventional ways of constructing the similarity graphs, such as the k-MST, are problematic, as there could be multiple optimal graphs according to the graph construction rule and the results on different optimal graphs could be completely different. Song & Chen (2022) explored this problem on a phone-call network dataset studied by Chen & Zhang (2015). This phone-call network dataset was built from a study conducted by the MIT

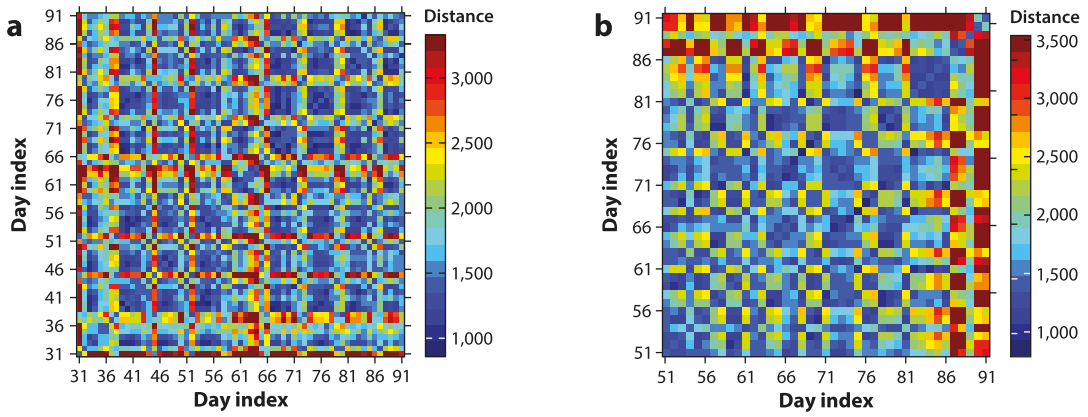

Figure 5

(*a*) Heatmap of $L_{1,1}$ distance matrix for the period of May 31–July 31, 2015 (indexed by 31, . . . , 92). (*b*) Heatmap of $L_{1,1}$ distance matrix for the period of October 21–November 30, 2015 (indexed by 51, . . . , 91). Figure adapted with permission from Chu & Chen (2022); © 2022 IEEE.

Table 4 Detected stopping times for dataset of NYC taxi pickups from JFK for October 21–November 30, 2015

	Reported stopping times	Estimated change-point
T_Z	Nov. 27–31	Nov. 21 (day 82)
T_W	Nov. 28	Nov. 21 (day 82)
T_M	Nov. 28	Nov. 21 (day 82)
T_S	Nov. 27–30	Nov. 21 (day 82), Nov. 23 (day 84)

The stopping times T_Z, T_W, T_M, and T_Z all detect an anomaly event. Abbreviations: JFK, John F. Kennedy International Airport; NA, not applicable; NYC, New York City. Table adapted with permission from Chu & Chen (2022); © 2022 IEEE.

Media Laboratory, in which the phone-call information of participating university students and staff was logged (Eagle et al. 2009). Chen & Zhang (2015) extracted the information of callers and callees and constructed phone-call networks on the participants for each day, with the nodes of the networks indicating the participants and an edge pointing from one node to another if there was a phone call between the two persons. The direction is from the caller to the callee (the phone-call networks on three days are shown in **Figure 6**). The study lasted for 330 days, so the length of the sequence is 330. However, there are only 290 distinct observations, i.e., the phone-call networks on some days are exactly the same, and thus there are many MSTs. Song & Chen (2022) checked the graph-based testing procedures on three randomly chosen MSTs, and their p-values are provided in **Table 5**. It is clear that the conclusion could be completely different even for the same testing procedure when a different MST is used.

To solve this problem, Song & Chen (2022) adopted ideas from Chen & Zhang (2013) and Zhang & Chen (2022) to combine information from these optimal graphs. In particular, the following algorithm is used to construct a candidate graph.

Algorithm 1 (Graph construction for data with repeated observations).

1. Construct a graph on distinct values and denote this by C_0.
2. For each distinct value, randomly choose one observation and connect them based on C_0.
3. For each distinct value with more than one observation, connect those observations by a tree.

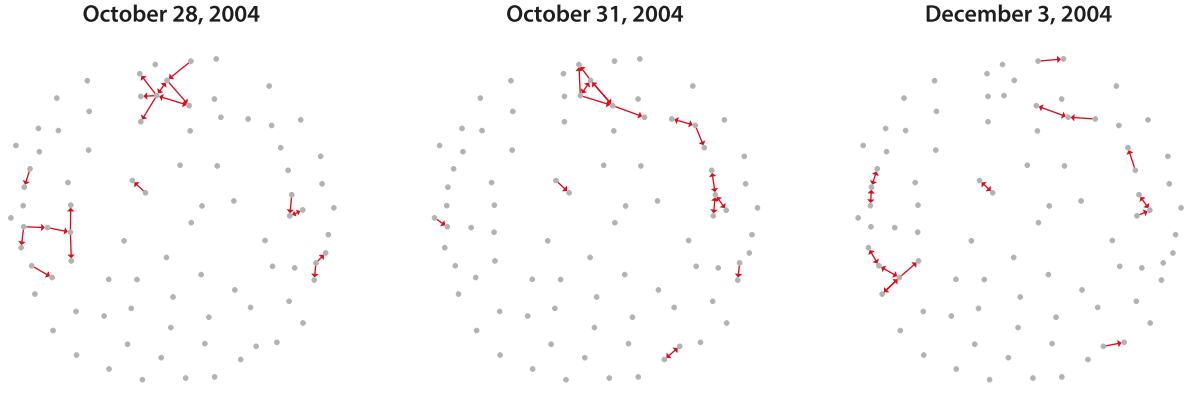

Figure 6

Phone-call networks on three representative days. Nodes indicate the participants; an edge points from one node to another if there was a phone call between the two persons, and the direction is from the caller to the callee.

Table 5 ***p*-Values and corresponding test statistics (in parentheses) for four testing procedures proposed by Chen & Zhang (2015) and Chu & Chen (2019)**

	MST #1	MST #2	MST #3
$\max_{n_0 \le t \le n_1} Z_0(t)$	0.09 (2.32)	0.91 (0.92)	0.51 (1.57)
$\max_{n_0 \le t \le n_1} S(t)$	0.04 (13.61)	0.08 (12.31)	0.01 (16.36)
$\max_{n_0 \le t \le n_1} Z_w(t)$	0.44 (2.11)	0.02 (3.49)	0.88 (1.54)
$\max_{n_0 \le t \le n_1} M(t)$	0.09 (3.05)	0.02 (3.49)	0.05 (3.27)

$\max_{n_0 \le t \le n_1}$ takes the maximum value for t in the range of (n_0, n_1). Test statistics are Z_0, original; S, generalized; Z_w, weighted; and M, max-type. Abbreviation: MST, minimum spanning tree. Table adapted from Song & Chen (2022), "Asymptotic Distribution-Free Changepoint Detection for Data with Repeated Observations," *Biometrika* 109(3), pp. 783–98, by permission of Oxford University Press.

One good choice of C_0 is the union of all MSTs on the distinct values. There are still many graphs constructed by Algorithm 1. Song & Chen (2022) studied two approaches to make use of these graphs:

1. Averaging: compute the test statistic for each graph from Algorithm 1 and take the average
2. Union: take the union of all graphs from Algorithm 1 and compute the test statistic on the union graph

Since there are millions of graphs constructed from Algorithm 1, it is infeasible to obtain all the graphs to conduct either the averaging or union approach. Song & Chen (2022) worked out analytic expressions for both approaches. Here, we use slightly different notations as there are repeated observations. Suppose that there are K distinct values. Each time, t, divides the sequence into two groups, before or at time t (group 1) and after time t (group 2). Let $n_{ik}(t)$ be the number of observations in group $i\,(i = 1, 2)$ and category $k\,(k = 1, \ldots, K)$ and $m_k\,(k = 1, \ldots, K)$ be the number of observations in category k. Notice that $m_k = n_{1k}(t) + n_{2k}(t)\,(k = 1, \ldots, K)$, $\sum_{k=1}^{K} m_k = n$, $\sum_{k=1}^{K} n_{1k}(t) = t$, and $\sum_{k=1}^{K} n_{2k}(t) = n - t$.

The new scan statistics under the averaging approach are

$$\max_{n_0 \le t \le n_1} Z_{0,(a)}(t), \quad \max_{n_0 \le t \le n_1} Z_{w,(a)}(t), \quad \max_{n_0 \le t \le n_1} S_{(a)}(t), \quad \text{and} \quad \max_{n_0 \le t \le n_1} M_{(a)}(t),$$

where

$$Z_{0,(a)}(t) = \frac{R_{0,(a)}(t) - \mathrm{E}(R_{0,(a)}(t))}{\sqrt{\mathrm{Var}(R_{0,(a)}(t))}},$$

$$Z_{w,(a)}(t) = \frac{R_{w,(a)}(t) - \mathrm{E}(R_{w,(a)}(t))}{\sqrt{\mathrm{Var}(R_{w,(a)}(t))}}, \quad R_{w,(a)}(t) = \frac{n-t-1}{n-1} R_{1,(a)}(t) + \frac{t-1}{n-2} R_{2,(a)}(t), \quad \text{and}$$

$$Z_{\mathrm{diff},(a)}(t) = \frac{R_{\mathrm{diff},(a)}(t) - \mathrm{E}\left(R_{\mathrm{diff},(a)}(t)\right)}{\sqrt{\mathrm{Var}(R_{\mathrm{diff},(a)}(t))}}, \quad R_{\mathrm{diff},(a)}(t) = R_{1,(a)}(t) - R_{2,(a)}(t),$$

with

$$R_{0,(a)}(t) = \sum_{k=1}^{K} \frac{2n_{1k}(t)n_{2k}(t)}{m_k} + \sum_{(u,v)\in C_0} \frac{n_{1u}(t)n_{2v}(t) + n_{1v}(t)n_{2u}(t)}{m_u m_v},$$

$$R_{1,(a)}(t) = \sum_{k=1}^{K} \frac{n_{1k}(t)\,(n_{1k}(t)-1)}{m_k} + \sum_{(u,v)\in C_0} \frac{n_{1u}(t)n_{1v}(t)}{m_u m_v}, \quad \text{and}$$

$$R_{2,(a)}(t) = \sum_{k=1}^{K} \frac{n_{2k}(t)\,(n_{2k}(t)-1)}{m_k} + \sum_{(u,v)\in C_0} \frac{n_{2u}(t)n_{2v}(t)}{m_u m_v}.$$

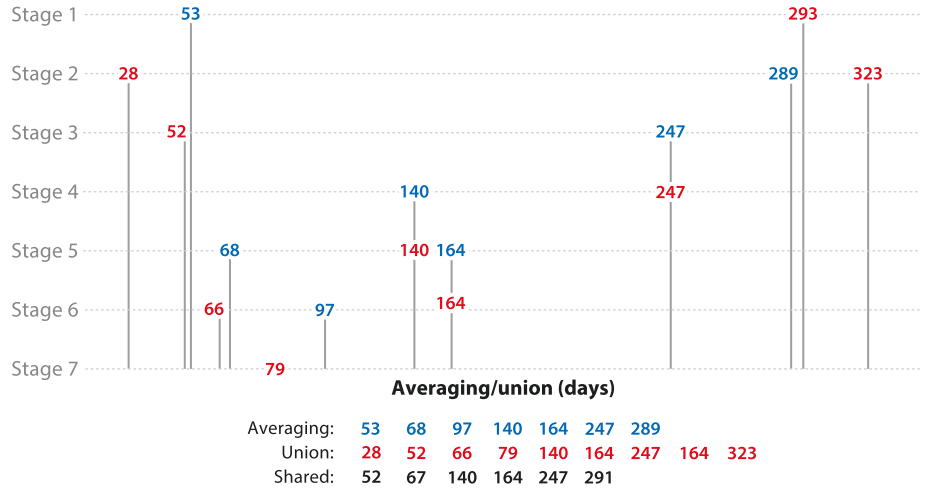

Figure 7

Estimated change-points and the order where change-points are detected in the averaging and union approaches, for the phone-call network dataset. Here, stage 1 is applying the approaches to the entire sequence, stage 2 is applying the approaches to the sub-sequences divided by the detected change-point from the previous stage, and so on. The method stops when no further change-points can be detected. Figure adapted from Song & Chen (2022), "Asymptotic Distribution-Free Changepoint Detection for Data with Repeated Observations," *Biometrika* 109(3), pp. 783–98, by permission of Oxford University Press.

The new scan statistics under the union approach are defined similarly, with

$$R_{0,(u)}(t) = \sum_{k=1}^{K} n_{1k}(t)n_{2k}(t) + \sum_{(u,v)\in C_0} \big(n_{1u}(t)n_{2v}(t) + n_{1v}(t)n_{2u}(t)\big),$$

$$R_{1,(u)}(t) = \sum_{k=1}^{K} \frac{n_{1k}(t)\,(n_{1k}(t)-1)}{2} + \sum_{(u,v)\in C_0} n_{1u}(t)n_{1v}(t), \quad \text{and}$$

$$R_{2,(u)}(t) = \sum_{k=1}^{K} \frac{n_{2k}(t)\,(n_{2k}(t)-1)}{2} + \sum_{(u,v)\in C_0} n_{2u}(t)n_{2v}(t).$$

Song & Chen (2022) further worked out analytic p-value approximations for the new scan statistics to make the new tests easy off-the-shelf tools for analyzing large datasets.

The new generalized scan statistics are applied to the phone-call network dataset, and the results are presented in **Figure 7**. A change-point $\hat{\tau}$ is defined to be detected by both approaches if they each find a change-point within the set $[\hat{\tau}-2, \hat{\tau}+2]$, and the shared change-point is computed as the floor of the average of the two change-points detected by the two approaches. Since the underlying distribution of the dataset is unknown, a sanity check is done with the distance matrix of the dataset (**Figure 8**). It is evident that there are some signals in this dataset and the results from the new tests are a reasonably good match to the signals.

5.2. A Faster Algorithm

When the graph is not given, the recommended graph to use in the graph-based change-point detection framework is the k-MST graph due to its high power and relatively low computational cost. However, when the sequence is long (e.g., hundreds of thousands of observations),

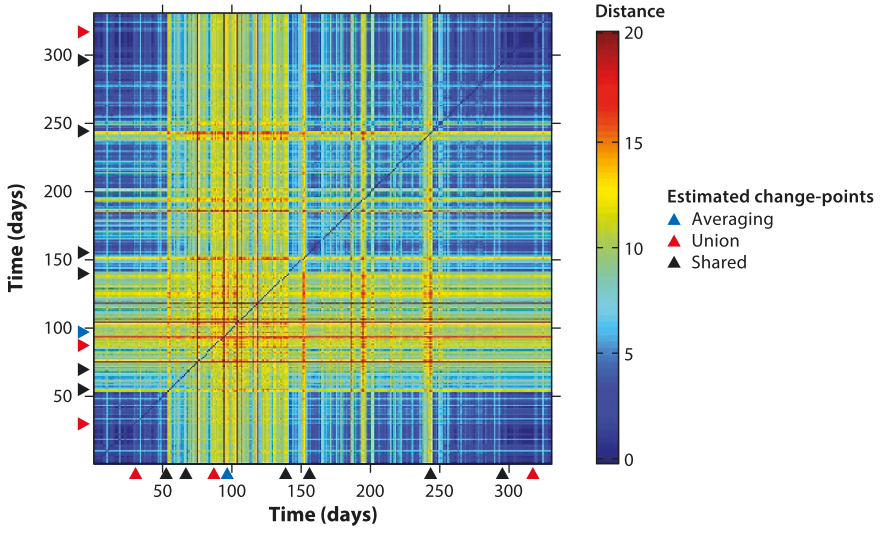

Figure 8

Heatmap of L_1 norm distance matrix corresponding to 330 networks in the phone-call network dataset. Change-points returned by the averaging and union tests, and the shared change-points, computed as the floor of the average of the two change-points detected by the two approaches, are shown. Figure adapted from Song & Chen (2022), "Asymptotic Distribution-Free Changepoint Detection for Data with Repeated Observations," *Biometrika* 109(3), pp. 783–98, by permission of Oxford University Press.

constructing the k-MST is time consuming: It requires $O(dn^2)$ time to compute the distance matrix among n d-dimensional observations, so it takes at least $O(dn^2)$ time to construct the k-MST from the original data when the pairwise distances were not provided in the beginning, which is usually the case. Recently, Liu & Chen (2022) extended the framework to directed approximate k-NN graphs. The new method on the directed approximate k-NN graph has power on par with the method on the k-MST graph, while the approximate k-NNs can be obtained in a much faster way. The time complexity of the entire method is $O(dn(\log n + k \log d) + nk^2)$, reducing time cost significantly.

The new test is applied to a functional MRI (fMRI) dataset (large $d = 96 \times 96 \times 48 = 442{,}368$ and moderate $n = 598$) (**Table 6**) and a Neuropixels dataset (moderate d, specified in **Table 7**, and large $n = 39{,}053$). The estimated change-points and runtimes are presented in **Tables 6** and **7**. We can see a clear improvement in runtime using the new approach based on the directed approximated 5-NN graph compared with the standard graph-based framework under the 5-MST.

Table 6 Results of two change-point detection methods on an fMRI dataset with two subjects

Subject	Method	$\hat{\tau}$	p-Value	Time cost (minutes)
SID-000005	New (d-a5NN)	437	<0.001	3.8
	5-MST	437	<0.001	120.9
SID-000024	New (d-a5NN)	260	<0.001	3.9
	5-MST	260	<0.001	147.8

The change-point is given by $\hat{\tau}$. Abbreviations: 5-MST, 5-minimum spanning tree; d-a5NN, directed approximated 5-nearest neighbors; fMRI, functional MRI. Table adapted with permission from Liu & Chen (2022); © 2022 IEEE.

Table 7 Results of two change-point detection methods on a Neuropixels dataset

Brain region	Method	$\hat{\tau}$	p-Value	Time cost (minutes)
Caudate putamen ($d = 176$)	New (d-a5NN)	35,148	<0.001	7.7
	5-MST	35,056	<0.001	96.1
Frontal motor ($d = 78$)	New (d-a5NN)	31,081	<0.001	6.0
	5-MST	32,242	<0.001	77.8
Hippocampus ($d = 265$)	New (d-a5NN)	4,109	<0.001	20.7
	5-MST	4,382	<0.001	159.1
Lateral septum ($d = 122$)	New (d-a5NN)	29,616	<0.001	11.4
	5-MST	29,636	<0.001	89.3
Midbrain ($d = 127$)	New (d-a5NN)	20,580	<0.001	13.9
	5-MST	20,590	<0.001	105.6
Superior colliculus ($d = 42$)	New (d-a5NN)	23,539	<0.001	4.0
	5-MST	31,328	<0.001	65.4
Somatomotor ($d = 91$)	New (d-a5NN)	30,316	<0.001	7.6
	5-MST	30,312	<0.001	81.9
Thalamus ($d = 227$)	New (d-a5NN)	28,613	<0.001	21.7
	5-MST	28,608	<0.001	146.1
V1 ($d = 334$)	New (d-a5NN)	30,226	<0.001	17.5
	5-MST	30,338	<0.001	173.8

The change-point is given by $\hat{\tau}$. Abbreviations: 5-MST, 5-minimum spanning tree; d-a5NN, directed approximated 5-nearest neighbors. Table adapted with permission from Liu & Chen (2022); © 2022 IEEE.

5.3. Finding Multiple Change-Points

It is common that a sequence has more than one change-point. The traditional approaches, binary segmentation (Vostrikova 1981) and circular binary segmentation (Olshen et al. 2004), have their drawbacks (Fryzlewicz 2014). Thus, Zhang & Chen (2021) adapt the idea of wild binary segmentation (Fryzlewicz 2014) and seeded binary segmentation (Kovács et al. 2020) to the graph-based framework to find a pool of candidate change-points. They then propose a pseudo–Bayesian information criterion for change-point selection. Simulation studies show that this approach has superb performance compared with other state-of-the-art methods in dealing with high-dimensional/non-Euclidean data.

5.4. Dealing with Locally Dependent Data

Local dependency is common in time-series data and spatial data. For example, in social networks, relationships among people last over an extended time, and in neuroimaging, neurons react over an extended time, resulting in similar images over consecutive time points. Modeling and adjusting for local dependence in change-point detection through traditional parametric approaches are not realistic in high-dimensional settings unless strong assumptions are imposed. Chen (2019a) proposed using circular block permutation to approximate the null distribution and achieved a good balance between preserving the power of the method and controlling the familywise error rate.

5.5. Other Related Work

Graph-based change-point detection methods have gained attention from a variety of fields due to their fast applicability and mild requirements on the data. For example, Dai et al. (2016) used the graph-based method to study brain functional connectivity, Pallotta et al. (2017) applied it to

analyze community evolution in network data, Shi et al. (2018) utilized the shortest Hamiltonian path graph to study the landing and departure times of bees' flower visit using video data, and Dong et al. (2020) used the graph-based change-point detection method to study gene coexpression dynamics. More recently, Nie & Nicolae (2021) and Zhou & Chen (2022) sought to add weights to the similarity graph, which could boost power when the weights reflect additional similarity information.

SUMMARY POINTS

1. The graph-based change-point detection framework adapts graph-based two-sample tests to the scan statistic setting.
2. Analytical type I error control is provided for both the offline and online settings. These are shown to work well for finite sample sizes, making the methods practical to apply for complex datasets.
3. Recent advancements have extended the framework to be useful for a broader class of settings, such as data with repeated observations or locally dependent data.
4. Graph-based change-point detection can detect a variety of types of changes, offering power and flexibility for a wide range of applications.

FUTURE ISSUES

1. Uncertainty quantification for the accuracy of the estimated change-points remains an issue. Specifically, the construction of confidence intervals with coverage guarantees in a nonparametric setting needs to be worked out. Chen & Zhang (2015) proposed a modified version of the Cox-Spjøtvoll–type confidence region (Cox & Spjøtvoll 1982), but this region is quite conservative and we aim to develop more accurate procedures.
2. The choice of the optimal graph construction and its density can affect the power and accuracy of the graph-based framework. Moreover, the choice of similarity measure is not always obvious and could be estimated in a data-driven fashion. Further exploration and studies will need to be done to better incorporate these decisions into the framework.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors were supported in part by National Science Foundation (NSF) Awards DMS-1513653 and DMS-1848579 when conducting some of the research activities covered in this manuscript.

LITERATURE CITED

Arlot S, Celisse A, Harchaoui Z. 2019. A kernel multiple change-point algorithm via model selection. *J. Mach. Learn. Res.* 20(162)

Barabâsi AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. 2002. Evolution of the social network of scientific collaborations. *Phys. A Stat. Mech. Appl.* 311(3):590–614

Basseville M, Nikiforov IV. 1993. *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ: Prentice Hall

Bolton RJ, Hand DJ. 2001. Unsupervised profiling methods for fraud detection. In *Proceedings of Credit Scoring and Credit Control VII*, pp. 235–55. Edinburgh: Univ. Edinburgh Credit Res. Cent.

Brodsky E, Darkhovsky BS. 1993. *Nonparametric Methods in Change Point Problems*. New York: Springer

Cabeza R, Nyberg L. 2000. Imaging cognition II: An empirical review of 275 PET and fMRI studies. *J. Cogn. Neurosci.* 12(1):1–47

Carlstein EG, Müller HG, Siegmund D. 1994. *Change-point problems: Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College*. N.p.: Inst. Math. Stat.

Celisse A, Marot G, Pierre-Jean M, Rigaill G. 2018. New efficient algorithms for multiple change-point detection with reproducing kernels. *Comput. Stat. Data Anal.* 128:200–20

Chan HP, Walther G. 2015. Optimal detection of multi-sample aligned sparse signals. *Ann. Stat.* 43(5):1865–95

Chandola V, Banerjee A, Kumar V. 2009. Anomaly detection: a survey. *ACM Comput. Surv.* 41(3):15

Chang WC, Li CL, Yang Y, Póczos B. 2019. Kernel change-point detection with auxiliary deep generative models. arXiv:1901.06077 [stat.ML]

Chen H. 2019a. Change-point detection for multivariate and non-Euclidean data with local dependency. arXiv:1903.01598 [stat.ME]

Chen H. 2019b. Sequential change-point detection based on nearest neighbors. *Ann. Stat.* 47(3):1381–407

Chen H, Chen X, Su Y. 2018. A weighted edge-count two-sample test for multivariate and object data. *J. Am. Stat. Assoc.* 113:1146–55

Chen H, Friedman JH. 2017. A new graph-based two-sample test for multivariate and object data. *J. Am. Stat. Assoc.* 112(517):397–409

Chen H, Ren H, Yao F, Zou C. 2021. Data-driven selection of the number of change-points via error rate control. *J. Am. Stat. Assoc.* **https://doi.org/10.1080/01621459.2021.1999820**

Chen H, Zhang N. 2015. Graph-based change-point detection. *Ann. Stat.* 43(1):139–76

Chen H, Zhang NR. 2013. Graph-based tests for two-sample comparisons of categorical data. *Stat. Sin.* 23:1479–503

Chen J, Gupta AK. 2000. *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. New York: Springer

Chen LHY, Shao QM. 2005. Stein's method for normal approximation. In *An Introduction to Stein's Method*, ed. AD Barbour, LHY Chen, pp. 1–59. Singapore: World Sci.

Chu L, Chen H. 2019. Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Ann. Stat.* 47(1):382–414

Chu L, Chen H. 2022. Sequential change-point detection for high-dimensional and non-Euclidean data. arXiv:1810.05973 [stat.ME]

Collins RT, Lipton A, Kanade T, Fujiyoshi H, Duggins D, et al. 2000. *A system for video surveillance and monitoring*. Work. Pap. CMU-RI-TR-00-12, Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA

Cox D, Spjøtvoll E. 1982. On partitioning means into groups. *Scand. J. Stat.* 9(3):147–52

Csörgö M, Horváth L. 1997. *Limit Theorems in Change-Point Analysis*. New York: Wiley

Dai M, Zhang Z, Srivastava A. 2016. Testing stationarity of brain functional connectivity using change-point detection in fMRI data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 19–27. New York: IEEE

Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, et al. 2020. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* 369:6500

Desobry F, Davy M, Doncarli C. 2005. An online kernel change detection algorithm. *IEEE Trans. Sign. Proc.* 53(8):2961–74

Dong F, He Y, Wang T, Han D, Lu H, Zhao H. 2020. Predicting viral exposure response from modeling the changes of co-expression networks using time series gene expression data. *BMC Bioinformatics* 21:370

Eagle N, Pentland A, Lazer D. 2009. Inferring friendship network structure by using mobile phone data. *PNAS* 106(36):15274–78

Frick K, Munk A, Sieling H. 2014. Multiscale change point inference. *J. R. Stat. Soc. Ser. B* 76(3):495–580
Friedman J, Rafsky L. 1979. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Stat.* 7(4):697–717
Fryzlewicz P. 2014. Wild binary segmentation for multiple change-point detection. *Ann. Stat.* 42(6):2243–81
Fryzlewicz P. 2020. Narrowest significance pursuit: inference for multiple change-points in linear models. arXiv:2009.05431 [stat.ME]
Garreau D, Arlot S. 2018. Consistent change-point detection with kernels. *Electron. J. Stat.* 12(2):4440–86
Guia J, Wittlin C. 1999. *Nine Problem Areas Concerning Tirant lo Blanc*. London: Tamesis
Guimarães SJF, Couprie M, de Albuquerque Araújo A, Leite NJ. 2003. Video segmentation based on 2D image analysis. *Pattern Recognit. Lett.* 24(7):947–57
Harchaoui Z, Cappé O. 2007. Retrospective multiple change-point estimation with kernels. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pp. 768–72. New York: IEEE
Harchaoui Z, Moulines E, Bach FR. 2009. Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, ed. D Koller, D Schuurmans, Y Bengio, L Bottou, pp. 609–16. Red Hook, NY: Curran
Henze N. 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Stat.* 16(2):772–83
Hu X, Wang Y, Wu Q. 2014. Multiple authors detection: a quantitative analysis of *Dream of the Red Chamber*. *Adv. Adapt. Data Anal.* 6(04):1450012
Huang S, Ernberg I, Kauffman S. 2009. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* 20:869–76
James B, James K, Siegmund D. 1992. Asymptotic approximations for likelihood ratio tests and confidence regions for a change-point in the mean of a multivariate Gaussian process. *Stat. Sin.* 2(1):69–90
Jun JJ, Steinmetz NA, Siegle JH, Denman DJ, Bauza M, et al. 2017. Fully integrated silicon probes for high-density recording of neural activity. *Nature* 551(7679):232–36
Kappenman J. 2012. A perfect storm of planetary proportions. *IEEE Spectrum* 49(2):26–31
Killick R, Fearnhead P, Eckley IA. 2012. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* 107(500):1590–98
Koprinska I, Carrato S. 2001. Temporal video segmentation: a survey. *Signal Proc. Image Commun.* 16(5):477–500
Kossinets G, Watts D. 2006. Empirical analysis of an evolving social network. *Science* 311(5757):88–90
Kovács S, Li H, Bühlmann P, Munk A. 2020. Seeded binary segmentation: a general methodology for fast and optimal change point detection. arXiv:2002.06633 [stat.ME]
Li ZN, Zhong X, Drew MS. 2002. Spatial–temporal joint probability images for video segmentation. *Pattern Recognit.* 35(9):1847–67
Liu YW, Chen H. 2022. A fast and efficient change-point detection framework based on approximate *k*-nearest neighbor graphs. *IEEE Trans. Signal Proc.* 70:1976–86
Londschien M, Bühlmann P, Kovács S. 2022. Random forests for change point detection. arXiv:2205.04997 [stat.ME]
Long DG, Drinkwater MR, Holt B, Saatchi S, Bertoia C. 2001. Global ice and land climate studies using scatterometer image data. *Eos* 82(43):503
Lung-Yut-Fong A, Lévy-Leduc C, Cappé O. 2011. Homogeneity and change-point detection tests for multivariate data using rank statistics. arXiv:1107.1971 [math.ST]
Malladi R, Kalamangalam GP, Aazhang B. 2013. Online Bayesian change point detection algorithms for segmentation of epileptic activity. In *2013 Asilomar Conference on Signals, Systems and Computers*, pp. 1833–37. New York: IEEE
Matteson DS, James NA. 2014. A nonparametric approach for multiple change point analysis of multivariate data. *J. Am. Stat. Assoc.* 109(505):334–45
Nie L, Nicolae DL. 2021. Weighted-graph-based change point detection. arXiv:2103.02680 [stat.ME]
Niu YS, Zhang H. 2012. The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* 6(3):1306

Olshen AB, Venkatraman E, Lucito R, Wigler M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4):557–72

Pallotta G, Konjevod G, Cadena J, Nguyen P. 2017. Context-aided analysis of community evolution in networks. *Stat. Anal. Data Min.* 10(5):290–311

Pastor-Satorras R, Smith E, Solé RV. 2003. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* 222(2):199–210

Pervaiz F, Pervaiz M, Rehman NA, Saif U. 2012. FluBreaks: early epidemic detection from Google flu trends. *J. Med. Internet Res.* 14(5):e125

Qu M, Shih FY, Jing J, Wang H. 2005. Automatic solar filament detection using image processing techniques. *Solar Phys.* 228(1–2):119–35

Radke RJ, Andra S, Al-Kofahi O, Roysam B. 2005. Image change detection algorithms: a systematic survey. *IEEE Trans. Image Proc.* 14(3):294–307

Riba A, Ginebra J. 2006. Diversity of vocabulary and homogeneity of literary style. *J. Appl. Stat.* 33(7):729–41

Rosenbaum PR. 2005. An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B* 67(4):515–30

Shi X, Wu Y, Rao CR. 2018. Consistent and powerful non-Euclidean graph-based change-point test with applications to segmenting random interfered video data. *PNAS* 115(23):5914–19

Siegmund D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer

Siegmund D, Yakir B. 2007. *The Statistics of Gene Mapping*, Vol. 1. New York: Springer

Siegmund D, Yakir B, Zhang N. 2011. Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.* 5(2A):645–68

Song H, Chen H. 2022. Asymptotic distribution-free change-point detection for data with repeated observations. *Biometrika* 109:783–98

Srivastava M, Worsley K. 1986. Likelihood ratio tests for a change in the multivariate normal mean. *J. Am. Stat. Assoc.* 81:199–204

Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD. 2019. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* 364(6437):255

Tartakovsky AG, Rozovskii BL, Blazek RB, Kim H. 2006. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. *IEEE Trans. Signal Proc.* 54(9):3372–82

Tian YL, Lu M, Hampapur A. 2005. Robust and efficient foreground analysis for real-time video surveillance. In *Computer Vision and Pattern Recognition, 2005*, Vol. 1, pp. 1182–87. New York: IEEE

Tsirigos A, Rigoutsos I. 2005. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.* 33(3):922–33

Vostrikova LY. 1981. Detecting "disorder" in multidimensional random processes. *Doklady Akad. Nauk.* 259:270–74

Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18(7):1283–92

Wang H, Tang M, Park Y, Priebe CE. 2014. Locality statistics for anomaly detection in time series of graphs. *IEEE Trans. Signal Proc.* 62(3):703–17

Wang T, Samworth RJ. 2018. High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B* 80(1):57–83

Wong WK, Moore AW, Cooper GF, Wagner MM. 2003. Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, ed. T Fawcett, N Mishra, pp. 808–15. Menlo Park, CA: AAAI

Xie M, Han S, Tian B, Parvin S. 2011. Anomaly detection in wireless sensor networks: a survey. *J. Netw. Comput. Appl.* 34(4):1302–25

Xie Y, Siegmund D. 2013. Sequential multi-sensor change-point detection. In *Information Theory and Applications Workshop (ITA), 2013*, pp. 448–67. New York: IEEE

Zhang J, Chen H. 2022. Graph-based two-sample tests for data with repeated observations. *Stat. Sin.* 32:391–415

Zhang M, Raghunathan A, Jha NK. 2013. MedMon: securing medical devices through wireless monitoring and anomaly detection. *IEEE Trans. Biomed. Circ. Syst.* 7(6):871–81

Zhang N, Siegmund D, Ji H, Li J. 2010. Detecting simultaneous changepoints in multiple sequences. *Biometrika* 97(3):631–45

Zhang Y, Chen H. 2021. Graph-based multiple change-point detection. arXiv:2110.01170 [stat.ME]

Zhou D, Chen H. 2022. RING-CPD: asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. arXiv:2206.03038 [stat.ME]

Zou C, Wang G, Li R. 2020. Consistent selection of the number of change-points via sample-splitting. *Ann. Stat.* 48(1):413–39

Annual Review of Statistics and Its Application

Volume 10, 2023

Contents

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics