The Annals of Statistics
2019, Vol. 47, No. 1, 382–414
https://doi.org/10.1214/18-AOS1691

ASYMPTOTIC DISTRIBUTION-FREE CHANGE-POINT DETECTION FOR MULTIVARIATE AND NON-EUCLIDEAN DATA[1]

BY LYNNA CHU AND HAO CHEN

University of California, Davis

We consider the testing and estimation of change-points, locations where the distribution abruptly changes, in a sequence of multivariate or non-Euclidean observations. We study a nonparametric framework that utilizes similarity information among observations, which can be applied to various data types as long as an informative similarity measure on the sample space can be defined. The existing approach along this line has low power and/or biased estimates for change-points under some common scenarios. We address these problems by considering new tests based on similarity information. Simulation studies show that the new approaches exhibit substantial improvements in detecting and estimating change-points. In addition, under some mild conditions, the new test statistics are asymptotically distribution-free under the null hypothesis of no change. Analytic p-value approximations to the significance of the new test statistics for the single change-point alternative and changed interval alternative are derived, making the new approaches easy off-the-shelf tools for large datasets. The new approaches are illustrated in an analysis of New York taxi data.

1. Introduction. Change-point analysis is regaining attention as we enter the big data era. Massive amounts of data are collected in many fields 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, which divides the sequence into homogeneous temporal or spatial segments. In this paper, we consider this segmentation problem. Let the sequence of observations be $\{\mathbf{y}_i : i = 1, \ldots, n\}$, indexed by time or some other meaningful orderings. We are concerned with testing the null hypothesis:

$$H_0 : \mathbf{y}_i \sim F_0, \qquad i = 1, \ldots, n \tag{1.1}$$

against the single change-point alternative

$$H_1 : \exists 1 \le \tau < n, \qquad \mathbf{y}_i \sim \begin{cases} F_0, & i \le \tau, \\ F_1, & \text{otherwise} \end{cases} \tag{1.2}$$

Received June 2017; revised February 2018.
[1]Supported in part by NSF award DMS-15-13653.
MSC2010 subject classifications. Primary 62G32; secondary 60K35.
Key words and phrases. Change-point, graph-based tests, nonparametric, scan statistic, tail probability, high-dimensional data, network data, non-Euclidean data.

or the changed interval alternative

$$H_2 : \exists 1 \le \tau_1 < \tau_2 < n, \qquad \mathbf{y}_i \sim \begin{cases} F_0, & i = \tau_1 + 1, \ldots, \tau_2, \\ F_1, & \text{otherwise}, \end{cases} \tag{1.3}$$

where F_0 and F_1 are two different probability measures. We consider the problem that observations are independent over time. (More discussions on violation of this independence assumption can be found in Supplement I [Chu and Chen (2019)].)

The segmentation problem has been widely studied for *univariate* data. See monograph Carlstein, Müller and Siegmund (1994) for a survey. However, in many modern applications, $\{\mathbf{y}_i\}$'s could be a sequence of vectors [e.g., cross-sample copy number variation analysis, Zhang et al. (2010)], images [e.g., brain image, Park et al. (2015)], or networks [e.g., social network, Kossinets and Watts (2006)].

When $\mathbf{y}_i \in \mathbb{R}^d$ and the d dimensions are independent, the problem becomes the analysis of d independent sequences and it has been studied in a number of works; see, for example, Zhang et al. (2010) and Siegmund (2013). For more generic multivariate observations, most existing methods are based on parametric models [see, e.g., Chen and Gupta (2012), Csörgő and Horváth (1997) and references therein]. Parametric methods have been proposed for network data sequences as well. For example, Heard et al. (2010) designed a two-stage Bayesian method to detect anomalies by modeling the communication between nodes over time as a counting process where increments of the process follow a Bayesian probability model. Wang et al. (2014) designed locality-based scan statistics to detect change arising in the connectivity matrix of networks generated by a stochastic block model where the block membership of the vertices are fixed across time. All of these parametric methods provide useful tools when the assumptions made in the paper are reasonably true. However, these assumptions are many times too strong in real applications.

Nonparametric methods have been proposed for the change-point detection problem for multivariate/non-Euclidean observations as well [Jirak (2015), Matteson and James (2014), Lung-Yut-Fong, Lévy-Leduc and Cappé (2015), Cule, Samworth and Stewart (2010), Desobry, Davy and Doncarli (2005)]. Nonparametric methods are usually more flexible in terms of model specification. However, it is in general more difficult to conduct theoretical analysis, such as controlling the type I error.

Recently, Chen and Zhang (2015) proposed a nonparametric approach that can be applied to data in arbitrary dimension and to non-Euclidean data. They also provided *analytical p-value approximations* for type I error control, making their approach easy to be applied to large data sets. Through simulation studies, they showed that their approach achieves substantial power gains when dimension is moderate to high compared with existing parametric change-point methods.

However, while the method proposed by Chen and Zhang (2015) is effective for locational alternatives, it is less effective for scale alternatives and even worse

provides biased estimates for the location of the change-point when detected. Also, if the change-point is not in the middle of the sequence, the detection power could be low (more details of these problems are discussed in Section 2).

In this paper, we improve upon the limitations of the test statistic in Chen and Zhang (2015) and propose three new test statistics. The new test statistics exhibit better estimates to the location of the change-points for a wider range of alternatives and also exhibit substantial power gains when the change is not in the middle of the sequence. In addition, under some mild regularity conditions, the new statistics are asymptotically distribution-free under the null hypothesis of no change. The new approaches are implemented in an `R` package `gSeg`.

The organization of the rest of the paper is as follows. In Section 2, we describe and explain in more details the problems of the method in Chen and Zhang (2015). To tackle the problems, three new scan statistics are proposed in Section 3. The asymptotic behaviors of the new test statistics are studied and analytical p-value approximations for the tests are provided in Section 4. Section 5 examines the performance of the new test statistics under more simulation settings. The new methods are illustrated in the analysis of New York taxi data in Section 6. We conclude with discussion in Section 7.

2. Restrictions of the method in Chen and Zhang (2015). In this section, we state the restrictions of the method in Chen and Zhang (2015) and explore the underlying reasons for these restrictions.

2.1. *Scenarios when the method breaks down.* The method in Chen and Zhang (2015) for detecting change-point is a typical scan statistic $\max_t Z(t)$, with $Z(t)$ a standardized two-sample test statistic for comparing $\{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ and $\{\mathbf{y}_{t+1}, \ldots, \mathbf{y}_n\}$. Ideally, when the method works, $Z(t)$ would be maximized around the true change-point. Figure 1 plots $Z(t)$ from typical simulation runs under three different scenarios: (a) a mean change, (b) a change in both mean and variance with the variance larger after the change and (c) a change in both mean and variance with the variance smaller after the change. In each scenario, the change occurs at the center of the sequence, indicated by a blue dashed vertical line in each plot. The estimated change-point is indicated by a black solid vertical line in each plot. From the plots, the method works perfectly well in scenario (a). However, it has serious problems in correctly estimating the location of the change-point in scenarios (b) and (c). From the plots, we see that the estimated change-point is biased toward the direction with a larger variance.

In addition, even when the change is only in mean, the method also has biased change-point estimates along with power loss when the change is not near the middle of the sequence. Table 1 shows the performance of the method in Chen and Zhang (2015) under two choices of the location of the change-point (middle versus one-third of the sequence). It lists the number of trials, out of 100, that the null hypothesis of homogeneity is rejected at the 0.05 level with the number in the

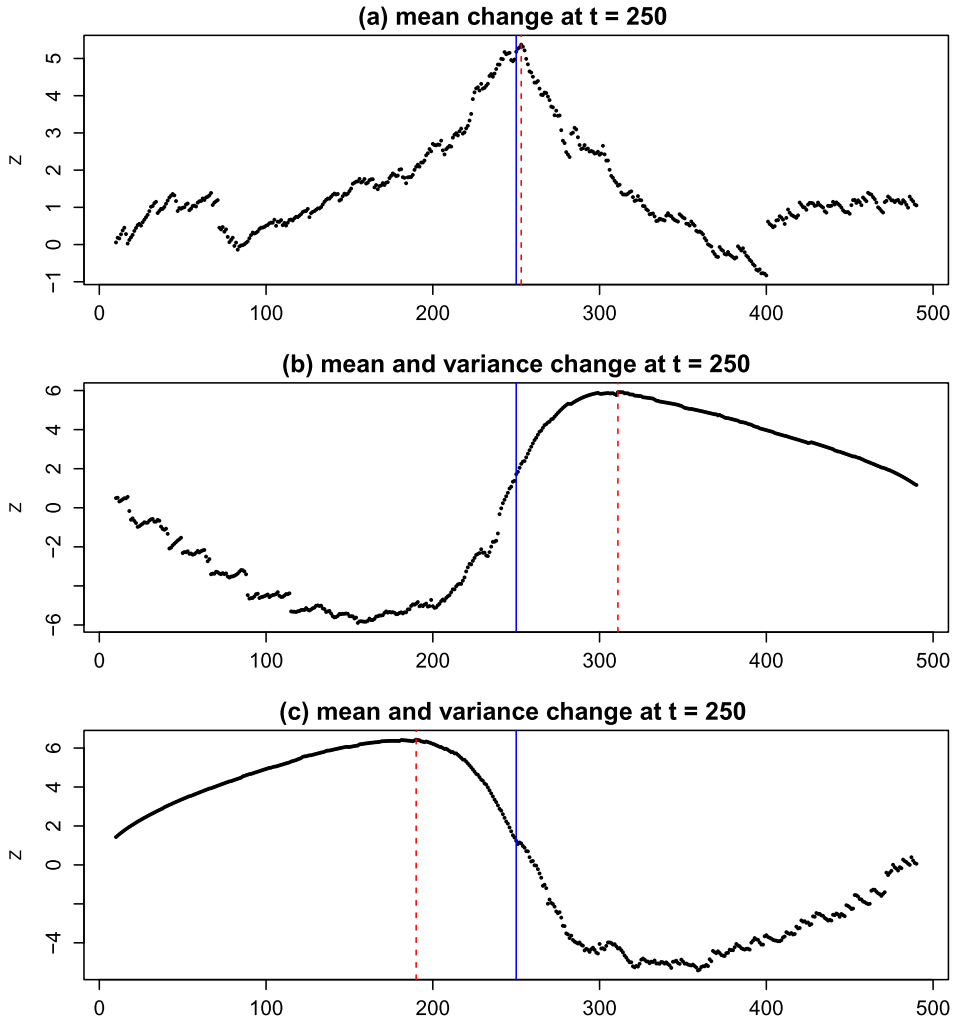


FIG. 1. *Plots of the scan statistic for the method in Chen and Zhang (2015). Multivariate Gaussian data, $d = 100$, $n = 500$. Before the change, the data is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I_d})$ and after the change, data drawn from* (a) $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$ *where* $\|\mu\|_2 = 1.4$, (b) $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I_d})$ *where* $\|\mu\|_2 = 1.4$, $\sigma = 1.2$ *and* (c) $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$ *where* $\|\mu\|_2 = 1.4$, $\sigma = 0.8$. *The change occurs at* $t = 250$ *for all scenarios. The solid vertical line indicates the true change-point. The dashed vertical line indicates the estimated change–point by the method in Chen and Zhang (2015). The similarity graphs are the 5-MST constructed using the Euclidean distance.*

parentheses those trials both rejecting the null and estimating the location of the change-point reasonably well (within 20 indices from the true change-point). In both settings, the change happens at $\tau = 250$ and the change is in the mean only [$\mathcal{N}(\mathbf{0}, \mathbf{I_d})$ versus $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$ where $\|\mu\|_2 = 1.4$ and $d = 100$].

TABLE 1

The number of trials, out of 100, *that the null hypothesis is rejected at* 0.05 *significance level with the number in the parentheses the number of trials that the null hypothesis is rejected and the index difference between the estimated change-point and true change-point less than* 20. *The change happens at* $\tau = 250$. *The length of the sequence is* n. *Before the change, the observations are drawn from* $\mathcal{N}(\mathbf{0}, \mathbf{I_d})$, $d = 100$; *after the change, the observations are drawn from* $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I_d})$, $\|\mu\|_2 = 1.4$

$n = 500$	$n = 750$
97	83
(89)	(24)

When the length of the sequence is $n = 500$, the change happens at the middle of the sequence, and the method does very well. When $n = 750$, since the change happens at $\tau = 250$, there are twice as many observations after the change compared to $n = 500$. Intuitively, the increase in sample size should increase the power of the test. However, different from what we would expect, the performance of the test becomes worse ($97 \to 83$). Even worse is the dramatic decrease in the number in the parentheses ($89 \to 24$), indicating the poor ability of the the method in estimating the location of the change-point correctly when the change does not happen in the middle of the sequence.

2.2. *Understanding the graph-based approach.* Here, we look closer at the method in Chen and Zhang (2015). It is a scan statistic $Z(t)$ calculated based on a graph-based two-sample test. First, a similarity graph G is constructed on the observations based on a distance measure defined on the sample space up to a criterion. For example, G could be a minimum spanning tree (MST), which is a tree connecting all observations such that the sum of the distances of edges in the tree is minimized; G could also be a nearest neighbor graph (NNG) where each observation connects to its nearest neighbors. Then the number of edges in G that connect observations before t and observations after t are counted. A relative low count indicates the observations before and after t are less mixed, which implies distributional difference. This graph-based two-sample test was first proposed by Friedman and Rafsky (1979) and the intuition behind this is that if the distributions of the two samples are different, observations would tend to be closer to those from the same sample. Thus, edges in the similarity graph would be more likely to connect observations within the same sample. Chen and Zhang (2015) adapts this graph-based two-sample test to the change-point setting and $Z(t)$ is a standardized version of the raw count by the mean and standard deviation of the raw count (with a sign flip so that large $Z(t)$ values imply change-points). We refer to this underlying graph-based two-sample test as the *edge-count two-sample test* for easy reference.

The rationale of the edge-count two-sample test holds for low-dimensional data. However, when the dimension is high, the edge-count two-sample test can

be powerless for some very common types of alternatives due to the curse-of-dimensionality [Chen and Friedman (2017)]. For example, if two distributions differ in variance and when the dimension is moderate to high, such as $d = 50$, the two samples would be separated into two layers with the sample with a smaller variance in the inner layer and the other sample in the outer layer. Since the volume of a d-dimensional space increases exponentially in d, the phenomenon that points in the outer layer find themselves to be closer to points in the inner layer than other points in the outer layer is common unless the number of points in the outer layer is extremely large (exponential in d). Then, for typical sample sizes, the between-sample edge-count is still high under this alternative and the edge-count two-sample test is unable to reject the null hypothesis. To address this issue, Chen and Friedman (2017) proposed a *generalized edge-count two-sample test*.

Meanwhile, Chen, Dou and Qiao (2014) found that, starting from the equal sample size scenario, the estimated power of the edge-count two-sample test decreased when one sample size was doubled and the other kept the same. As seen in Table 1, even for locational alternatives, this is counterintuitive since increasing the sample size adds more information, which should increase the power of the test. They found that the decrease in power is due to a variance boosting problem when the sample sizes are unequal. To address this issue, Chen, Dou and Qiao (2014) proposed a *weighted edge-count two-sample test*.

In the following, we adapt these two extended graph-based two-sample tests, *generalized edge-count two-sample test* and the *weighted edge-count two-sample test*, as well as a new version of the edge-count two-sample test, which we refer to as the *max-type edge-count two-sample test*, to the change-point setting.

3. New test statistics. The new test statistics for testing the null H_0 (1.1) versus the single change-point alternative H_1 (1.2) and versus the changed interval alternative H_2 (1.3) are presented below. Under the null hypothesis H_0 (1.1), the joint distribution of the observations in the sequence is the same if we permute the order of the observations. In the following, we work under the permutation null distribution that places $1/n!$ probability on each of the $n!$ permutations of $\{\mathbf{y}_i : i = 1, \ldots, n\}$. With no further specification, we use $\mathbf{P}, \mathbf{E}$, **Var**, and **Cov** to denote probability, expectation, variance, and covariance, respectively, under the permutation null distribution.

3.1. *Generalized edge-count scan statistic for single change-point alternative.* Here, we define the test statistic for the generalized edge-count two-sample test when testing the null H_0 (1.1) versus the single change-point alternative H_1 (1.2).

Each possible value of τ divides the sequence of observations into two groups: Observations that come before or at τ and observations that come after τ. Let G be the similarity graph on $\mathbf{y}_i$. We use G to denote both the graph and its set of edges when its vertex set is implicitly obvious. For more discussions on the choice of G, see Chen and Zhang (2015). For any event x let I_x be the indicator function that

takes value 1 if x is true and 0 otherwise. We define $g_i(t)$ as an indicator function for the event that $\mathbf{y}_i$ is observed after t, $g_i(t) = I_{i>t}$. For an edge $e = (i, j)$, we define

$$J_e(t) = \begin{cases} 0 & \text{if } g_i(t) \neq g_j(t), \\ 1 & \text{if } g_i(t) = g_j(t) = 0, \\ 2 & \text{if } g_i(t) = g_j(t) = 1. \end{cases}$$

For any candidate value t of τ, we define

$$R_k(t) = \sum_{e \in G} I_{J_e(t)=k}, \qquad k = 0, 1, 2. \tag{3.1}$$

Then $R_0(t)$ is the number of edges connecting observations before and after t (which is the test statistic for the edge-count two-sample test), $R_1(t)$ is the number of edges connecting observations prior to t and $R_2(t)$ is the number of edges that connect observations after t.

The generalized edge-count two-sample test at t is defined as

$$S(t) = \begin{pmatrix} R_1(t) - \mathbf{E}\big(R_1(t)\big) \\ R_2(t) - \mathbf{E}\big(R_2(t)\big) \end{pmatrix}^T \boldsymbol{\Sigma}^{-1}(t) \begin{pmatrix} R_1(t) - \mathbf{E}\big(R_1(t)\big) \\ R_2(t) - \mathbf{E}\big(R_2(t)\big) \end{pmatrix}. \tag{3.2}$$

Here, $\boldsymbol{\Sigma}(t)$ is the covariance matrix of the vector $(R_1(t), R_2(t))^T$ under the permutation null distribution. The test statistic is defined in this way so that either direction of deviations of the number of within-group edges from its null expectation would contribute to the test statistic. Under location alternatives, we would expect both $R_1(t)$ and $R_2(t)$ to be larger than their null expectations, which would lead to a large $S(t)$. Under scale alternatives, the group with the smaller variance would have a within-edge count larger than its null expectation and the group with the larger variance would have a within-edge count smaller than its null expectation, which would also lead to a large $S(t)$. Therefore, this test is powerful for both location and scale alternatives.

Figure 2 illustrates the computation of $R_1(t)$ and $R_2(t)$ on a small artificial dataset of length $n = 40$. The first 20 observations are generated from $\mathcal{N}(0, I_2)$. The second 20 observations are generated from $\mathcal{N}((2, 2)^T, I_2)$ (the 2-dimensional data is chosen for illustration purposes, while the method is not limited by dimensionality). The similarity graph G is the MST on Euclidean distance. Notice that the graph G is determined by the values of $\mathbf{y}_i$'s and not the order of their appearance. Thus it remains constant under permutation. As t changes, the group identify of some points changes.

Under the permutation null, the analytic expressions for $\mathbf{E}(R_1(t))$, $\mathbf{E}(R_2(t))$ and $\boldsymbol{\Sigma}(t) = (\Sigma_{i,j}(t))_{i,j=1,2}$ can be calculated through combinatorial analysis, and they can be obtained straightforwardly following Chen and Friedman (2017). Their expressions are listed below. Let G_i be the subgraph of G containing all edges that

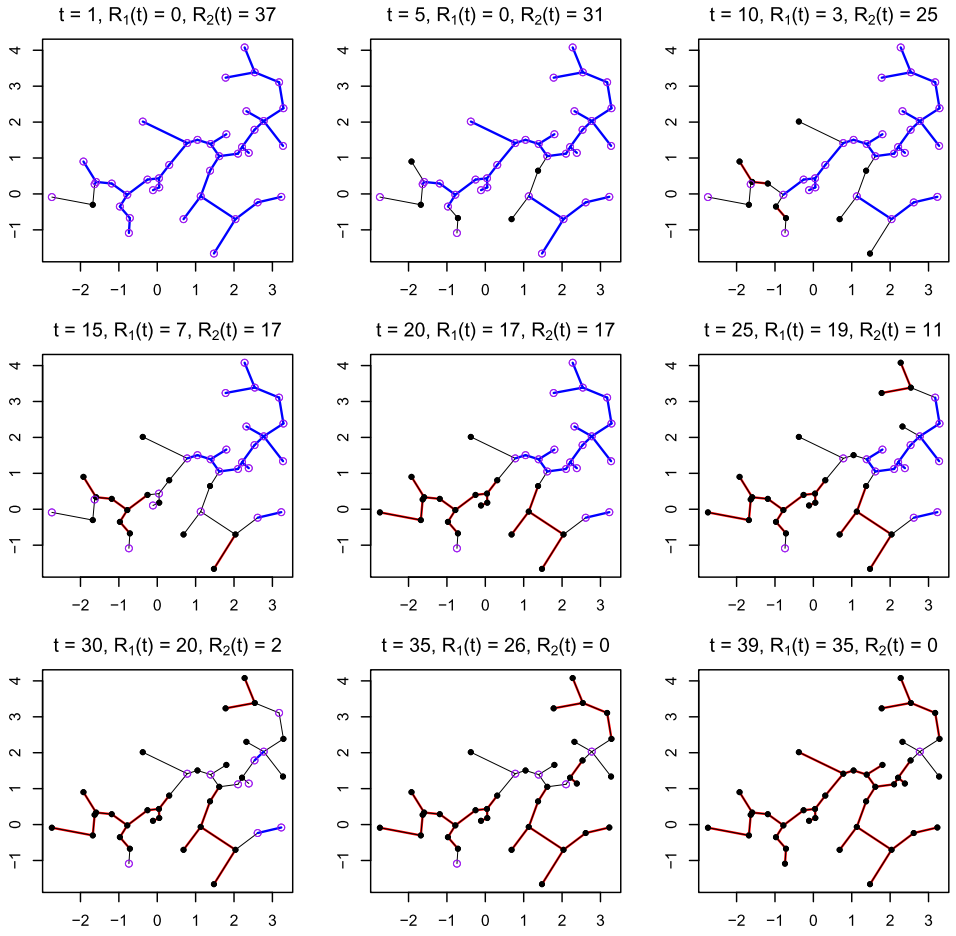


FIG. 2. *The computation of $R_1(t)$ and $R_2(t)$ for nine different values of t. The first* 20 *observations are generated from $\mathcal{N}(0, I_2)$. The second* 20 *observations are generated from $\mathcal{N}((2,2)^T, I_2)$. The similarity graph G shown here is the MST on Euclidean distance. Each t divides the observations into two groups: one group for observations before t (shown as solid circles) and the other group for observations shown after t (shown as open circles). Edges in red connect observations before t and the number of these edges is $R_1(t)$. Edges in blue connect observations after t and the number of these edges is $R_2(t)$. Notice that as t changes, the group identities change but the graph G does not change.*

connect to node $\mathbf{y}_i$. Then $|G_i|$ is the number of edges in G_i or the degree of node $\mathbf{y}_i$ in G. We have

$$\mathbf{E}\big(R_1(t)\big) = |G|\frac{t(t-1)}{n(n-1)},$$

$$\mathbf{E}\big(R_2(t)\big) = |G|\frac{(n-t)(n-t-1)}{n(n-1)},$$

$$
\begin{aligned}
\Sigma_{11}(t) &= \mathbf{E}\big(R_1(t)\big)\big(1-\mathbf{E}\big(R_1(t)\big)\big) \\
&\quad + \frac{t(t-1)(t-2)(\sum_{i=1}^{n}|G_i|^2-2|G|)}{n(n-1)(n-2)} \\
&\quad + \frac{t(t-1)(t-2)(t-3)(|G|^2-\sum_{i=1}^{n}|G_i|^2+|G|)}{n(n-1)(n-2)(n-3)}, \\
\Sigma_{22}(t) &= \mathbf{E}\big(R_2(t)\big)\big(1-\mathbf{E}\big(R_2(t)\big)\big) \\
&\quad + \frac{(n-t)(n-t-1)(n-t-2)(\sum_{i=1}^{n}|G_i|^2-2|G|)}{n(n-1)(n-2)} \\
&\quad + \Bigg[(n-t)(n-t-1)(n-t-2)(n-t-3) \\
&\quad \times \Bigg(|G|^2-\sum_{i=1}^{n}|G_i|^2+|G|\Bigg)\Bigg] \\
&\quad \times \big[n(n-1)(n-2)(n-3)\big]^{-1} \\
\Sigma_{12}(t) &= \Sigma_{21}(t) \\
&= \frac{t(t-1)(n-t)(n-t-1)(|G|^2-\sum_{i=1}^{n}|G_i|^2+|G|)}{n(n-1)(n-2)(n-3)} \\
&\quad - \mathbf{E}\big(R_1(t)\big)\mathbf{E}\big(R_2(t)\big).
\end{aligned}
$$

To test H_0 versus H_1, we use the following scan statistic:

$$\max_{n_0\le t\le n_1} S(t), \tag{3.3}$$

where n_0 and n_1 are pre-specified constraints for the range of τ, such as $n_0 = 20$, $n_1 = n - n_0$, as we need some observations in each group to 'represent' the distribution. The null hypothesis is rejected if the maxima is greater than a threshold. Details about how to choose the threshold to control the type I error rate are discussed in Section 4.

Figure 3 shows the $S(t)$ process for the dataset in Figure 2 where there is a change-point in the middle (left) and by contrast a typical result when there is no change (right). It is clear that the $\max_{n_0\le t\le n_1} S(t)$ in the left panel is much larger.

3.2. *Weighted edge-count scan statistic for single change-point alternative.* Here, we present the weighted edge-count two-sample test statistic for testing the null H_0 (1.1) versus the single change-point alternative H_1 (1.2). Following the same notation in Section 3.1, for any candidate value t of τ, the weighted edge-count two-sample test statistic is

$$R_w(t) = q(t)\sum_{e\in G} I_{J_e(t)=1} + p(t)\sum_{e\in G} I_{J_e(t)=2} = q(t)R_1(t) + p(t)R_2(t),$$

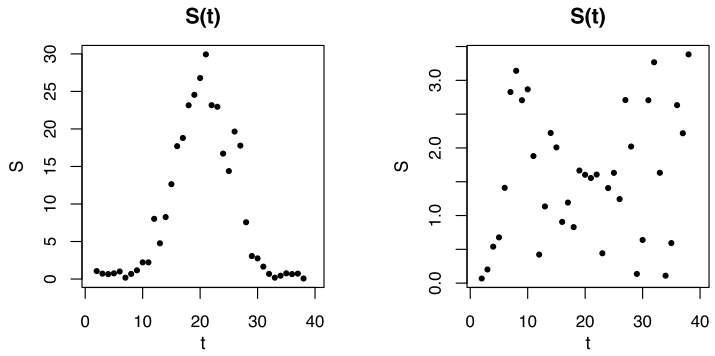


FIG. 3. *On the left, the profile of $S(t)$ against t for the example data set in Figure 2. On the right, the profile of $S(t)$ against t on a sequence of points with no change-point in a typical simulation run.* (*The scale on the y-axis are different in the two plots.*)

where $p(t) = \frac{t-1}{n-2}$ and $q(t) = 1 - p(t)$. As it is more difficult for the sample with a smaller sample size to form an edge within the sample, $R_1(t)$ and $R_2(t)$ are weighted by the inverse of their corresponding sample sizes. The test statistic defined in this way resolves the variance boosting problem [Chen, Dou and Qiao (2014)]. Relatively large values of $R_w(t)$ are evidence against the null hypothesis.

Since the null distribution of $R_w(t)$ depends on t, $R_w(t)$ is standardized so that it is comparable across t. Let

$$Z_w(t) = \frac{R_w(t) - \mathbf{E}[R_w(t)]}{\sqrt{\mathbf{Var}[R_w(t)]}}. \tag{3.4}$$

Analytic formulas for $\mathbf{E}(R_w(t))$ and $\mathbf{Var}(R_w(t))$ are given below:

$$\begin{aligned}\mathbf{E}\big(R_w(t)\big) &= |G|\frac{(t-1)(n-t-1)}{(n-1)(n-2)},\\ \mathbf{Var}\big(R_w(t)\big) &= \frac{t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)}\\ &\quad\times\left(|G| - \frac{\sum_{i=1}^{n}|G_i|^2}{(n-2)} + \frac{2|G|^2}{(n-1)(n-2)}\right).\end{aligned}$$

To test H_0 versus H_1, the following scan statistic is used:

$$\max_{n_0 \le t \le n_1} Z_w(t), \tag{3.5}$$

where n_0 and n_1 are pre-specified constraints for the range of τ. The null hypothesis is rejected if the maxima is greater than a threshold. Details about how to choose the threshold to control the type I error are discussed in Section 4.

For illustration, Figure 4 shows the $Z_w(t)$ processes for the same illustration dataset as in Figure 2. We see that $Z_w(t)$ peaks at the true change-point $\tau = 20$.

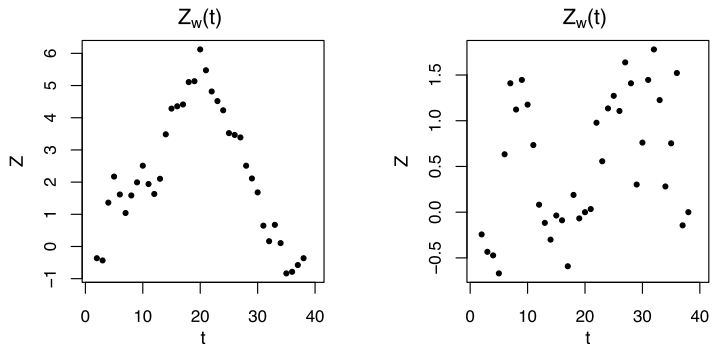


FIG. 4. *On the left, the profile of $Z_w(t)$ against t for the example data set in Figure 2. On the right, the profile of $Z_w(t)$ against t on a sequence of points with no change-point.* (*The scale on the y-axis are different in the two plots.*)

For contrast, when there is no change-point, $Z_w(t)$ exhibits random fluctuation and attains a much smaller maximum value compared to when there is a change-point.

3.3. *Scan statistics for changed interval alternative.* For testing the changed interval alternative H_2 (1.3), each possible interval $(t_1, t_2]$ partitions the observations into two groups: one group containing all observations observed during $(t_1, t_2]$, and the other group containing all observations observed outside of this interval. Then, for any candidate changed interval $(t_1, t_2]$, we have that $R_0(t_1, t_2)$ is the number of edges connecting observations within and outside the interval $(t_1, t_2]$, $R_1(t_1, t_2)$ is the number of edges connecting observations outside of the interval $(t_1, t_2]$, and $R_2(t_1, t_2)$ is the number of edges connecting observations within the interval $(t_1, t_2]$. Then the two-sample test statistics for testing the changed interval alternative can be defined in a similar manner to the single change-point case in Sections 3.1 and 3.2. For example, the generalized edge-count two-sample test statistic, $S(t_1, t_2)$, for testing H_0 (1.1) versus H_2 (1.3) is defined as

$$\begin{pmatrix} R_1(t_1,t_2) - \mathbf{E}\big(R_1(t_1,t_2)\big) \\ R_2(t_1,t_2) - \mathbf{E}\big(R_2(t_1,t_2)\big) \end{pmatrix}^T \boldsymbol{\Sigma}^{-1}(t_1,t_2) \begin{pmatrix} R_1(t_1,t_2) - \mathbf{E}\big(R_1(t_1,t_2)\big) \\ R_2(t_1,t_2) - \mathbf{E}\big(R_2(t_1,t_2)\big) \end{pmatrix}.$$

Under the permutation null, the explicit expression for $\mathbf{E}(R_1(t_1, t_2))$, $\mathbf{E}(R_2(t_1, t_2))$ and the covariance matrix can be obtained similarly as in the single change-point setting. The explicit expressions can be found in Supplement A.1 [Chu and Chen (2019)]. The scan statistic involves a maximization over t_1 and t_2, that is, $\max_{1\le t_1<t_2\le n, l_0\le t_2-t_1\le l_1} S(t_1, t_2)$, where l_0 and l_1 are constraints on the window size. For example, we can set $l_1 = n - l_0$ so that only alternatives where the numbers of observations in either group is larger than l_0 are considered.

Complete details of the generalized edge-count scan statistic and weighted edge-count scan statistic for the changed alternative are given in Supplement A.1 and A.2, respectively [Chu and Chen (2019)].

3.4. *Max-type edge-count two-sample test.* Here, we present a new test statistic, based on the following lemma.

LEMMA 3.1. *The generalized edge-count scan statistic can be expressed as*

$$S(t) = Z_w^2(t) + Z_{diff}^2(t),$$

$$S(t_1, t_2) = Z_w^2(t_1, t_2) + Z_{diff}^2(t_1, t_2),$$

where $Z_w(t)$, $Z_w(t_1, t_2)$ *are the standardized weighted edge-count two-sample test statistic defined in* (3.4) *and* (A.3), *respectively, and*

$$Z_{\text{diff}}(t) = \frac{R_{\text{diff}}(t) - \mathbf{E}(R_{\text{diff}}(t))}{\sqrt{\mathbf{Var}(R_{\text{diff}}(t))}}, \tag{3.6}$$

$$Z_{\text{diff}}(t_1, t_2) = \frac{R_{\text{diff}}(t_1, t_2) - \mathbf{E}(R_{\text{diff}}(t_1, t_2))}{\sqrt{\mathbf{Var}(R_{\text{diff}}(t_1, t_2))}}, \tag{3.7}$$

with $R_{\text{diff}}(t) = R_1(t) - R_2(t)$ *and* $R_{\text{diff}}(t_1, t_2) = R_1(t_1, t_2) - R_2(t_1, t_2)$.

The proof of this lemma is in Supplement C.1 [Chu and Chen (2019)]. The analytical expressions for the expectation and variance of $R_{\text{diff}}(t)$ and $R_{\text{diff}}(t_1, t_2)$ under the permutation null are

$$\mathbf{E}\big(R_{\text{diff}}(t)\big) = |G|\frac{(2t-n)}{n},$$

$$\mathbf{E}\big(R_{\text{diff}}(t_1, t_2)\big) = |G|\frac{(2(t_2-t_1)-n)}{n},$$

$$\mathbf{Var}\big(R_{\text{diff}}(t)\big) = \frac{t(n-t)\big(\sum_{i=1}^{n}|G_i|^2 - \frac{4|G|^2}{n}\big)}{n(n-1)},$$

$$\mathbf{Var}\big(R_{\text{diff}}(t_1, t_2)\big) = \frac{(t_2-t_1)(n+t_2-t_1)\big(\sum_{i=1}^{n}|G_i|^2 - \frac{4|G|^2}{n}\big)}{n(n-1)}.$$

From the above lemma, we can see that $S(t)$ is the sum of squares of two uncorrelated quantities (these two quantities are further asymptotically independent, see in Section 4). Here, $Z_w(t)$ tends to be sensitive to locational alternatives. When the change is locational, $Z_w(t)$ tends to be large. On the other hand, $Z_{\text{diff}}(t)$ tends to be sensitive to scale alternative. When the change is in the spread of the distribution,

$|Z_{\text{diff}}(t)|$ tends to be large. The sign of $Z_{\text{diff}}(t)$ depends on whether the distribution after the change has a larger spread or not. Hence, we propose the following max-type edge-count two-sample test statistic:

$$M(t) = \max\bigl(\bigl|Z_{\text{diff}}(t)\bigr|, Z_w(t)\bigr) \tag{3.8}$$

for the single change-point alternative and

$$M(t_1, t_2) = \max\bigl(\bigl|Z_{\text{diff}}(t_1, t_2)\bigr|, Z_w(t_1, t_2)\bigr) \tag{3.9}$$

for the changed-interval alternative. The corresponding scan statistics are

$$\max_{n_0 \le t \le n_1} M(t), \tag{3.10}$$

for the single change-point alternative and

$$\max_{\substack{1 \le t_1 < t_2 \le n \\ l_0 \le t_2 - t_1 \le l_1}} M(t_1, t_2), \tag{3.11}$$

for the changed-interval alternative.

As it will come later, this max-type statistic is of particular interest as its performance is similar to $S(t)$ and we can obtain more accurate p-value approximations (details in Section 4).

A more detailed discussion on the relationship of the three test statistics (S, Z_w, M) and an extension to the max-type statistic can be found in Supplement H [Chu and Chen (2019)].

4. Analytical p-value approximations. Given the scan statistics, the next question is how large do they need to be to constitute sufficient evidence against the null hypothesis of homogeneity. In order words, we are concerned with the tail probability of the scan statistics under H_0. For the generalized edge-count two-sample test, that is,

$$\mathbf{P}\Bigl(\max_{n_0 \le t \le n_1} S(t) > b\Bigr) \tag{4.1}$$

for the single change-point alternative, and

$$\mathbf{P}\Bigl(\max_{\substack{1 \le t_1 < t_2 \le n \\ l_0 \le t_2 - t_1 \le l_1}} S(t_1, t_2) > b\Bigr) \tag{4.2}$$

for the changed interval alternative. For the weighted edge-count two-sample test, that is,

$$\mathbf{P}\Bigl(\max_{n_0 \le t \le n_1} Z_w(t) > b\Bigr) \tag{4.3}$$

for the single change-point alternative, and

$$\mathbf{P}\Bigl(\max_{\substack{1 \le t_1 < t_2 \le n \\ l_0 \le t_2 - t_1 \le l_1}} Z_w(t_1, t_2) > b\Bigr) \tag{4.4}$$

for the changed interval alternative. For the max-type edge-count two-sample test, that is,

$$\mathbf{P}\Big(\max_{n_0\le t\le n_1} M(t) > b\Big) \tag{4.5}$$

for the single change-point alternative, and

$$\mathbf{P}\Big(\max_{\substack{1\le t_1<t_2\le n\\ l_0\le t_2-t_1\le l_1}} M(t_1,t_2) > b\Big) \tag{4.6}$$

for the changed interval alternative.

For small n, we can directly sample from the permutation distribution to approximate (4.1)–(4.6). However, when n is large, permutation is very time consuming. Therefore, to make the method instantly applicable, we derive analytical expressions to approximate these tail probabilities.

To derive the analytical expressions, we study the asymptotic properties of the stochastic processes $\{S(t)\}$, $\{S(t_1,t_2)\}$, $\{Z_w(t)\}$, $\{Z_w(t_1,t_2)\}$, $\{M(t)\}$ and $\{M(t_1,t_2)\}$, and then make adjustments for finite samples. By Lemma 3.1 and how $M(t)$ is defined, these stochastic processes boil down to two pairs of basic processes: $\{Z_{\mathrm{diff}}(t)\}$ and $\{Z_w(t)\}$ for the single change-point case and $\{Z_{\mathrm{diff}}(t_1,t_2)\}$ and $\{Z_w(t_1,t_2)\}$ for changed-interval. So we first study the properties of these basic stochastic processes.

4.1. *Asymptotic null distributions of the basic processes.* In this section, we derive the limiting distributions of $\{Z_{\mathrm{diff}}([nu]) : 0 < u < 1\}$ and $\{Z_w([nu]) : 0 < u < 1\}$ for the single change-point alternative, and $\{Z_{\mathrm{diff}}([nu],[nv]) : 0 < u < v < 1\}$ and $\{Z_w([nu],[nv]) : 0 < u < v < 1\}$ for the changed-interval alternative (we use $[x]$ to denote the largest integer that is no larger than x).

We first introduce some notation. For edge $e = (e_-, e_+)$, where $e_- < e_+$ are the indices of the nodes connected by the edge e, let

$$A_e = G_{e_-} \cup G_{e_+}, \tag{4.7}$$

be the subgraph in G that connect to either node e_- or node e_+, and

$$B_e = \bigcup_{e^*\in A_e} A_{e^*}, \tag{4.8}$$

be the subgraph in G that connect to any edge in A_e.

In the following, we write $a_n = O(b_n)$ when a_n has the same order as b_n, and write $a_n = o(b_n)$ when a_n has order smaller than b_n.

THEOREM 4.1. *When* $|G| = O(n^\alpha)$, $1 \le \alpha < 1.5$, $\sum_{e\in G} |A_e||B_e| = o(n^{1.5\alpha})$, $\sum_{e\in G} |A_e|^2 = o(n^{\alpha+0.5})$, *and* $\sum_{i=1}^n |G_i|^2 - \frac{4|G|^2}{n} = O(\sum_{i=1}^n |G_i|^2)$, *as* $n\to\infty$:

1. $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$ *and* $\{Z_w([nu]) : 0 < u < 1\}$ *converge to independent Gaussian processes in finite dimensional distributions, which we denote as* $\{Z^*_{\text{diff}}(u) : 0 < u < 1\}$ *and* $\{Z^*_w(u) : 0 < u < 1\}$, *respectively.*
2. $\{Z_{\text{diff}}([nu], [nv]) : 0 < u < v < 1\}$ *and* $\{Z_w([nu], [nv]) : 0 < u < v < 1\}$ *converge to independent two-dimension Gaussian random fields in finite dimensional distributions, which we denote as* $\{Z^*_{\text{diff}}(u, v) : 0 < u < v < 1\}$ *and* $\{Z^*_w(u, v) : 0 < u < v < 1\}$, *respectively.*

The proof for this theorem utilizes Stein's method [Chen and Shao (1994)] and the details of the proof are in Supplement C.2 [Chu and Chen (2019)].

REMARK 4.2. The condition $|G| = O(n^\alpha), 1 \le \alpha < 1.5$ ensures that the graph is dense enough but not too dense. The conditions $\sum_{e\in G} |A_e||B_e| = o(n^{1.5\alpha})$ and $\sum_{e\in G} |A_e|^2 = o(n^{\alpha+0.5})$ ensure that the graph does not have a large hub or a cluster of small hubs, where a hub is a node with a large degree. The condition $\sum_{i=1}^n |G_i|^2 - \frac{4|G|^2}{n} = O(\sum_{i=1}^n |G_i|^2)$ ensures Z_{diff} to be well defined.

These conditions are quite mild. For example, for k-MST, when $k = O(1)$, we have $|G| = k(n-1) = O(n)$, and the conditions boil down to $\sum_{e\in G} |A_e||B_e| = o(n^{1.5})$ and $\sum_{i=1}^n |G_i|^2 - \frac{4|G|^2}{n} = O(\sum_{i=1}^n |G_i|^2)$. Based on Theorems 5.1 and 5.2 in Chen and Friedman (2017), both conditions are satisfied for k-MST constructed on Euclidean distance for $k = O(1)$.

More discussions on the conditions of the graph can be found in Supplement G [Chu and Chen (2019)].

Let $\rho^*_w(u, v) = \mathbf{Cov}(Z^*_w(u), Z^*_w(v))$ and $\rho^*_{\text{diff}}(u, v) = \mathbf{Cov}(Z^*_{\text{diff}}(u), Z^*_{\text{diff}}(v))$. The next theorem state explicitly the covariance functions of the limiting Gaussian processes, $\{Z^*_w(u), 0 < u < 1\}$ and $\{Z^*_{\text{diff}}(u), 0 < u < 1\}$.

THEOREM 4.3. *The exact expressions for* $\rho^*_{\text{diff}}(u, v)$ *and* $\rho^*_w(u, v)$ *are*

$$\rho^*_w(u, v) = \frac{(u \wedge v)(1 - (u \vee v))}{(u \vee v)(1 - (u \wedge v))},$$

$$\rho^*_{\text{diff}}(u, v) = \frac{(u \wedge v)(1 - (u \vee v))}{\sqrt{(u \wedge v)(1 - (u \wedge v))(u \vee v)(1 - (u \vee v))}},$$

where $u \wedge v = \min(u, v)$ *and* $u \vee v = \max(u, v)$.

The above theorem is proved through combinatorial analysis and details are given in the Supplement C.3 [Chu and Chen (2019)]. From the above theorem, we see that the limiting processes, $\{Z^*_w(u), 0 < u < 1\}$ and $\{Z^*_{\text{diff}}(u), 0 < u < 1\}$, do not depend on G at all.

4.2. *Asymptotic p-value approximations*. We now examine the asymptotic behavior of the tail probabilities (4.1)–(4.6). Our approximations require the function $\nu(x)$ defined as

$$\nu(x) = 2x^{-2}\exp\biggl(-2\sum_{m=1}^{\infty} m^{-1}\Phi\Bigl(-\frac{1}{2}xm^{1/2}\Bigr)\biggr), \qquad x > 0. \tag{4.9}$$

This function is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A simple approximation given in Siegmund and Yakir (2007) is sufficient for numerical purpose:

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2) - 0.5)}{(x/2)\Phi(x/2) + \phi(x/2)}, \tag{4.10}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative density function and standard normal density function, respectively. Following similar arguments in the proof for Proposition 3.4 in Chen and Zhang (2015), when the conditions on G in Theorem 4.1 hold, $n, b, n_0, n_1 \to \infty$ in a way such that for some $b_0 > 0$ and $0 < x_0 < x_1 < 1$, $b/\sqrt{n} \to b_0$, $\frac{n_0}{n} \to x_0$ and $\frac{n_1}{n} \to x_1$, then as $n \to \infty$, we have

$$\begin{aligned}
&\mathbf{P}\Bigl(\max_{n_0 \le t \le n_1} Z_w^*(t/n) > b\Bigr) \\
&\qquad \sim b\phi(b)\int_{x_0}^{x_1} h_w^*(x)\nu\bigl(b_0\sqrt{2h^*(x)}\bigr)\,dx, \\
&\mathbf{P}\Bigl(\max_{n_0 \le t_2 - t_1 \le n_1} Z_w^*(t_1/n, t_2/n) > b\Bigr) \\
&\qquad \sim b^3\phi(b)\int_{x_0}^{x_1} \Bigl(h_w^*(x)\nu\Bigl(b_0\sqrt{2h_w^*(x)}\Bigr)\Bigr)^2(1-x)\,dx, \\
&\mathbf{P}\Bigl(\max_{n_0 \le t \le n_1} |Z_{\mathrm{diff}}^*(t/n)| > b\Bigr) \\
&\qquad \sim 2b\phi(b)\int_{x_0}^{x_1} h_{\mathrm{diff}}^*(x)\nu\Bigl(b_0\sqrt{2h_{\mathrm{diff}}^*(x)}\Bigr)\,dx, \\
&\mathbf{P}\Bigl(\max_{n_0 \le t_2 - t_1 \le n_1} |Z_{\mathrm{diff}}^*(t_1/n, t_2/n)| > b\Bigr) \\
&\qquad \sim 2b^3\phi(b)\int_{x_0}^{x_1} \Bigl(h_{\mathrm{diff}}^*(x)\nu\Bigl(b_0\sqrt{2h_{\mathrm{diff}}^*(x)}\Bigr)\Bigr)^2(1-x)\,dx,
\end{aligned}$$

where

$$\begin{aligned}
h_w^*(x) &= \lim_{u \nearrow x} \frac{\partial \rho_w^*(u,x)}{\partial u} \equiv -\lim_{u \searrow x} \frac{\partial \rho_w^*(u,x)}{\partial u}, \\
h_{\mathrm{diff}}^*(x) &= \lim_{u \nearrow x} \frac{\partial \rho_{\mathrm{diff}}^*(u,x)}{\partial u} \equiv -\lim_{u \searrow x} \frac{\partial \rho_{\mathrm{diff}}^*(u,x)}{\partial u}.
\end{aligned}$$

It can be shown that

$$h_w^*(x) = \frac{1}{x(1-x)}, \tag{4.11}$$

$$h_{\mathrm{diff}}^*(x) = \frac{1}{2x(1-x)}. \tag{4.12}$$

Since Z_w^* and Z_{diff}^* are independent, we have

$$\begin{aligned}
&\mathbf{P}\Big(\max_{n_0\le t\le n_1} M^*(t/n) > b\Big)\\
&\quad = 1 - \mathbf{P}\Big(\max_{n_0\le t\le n_1} \big|Z_{\mathrm{diff}}^*(t)\big| < b\Big)\mathbf{P}\Big(\max_{n_0\le t\le n_1} Z_w^*(t) < b\Big),\\
&\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} M^*(t_1/n, t_2/n) > b\Big)\\
&\quad = 1 - \mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} \big|Z_{\mathrm{diff}}^*(t_1,t_2)\big| < b\Big)\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} Z_w^*(t_1,t_2) < b\Big).
\end{aligned}$$

For the tail probabilities for $\max_{n_0\le t\le n_1} S(t)$ and $\max_{l_0\le t_2-t_2\le l_1} S(t_1,t_2)$, some additional works are needed and the results are stated in the following proposition.

PROPOSITION 4.4. *Assume that* $|G| = O(n^\alpha)$, $1 \le \alpha < 1.5$, $\sum_{e\in G}|A_e||B_e| = o(n^{1.5\alpha})$, $\sum_{e\in G}|A_e|^2 = o(n^{\alpha+0.5})$, *and* $\sum_{i=1}^n |G_i|^2 - \frac{4|G|^2}{n} = O(\sum_{i=1}^n |G_i|^2)$, $n, b, n_0, n_1 \to \infty$ *in a way such that for some* $b_1 > 0$ *and* $0 < x_0 < x_1 < 1$, $b/n \to b_1$, $\frac{n_0}{n} \to x_0$ *and* $\frac{n_1}{n} \to x_1$, *then as* $n \to \infty$,

$$\begin{aligned}
&\mathbf{P}\Big(\max_{n_0\le t\le n_1} S^*(t/n) > b\Big)\\
&\quad \approx \frac{b\,e^{-b/2}}{2\pi}\int_0^{2\pi}\int_{x_0}^{x_1} u^*(x,\omega)\nu\big(\sqrt{2b_1u^*(x,\omega)}\big)\,dx\,d\omega,
\end{aligned} \tag{4.13}$$

$$\begin{aligned}
&\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} S^*(t_1/n, t_2/n) > b\Big)\\
&\quad \approx \frac{b^2e^{-b/2}}{\pi}\int_0^{2\pi}\int_{x_0}^{x_1} \big(u^*(x,\omega)\nu\big(\sqrt{2b_1u^*(x,\omega)}\big)\big)^2(1-x)\,dx\,d\omega,
\end{aligned} \tag{4.14}$$

where $u^*(x,\omega) = h_w^*(x)\sin^2(\omega) + h_{\mathrm{diff}}^*(x)\cos^2(\omega)$, *with* $h_w^*(x)$ *and* $h_{\mathrm{diff}}^*(x)$ *provided in* (4.11) *and* (4.12), *respectively.*

The proof of this proposition is in Supplement C.4 [Chu and Chen (2019)].

Based on the above results, we can approximate the tail probabilities (4.1)–(4.6) by

$$\mathbf{P}\Big(\max_{n_0\le t\le n_1} S(t) > b\Big) \approx \frac{be^{-b/2}}{2\pi}\int_0^{2\pi}\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} u^*(x,\omega)\nu\big(\sqrt{2bu^*(x,\omega)/n}\big)\,dx\,d\omega, \tag{4.15}$$

$$\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} S(t_1,t_2) > b\Big) \approx \frac{b^2e^{-b/2}}{\pi}\int_0^{2\pi}\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} \big(u^*(x,\omega)\nu\big(\sqrt{2bu^*(x,\omega)/n}\big)\big)^2(1-x)\,dx\,d\omega, \tag{4.16}$$

$$\mathbf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) > b\Big) \approx b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_w^*(x)\nu\big(b\sqrt{2h^*(x)/n}\big)\,dx, \tag{4.17}$$

$$\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} Z_w(t_1,t_2) > b\Big) \approx b^3\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} \Big(h_w^*(x)\nu\Big(b\sqrt{2h_w^*(x)/n}\Big)\Big)^2(1-x)\,dx, \tag{4.18}$$

$$\mathbf{P}\Big(\max_{n_0\le t\le n_1} M(t) > b\Big) = 1-\mathbf{P}\Big(\max_{n_0\le t\le n_1}\big|Z_{\mathrm{diff}}(t)\big| < b\Big)\mathbf{P}\Big(\max_{n_0\le t\le n_1} Z_w(t) < b\Big), \tag{4.19}$$

$$\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} M(t_1,t_2) > b\Big) = 1-\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1}\big|Z_{\mathrm{diff}}(t_1,t_2)\big| < b\Big)\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} Z_w(t_1,t_2) < b\Big), \tag{4.20}$$

where

$$\mathbf{P}\Big(\max_{n_0\le t\le n_1}\big|Z_{\mathrm{diff}}(t)\big| < b\Big) \approx 1-2b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} h_{\mathrm{diff}}^*(x)\nu\Big(b\sqrt{2h_{\mathrm{diff}}^*(x)/n}\Big)\,dx \tag{4.21}$$

and

$$\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1}\big|Z_{\mathrm{diff}}(t_1,t_2)\big|<b\Big)$$

$$\approx 1-2b^3\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}}\Big(h^*_{\mathrm{diff}}(x)\nu\Big(b\sqrt{2h^*_{\mathrm{diff}}(x)/n}\Big)\Big)^2(1-x)\,dx, \tag{4.22}$$

and $\mathbf{P}(\max_{n_0\le t\le n_1} Z_w(t)<b)$ and $\mathbf{P}(\max_{n_0\le t_2-t_1\le n_1} Z_w(t_1,t_2)<b)$ easily follow from (4.17) and (4.18), respectively.

REMARK 4.5. In practice, when using (4.15)–(4.20) to approximate the tail probabilities, we use $h_w(n,x)$ in place of $h^*_w(x)$, where $h_w(n,x)$ is the finite-sample equivalent of $h^*_w(x)$. That is,

$$h_w(n,x)=n\lim_{s\nearrow nx}\frac{\partial\rho_w(s,nx)}{\partial s},$$

with $\rho_w(s,t):=\mathbf{Cov}(Z_w(s),Z_w(t))$. The explicit expression for $h_w(n,x)$ can also be derived and simplified to be

$$h_w(n,x)=\frac{(n-1)(2nx^2-2nx+1)}{2x(1-x)(n^2x^2-n^2x+n-1)}. \tag{4.23}$$

It is clear from the above expression that $h_w(n,x)$ does not depend on the graph G as well. Also, it is easy to show that $\lim_{n\to\infty}h_w(n,x)=h^*_w(x)$.

The finite-sample equivalent version of $h^*_{\mathrm{diff}}(x)$ is exact the same as $h^*_{\mathrm{diff}}(x)$. That is,

$$h_{\mathrm{diff}}(n,x)=n\lim_{s\nearrow nx}\frac{\partial\,\mathbf{Cov}(Z_{\mathrm{diff}}(s),Z_{\mathrm{diff}}([nx]))}{\partial s}=\frac{1}{2x(1-x)}.$$

4.3. *Skewness correction.* Analytical approximations become less precise when the minimum window length decreases (see numerical results in Section 4.4). This is mainly because the convergence of $Z_w(t)$ and $Z_{\mathrm{diff}}(t)$ to normal is slow if t/n is close to 0 or 1 and the convergence of $Z_w(t_1,t_2)$ and $Z_{\mathrm{diff}}(t_1,t_2)$ to normal is slow if $\frac{t_2-t_1}{n}$ is close to 0 or 1. This problem becomes more severe when dimension is high. Figure 5 plots the skewness of $Z_w(t)$ and $Z_{\mathrm{diff}}(t)$ with G being MST constructed on the Euclidean distance. We can see from the plot that the statistic $Z_w(t)$ is right skewed. The p-value approximations (4.17) and (4.18) would then underestimate the true tail probabilities. On the other hand, $Z_{\mathrm{diff}}(t)$ is right skewed for small values of t and left skewed for large values of t, which would also affect the analytic p-value approximation derived based on asymptotic results.

Hence, we perform skewness correction to improve the analytical p-value approximations for finite sample sizes. As illustrated in Figure 5, the extent of the skewness depends on t, so we adopt a skewness correction approach discussed in

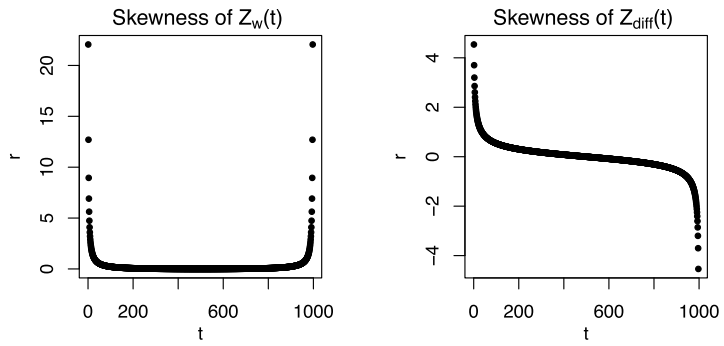


FIG. 5. *Plots of skewness of $Z_w(t)$ and of $Z_{\text{diff}}(t)$ against t for a sequence of* 1000 *points randomly generated from $\mathcal{N}(0, I_{100})$. The graph is MST constructed on Euclidean distance.*

Chen and Zhang (2015) that does the correction up to different extents based on the amount of skewness at each value of t. In particular, the approach provides a better approximation to the marginal probability, $\mathbf{P}(Z_w(t) \in b + dx/b)$, $\mathbf{P}(Z_{\text{diff}}(t) \in b + dx/b)$, $\mathbf{P}(Z_w(t_1, t_2) \in b + dx/b)$ and $\mathbf{P}(Z_{\text{diff}}(t_1, t_2) \in b + dx/b)$, through a cumulant generating function $\psi(\theta) = \log \mathbf{E}_P(e^{\theta z})$. By applying a change of measure $dQ_\theta = e^{\theta Z - \psi(\theta)}\, dP$, we can approximate the marginal probability by

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

where θ_b is chosen such that $\dot{\psi}(\theta_b) = b$. By a third Taylor approximation, we get $\theta_b \approx (-1 + \sqrt{1+2\gamma b})/\gamma$, where $\gamma := \mathbf{E}_{\mathbf{P}}(Z^3)$.

Notice that $\mathbf{E}(Z_w^3(t_1, t_2)) = \mathbf{E}(Z_w^3(t_2 - t_1))$ and $\mathbf{E}(Z_{\text{diff}}^3(t_1, t_2)) = \mathbf{E}(Z_{\text{diff}}^3(t_2 - t_1))$. Let $\gamma_w(t) = \mathbf{E}(Z_w^3(t))$ and $\gamma_{\text{diff}}(t) = \mathbf{E}(Z_{\text{diff}}^3(t))$. The p-value approximations, after correcting for skewness, are

$$\begin{aligned}&\mathbf{P}\Bigl(\max_{n_0 \le t \le n_1} Z_w(t) > b\Bigr)\\ &\approx b\phi(b) \int_{\frac{n_0}{n}}^{\frac{n_1}{n}} K_w(nx) h_w(n,x) \nu\bigl(b\sqrt{2h_w(n,x)/n}\bigr)\, dx,\end{aligned} \tag{4.24}$$

$$\begin{aligned}&\mathbf{P}\Bigl(\max_{n_0 \le t_2 - t_1 \le n_1} Z_w(t_1, t_2) > b\Bigr)\\ &\approx b^3\phi(b) \int_{\frac{n_0}{n}}^{\frac{n_1}{n}} K_w(nx) \bigl(h_w(n,x) \nu\bigl(b\sqrt{2h_w(n,x)/n}\bigr)\bigr)^2 (1-x)\, dx,\end{aligned} \tag{4.25}$$

where $K_w(t) = \frac{\exp(\frac{1}{2}(b-\hat{\theta}_{b,w}(t))^2+\frac{1}{6}\gamma_w(t)\hat{\theta}_{b,w}(t)^3)}{\sqrt{1+\gamma_w(t)\hat{\theta}_{b,w}(t)}}$ with $\hat{\theta}_{b,w}(t) = \frac{-1+\sqrt{1+2\gamma_w(t)b}}{\gamma_w(t)}$, and

$$\mathbf{P}\Big(\max_{n_0\le t\le n_1} Z_{\text{diff}}(t) > b\Big) \approx b\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} K_{\text{diff}}(nx)h_{\text{diff}}(n,x)v\big(b\sqrt{2h_{\text{diff}}(n,x)/n}\big)\,dx, \tag{4.26}$$

$$\mathbf{P}\Big(\max_{n_0\le t_2-t_1\le n_1} Z_{\text{diff}}(t_1,t_2) > b\Big) \approx b^3\phi(b)\int_{\frac{n_0}{n}}^{\frac{n_1}{n}} K_{\text{diff}}(nx)\big(h_{\text{diff}}(n,x)v\big(b\sqrt{2h_{\text{diff}}(n,x)/n}\big)\big)^2(1-x)\,dx, \tag{4.27}$$

where $K_{\text{diff}}(t) = \frac{\exp(\frac{1}{2}(b-\hat{\theta}_{b,\text{diff}}(t))^2+\frac{1}{6}\gamma_{\text{diff}}(t)\hat{\theta}_{b,\text{diff}}(t)^3)}{\sqrt{1+\gamma_{\text{diff}}(t)\hat{\theta}_{b,\text{diff}}(t)}}$ with $\hat{\theta}_{b,\text{diff}}(t) = \frac{-1+\sqrt{1+2\gamma_{\text{diff}}(t)b}}{\gamma_{\text{diff}}(t)}$.

The only unknown quantities in the above expressions are $\gamma_w(t)$ and $\gamma_{\text{diff}}(t)$. Since

$$\mathbf{E}\big[Z_w^3(t)\big] = \frac{\mathbf{E}(R_w^3(t)) - 3\mathbf{E}(R_w(t))\,\mathbf{Var}(R_w(t)) - \mathbf{E}^3(R_w(t))}{(\mathbf{Var}(R_w(t)))^{3/2}},$$

$$\mathbf{E}\big[Z_{\text{diff}}^3(t)\big] = \frac{\mathbf{E}(R_{\text{diff}}^3(t)) - 3\mathbf{E}(R_{\text{diff}}(t))\,\mathbf{Var}(R_{\text{diff}}(t)) - \mathbf{E}^3(R_{\text{diff}}(t))}{(\mathbf{Var}(R_{\text{diff}}(t)))^{3/2}},$$

and the analytic expressions for the expectation and variance of $R_w(t)$ and $R_{\text{diff}}(t)$ can be found in Section 3, we only need to figure out the analytic expressions of $\mathbf{E}(R_w^3(t))$ and $\mathbf{E}(R_w^3(t_1,t_2))$. The exact analytic expressions of $\mathbf{E}(R_w^3(t))$ and $\mathbf{E}(R_w^3(t_1,t_2))$ are quite long and they are provided in Appendix B.

REMARK 4.6. When the marginal distribution is highly left-skewed, it is possible that the third moment of the test statistic, $\gamma(t)$, is too small for $1+2\gamma(t)b$ to be positive. In order to obtain a better approximation to θ_b, higher moments are needed. However, since this problem usually occurs when t/n is close to 0 or 1, we apply a heuristic fix discussed in Chen and Zhang (2015) that extrapolates $\hat{\theta}$ by using its values outside the problematic region.

REMARK 4.7. Skewness corrected p-value approximations for $\max_{n_0\le t\le n_1} S(t) = \max_{0\le w\le 2\pi}\max_{n_0\le t\le n_1}(Z_w(t)\sin(w) + Z_{\text{diff}}(t)\cos(w))$ can be derived by jointly correcting for the marginal probabilities of $Z_w(t)$ and $Z_{\text{diff}}(t)$. After correcting for skewness, the integrand in (4.15) becomes

$$K_S(x,\omega)u(x,\omega)v\big(\sqrt{2bu(x,\omega)/n}\big),$$

where

$$K_S(t,\omega) = \Bigg[\exp\Bigg(\frac{1}{2}\big((\sqrt{b}\cos(\omega) - \hat{\theta}_{b,1}(t)\big)^2 + \big(\sqrt{b}\sin(\omega) - \hat{\theta}_{b,2}(t)\big)^2\big) + \frac{1}{6}\big(\gamma_1(t)\hat{\theta}_{b,1}(t)^3 + \gamma_2(t)\hat{\theta}_{b,2}(t)^3\big)\Bigg)\Bigg] \times \big[\sqrt{\big(1+\gamma_1(t)\hat{\theta}_{b,1}\big)\big(1+\gamma_2(t)\hat{\theta}_{b,2}\big)}\big]^{-1}$$

with $\gamma_1(t) = \mathbf{E}[Z_1^3(t)]$, $\hat{\theta}_{b,1}(t,\omega) = \frac{-1+\sqrt{1+2\gamma_1(t)\sqrt{b}\cos(\omega)}}{\gamma_1(t)}$, and $\gamma_2(t)$ and $\hat{\theta}_{b,2}(t,\omega)$ defined similarly. However, this integrand could easily be nonfinite in each quadrant in terms of w, and the method relies heavily on extrapolation. We thus do not perform skewness correction on $S(t)$.

4.4. *Checking p-value approximations for finite samples*. Here, we check how the p-value approximations based on asymptotic results directly and with skewness correction work for finite samples. To do so, we compare the critical values obtained from (4.15), (4.17), (4.19), (4.24) and (4.26) to the critical values obtained from doing 10,000 permutations directly, under various simulation settings. We here focus on the single change-point alternative here. For the changed interval alternative, the results are similar and details can be found in Supplement D.2 [Chu and Chen (2019)].

In each simulation, sequences of length 1000 were generated from a given distribution F_0 in $\mathbb{R}^d$. We considered three distributions (multivariate normal, multivariate t with 5 degrees of freedom and multivariate log-normal) under various dimensions ($d = 10$, $d = 100$, and $d = 1000$). Here, we present the results only for multivariate normal with $d = 10$ (denoted by (C1) in Tables 2, 3 and 4), multivariate t_5 with $d = 100$ [denoted by (C2)], and multivariate log-normal with $d = 1000$ [denoted by (C3)]. The complete tables showing all three distributions under these three dimensions with more cases are in Supplement D.1 [Chu and Chen (2019)]. The analytical approximations depend on constraints on the sequence in which the change-point is searched over (n_0 and n_1). To make things simple, we let $n_1 = n - n_0$.

Since the asymptotic p-value approximations (without skewness correction) do not depend on G, the critical value is determined by n, n_0 and n_1 only (here, n_1 is set to be $n - n_0$). The first table of Tables 2, 3 and 4 labeled "A1" presents the analytical critical values without skewness correction. On the other hand, the skewness corrected p-value approximations and permutation p-values depend on certain characteristics of the structure of the graph G. In this simulation, the MST is used. As the structure of MST depends on the observations, the critical value vary by simulation runs. We show results for 2 randomly simulated sequences in each setting. Two characteristics of the graph are also reported: the sum of squared node degrees ($\sum_i |G_i|^2$) and the maximum node degree ($d_{\max}$). These quantities

TABLE 2
Critical values for the single change-point scan statistic $\max_{n_0 \le t \le n_1} S(t)$ *based on MST at* 0.05 *significance level.* $n = 1000$

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
A1	13.10	13.38	13.70	14.11

	Critical values				Graph	
	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$		
	Per	Per	Per	Per	$\sum \lvert G_i \rvert^2$	$d_{\max}$
(C1)	12.87	13.29	14.04	15.17	5394	8
	13.02	13.42	13.71	15.65	5368	8
(C2)	13.47	14.20	15.48	17.81	14,302	42
	13.32	13.77	14.96	17.11	12,424	39
(C3)	14.50	15.83	18.14	21.96	46,876	83
	16.12	18.38	22.00	29.07	106,524	208

give some intuitions on the size and density of the hubs in the graph. The skewness corrected critical values are presented in Tables 3 and 4 under the column "A2." The column "Per" denotes critical values obtained through 10,000 random permutations directly.

We first focus on the results of the generalized edge-count test statistic $\max S(t)$. Since we do not perform skewness correction for $S(t)$, Table 2 compares these analytical critical values (A1) with the critical values obtained from doing 10,000 permutations (Per). The main factors that influence the approximation accuracy of the analytical critical values are the minimum window size (n_0) and the structure of the graph. We see that, when the graph is relatively flat [such as in (C1) that the largest degree in the graph is relatively small], the asymptotic p-value approximation is doing reasonably well when $n_0 \ge 50$. As the graph becomes to have larger and larger hubs, n_0 needs to be larger to achieve a similar degree of accuracy.

Table 3 shows the results for $\max Z_w(t)$. Similar to $S(t)$, as window size decreases and/or the maximum degree in the graph increases, the analytical critical values become less precise. However, the skewness corrected critical values perform much better than the critical values without skewness correction. Under (C1), the maximum degree of the graph is in general small and the skewness-corrected p-value approximations are doing reasonably well for n_0 as low as 25. When the maximum degree of the graph is less than 50, the skewness-corrected p-value approximations are doing quite well for $n_0 \ge 50$ and not bad for $n_0 = 25$. For even larger maximum degree scenarios, the skewness-corrected p-value approximations are somewhat less conservative for $n_0 \le 100$ but the discrepancy is not that bad.

TABLE 3
Critical values for the single change-point scan statistic $\max_{n_0 \le t \le n_1} Z_w(t)$ *based on MST at* 0.05 *significance level.* $n = 1000$

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
A1	2.98	3.02	3.08	3.14

	Critical values								Graph	
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$			
	A2	Per	A2	Per	A2	Per	A2	Per	$\sum \lvert G_i \rvert^2$	$d_{\max}$
(C1)	3.05	3.02	3.12	3.11	3.22	3.22	3.4	3.48	5518	10
	3.05	3.05	3.12	3.14	3.22	3.25	3.4	3.45	5442	8
(C2)	3.05	3.04	3.12	3.15	3.22	3.31	3.39	3.62	14,302	42
	3.05	3.06	3.12	3.13	3.22	3.29	3.39	3.54	12,424	39
(C3)	3.04	3.11	3.11	3.25	3.21	3.37	3.38	3.82	46,876	83
	3.03	3.20	3.10	3.40	3.19	3.61	3.35	3.99	106,524	208

TABLE 4
Critical values for the single change-point scan statistic $\max_{n_0 \le t \le n_1} M(t)$ *based on MST at* 0.05 *significance level.* $n = 1000$

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
A1	3.23	3.27	3.32	3.38

	Critical values								Graph	
	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$			
	A2	Per	A2	Per	A2	Per	A2	Per	$\sum \lvert G_i \rvert^2$	$d_{\max}$
(C1)	3.27	3.26	3.33	3.34	3.41	3.42	3.56	3.66	5518	10
	3.27	3.29	3.33	3.34	3.41	3.44	3.56	3.67	5442	8
(C2)	3.30	3.33	3.38	3.44	3.48	3.55	3.67	3.89	14,302	42
	3.29	3.31	3.36	3.40	3.46	3.54	3.64	3.85	12,424	39
(C3)	3.33	3.34	3.41	3.49	3.53	3.69	3.74	4.22	46,876	83
	3.39	3.51	3.49	3.75	3.63	4.06	3.88	4.58	106,524	208

Table 4 shows the results for $\max M(t)$. The pattern is somewhat similar to that for $\max Z_w(t)$ with the skewness-corrected p-value approximations for $\max M(t)$ slightly more tolerant for hubs. When the dimension is not too high [(C1) and (C2)], the maximum degree is less than 50, and the skewness-corrected p-value approximations are working very well when $n_0 \geq 50$. When the maximum degree is large (C3), the skewness-corrected p-value approximations are still doing pretty well for $n_0 \geq 75$ in general.

Overall, we see that the asymptotic critical values are on the right scale and are enough for detecting big changes. However, if one would like to have more accurate critical values, the skewness correction versions are recommended. When this is needed, it would be good to first check the structure of the graph, such as its maximum degree, so that we have a better idea on how well the critical values are.

5. Performance analysis. Here, we examine the performance of the three new test statistics under more settings through simulation studies. Since the proposed tests do not require the data to be from any specific distribution family, there are many possible alternatives. To have a good idea of the performance of the proposed tests, we examine the Gaussian data [$\mathbf{y_i} \sim N_d(\mu, \Sigma)$] where likelihood-based methods are available. We also checked other distributions to check the robustness of the tests in terms of the underlying distribution and these tables can be found in Supplement E [Chu and Chen (2019)].

Under the Gaussian setting, if one assumes that, at the change-point, only the mean ($\boldsymbol{\mu}$) may change, the scan statistic over Hotelling's T^2 statistics can be used: $\max_{n_0 \leq t \leq n_1} \mathrm{HT}(t)$, with $\mathrm{HT}(t) = \frac{t(n-t)}{n}(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}_t^*)^T \tilde{\Sigma}_t^{-1}(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}_t^*)$ where $\bar{\mathbf{y}}_t = \sum_{i=1}^t \mathbf{y}_i/t$, $\bar{\mathbf{y}}_t^* = \sum_{i=t+1}^n \mathbf{y}_i/(n-t)$, and $\tilde{\Sigma}_t = (\sum_{i=1}^t (\mathbf{y}_i - \bar{\mathbf{y}}_t)(\mathbf{y}_i - \bar{\mathbf{y}}_t)^T + \sum_{i=t+1}^n (\mathbf{y}_i - \bar{\mathbf{y}}_t^*)(\mathbf{y}_i - \bar{\mathbf{y}}_t^*)^T)/(n-2)$. If the variance may also change at the change-point, the scan statistic over the generalized likelihood ratio statistic can be used: $\max_{n_0 \leq t \leq n_1} \mathrm{GLR}(t)$ with $\mathrm{GLR}(t) = n \log|\hat{\Sigma}_n| - t \log|\hat{\Sigma}_t| - (n-t)\log|\hat{\Sigma}_t^*|$, where $\hat{\Sigma}_t = \frac{\sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}}_t)(\mathbf{y}_i - \bar{\mathbf{y}}_t)^T}{t}$ and $\hat{\Sigma}_t^* = \frac{\sum_{i=t+1}^n (\mathbf{y}_i - \bar{\mathbf{y}}_t^*)(\mathbf{y}_i - \bar{\mathbf{y}}_t^*)^T}{n-t}$.

In each simulation, we generated a sequence of $n = 200$ observations for various dimensions d with $\mathbf{y}_1, \ldots, \mathbf{y}_\tau \overset{\text{i.i.d.}}{\sim} F_0$ and $\mathbf{y}_{\tau+1}, \ldots, \mathbf{y}_n \overset{\text{i.i.d.}}{\sim} F_1$. Here, τ is the change-point. When there is a mean difference, we use Δ to denote the Euclidean distance of the means of F_0 and F_1. When there is a variance difference, to make the change less significant, only the first $[d/5]$ of the diagonal elements of the covariance matrix differ with a multiple of σ, and the rest are unchanged.

For the proposed methods, we also expand our study to denser graphs, the k-MST, which is the union of the 1st, $\ldots$, kth MSTs, where the 1st MST is the MST and the ith MST ($i > 1$) is the spanning tree with the sum of the distances of the edges in the tree minimized subject to the constraint that it does not use any of the edge in the 1st, $\ldots$, $(i-1)$th MST(s). Simulation studies show that the edge-count two-sample tests have higher power when the graph is slightly denser as

TABLE 5
Multivariate Gaussian data, mean difference, τ at center

d	10	50	100	150	175	500	2000
Δ	0.8	1	1.2	1.6	2	2.5	3.4
HT	**91 (83)**	**82 (72)**	**72 (60)**	65 51	38 (26)	–	–
GLR	22 (9)	4 (0)	–	–	–	–	–
Z	51 (46)	50 (44)	50 **(46)**	**84** (80)	**91 (88)**	**91 (91)**	87 **(84)**
Z_w	39 (28)	45 (31)	52 (32)	78 (66)	89 (79)	**91 (85)**	**88** (78)
S	32 (21)	33 (23)	37 (23)	68 (55)	80 (69)	84 (80)	81 (71)
M	35 (26)	36 (26)	41 (25)	74 (63)	86 (76)	87 (82)	86 (75)

it contains more similarity information. However, the optimal choice of k is still an open question. Chen and Friedman (2017) recommend to use 5-MST for the generalized edge-count two-sample test. In the following simulation settings, for simplicity, we set the graph to be the 5-MST constructed using Euclidean distance.

The performance of six methods are compared: two methods based on normal theory ($\max \mathrm{HT}(t)$, $\max \mathrm{GLR}(t)$), the method in Chen and Zhang (2015) ($\max Z(t)$), and three new tests ($\max S(t)$, $\max R_w(t)$, $\max M(t)$). The estimated power is calculated as the number of trials, out of 100, that the null hypothesis is rejected at 0.05 level for each of these methods, with p-values determined by 10,000 permutation runs for fairness in comparison. To examine the accuracy of the estimated change-point, the number of trials where the estimated change-point is within 20 from the true change-point is provided in parentheses. Under each setting, the specific alternative is chosen so that the tests have moderate power to be comparable. The best one for each scenario is made bold. 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.

Tables 5–8 show results for multivariate Gaussian data under various alternatives. When there is a mean change only (Tables 5 and 6), we see that in general HT outperforms all other methods in low to moderate dimensions. As dimension becomes larger, the graph-based tests take over. When the location change occurs in the middle of the sequence, the scan statistic $Z(t)$ from Chen and Zhang (2015) outperforms all other tests as dimension increases (Table 5) and the advantage of $Z_w(t)$ becomes evident (Table 6).

Results for scale change only can be found in Table E.1 in Supplement E [Chu and Chen (2019)]. Under this setting, when dimension is low GLR dominates in power. But starting at $d = 20$, the graph-based methods exceed GLR in power and $S(t)$ and $M(t)$ have much higher power among the graph-based methods. More details can be found in Supplement E [Chu and Chen (2019)].

TABLE 6
Multivariate Gaussian data, mean difference, τ at three quarters

d	10	50	100	150	175	500	2000
Δ	0.8	1	1.2	1.6	2	2.5	3.4
HT	**75 (63)**	**70 (66)**	**48 (38)**	40 (28)	34 (30)	–	–
GLR	16 (8)	12 (8)	–	–	–	–	–
Z	25 (15)	14 (5)	17 (7)	17 (6)	42 (12)	37 (14)	30 (11)
Z_w	29 (23)	25 (18)	31 (20)	**52 (42)**	**63 (55)**	**67 (55)**	**68 (62)**
S	25 (16)	17 (10)	25 (17)	35 (29)	50 (46)	49 (39)	48 (44)
M	25 (21)	20 (15)	29 (17)	41 (32)	53 (48)	62 (51)	58 (53)

When there is both location and scale change (Tables 7 and 8), we see that when dimension is low, the parametric-based scan statistics dominate in power. As dimension increases, the new graph-based methods exceed $Z(t)$ and the parametric methods in power. Depending on the size of the change, the best graph-based method is different. Generally, $M(t)$ seems to be most effective in detecting and estimating change-points for high dimension compared to the other graph-based test statistics.

The overall pattern of the power tables show that when d increases the graph-based statistics dominate the parametric tests. The new graph-based methods perform well under various scenarios. In general, $Z_w(t)$ dominates under the alterna-

TABLE 7
Multivariate Gaussian data, mean and scale difference, τ at center

d	10	50	100	150	175	500	2000
Δ	0.6	1	1.2	1.2	1.05	1	1
σ	1.3	1.3	1.1	1.1	1.1	1.1	1.05
HT	**49 (34)**	73 (60)	**65 (53)**	30 (16)	15 (5)	–	–
GLR	26 (17)	12 (0)	–	–	–	–	–
Z	38 (28)	**79** (66)	62 (52)	47 **(34)**	39 (29)	55 (33)	54 (18)
Z_w	30 (14)	62 (55)	55 (43)	38 (28)	29 (21)	15 (4)	18 (5)
S	30 (17)	**79 (70)**	49 (37)	**48** (30)	**44 (36)**	66 (42)	69 (44)
M	29 (12)	76 (67)	52 (40)	51 (29)	43 (34)	**69 (50)**	**74 (51)**

tive of location change away from the center of the sequence whereas $M(t)$ and $S(t)$ dominate under the alternatives of change not only in location. Even under the scenario that is well suited for the method in Chen and Zhang (2015) (location change at the center of the sequence), the new graph-based methods perform at a comparable level to the old method. Based on these results, if one is certain that the change is locational, the test based on $Z_w(t)$ is recommended; while for more general changes, the tests based on $S(t)$ and $M(t)$ are recommended.

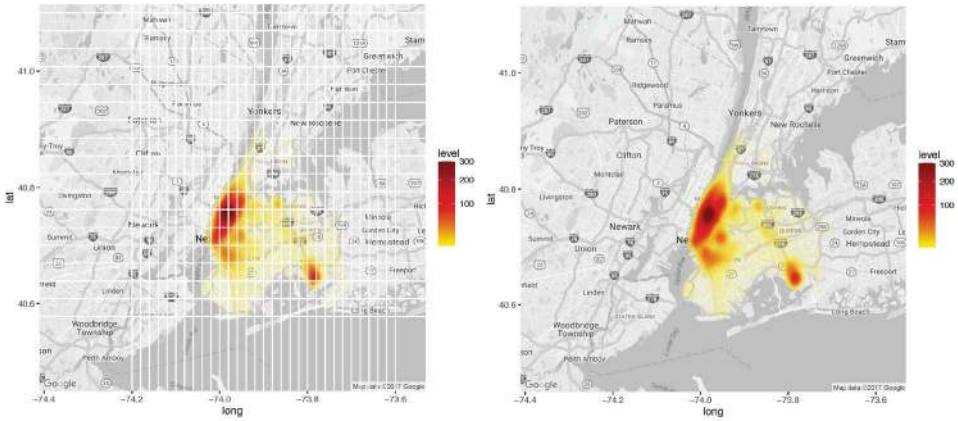
6. A real data example. We illustrate the new approaches on the yellow taxi trip records, which is publicly available on the NYC Taxi and Limousine Commission (TLC) website (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml). The trip records give information on the taxi pickup and drop-off date/times, longitude and latitude coordinates of pickup and drop-off locations, trip distances, fares, rate types, payments types and driver-reported passenger counts.

This dataset is very rich and many questions can be posed. Here, we illustrate the new approach in detecting changes in travel from the John F. Kennedy International Airport for the months October through December of 2015. For simplicity, the boundary of JFK airport was set to be 40.63 to 40.66 latitude and -73.80 to -73.77 longitude.

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 8
Multivariate Gaussian data, mean and scale difference, τ at three quarters

d	10	50	100	150	175	500	2000
Δ	**0.6**	**1**	**1.2**	**1.1**	**1.05**	**0.9**	**0.95**
σ	**1.3**	**1.15**	**0.9**	**0.85**	**0.85**	**0.8**	**0.6**
HT	**37 (28)**	**63 (53)**	**43 (36)**	11 (4)	9 (3)	–	–
GLR	17 (8)	8 (5)	–	–	–	–	–
Z	37 (23)	34 (20)	16 (0)	21 (0)	18 (0)	13 (0)	15 (0)
Z_w	23 (16)	21 (12)	34 (29)	36 (22)	34 (23)	15 (9)	4 (1)
S	25 (18)	22 (12)	36 (29)	44 (34)	**56 (45)**	54 (48)	**57** (52)
M	23 (15)	19 (10)	34 (27)	**48 (38)**	53 (41)	**58 (52)**	**57 (54)**



(a) October 15, 2015 (b) November 20, 2015

FIG. 6. *Density heatmap of taxi drop-offs for four randomly selected days.*

The NYC taxi dataset is immense in both size and information. To better visualize the dataset, we plot heatmaps of the frequency of taxi drop-offs for a small area of New York City that cover the drop-off locations. Figure 6, provides an illustration of the 30 by 30 grid we construct and two randomly selected days in our 3-month period: October 15 and November 20 for visualization. Heatmaps of additional days can be found in Supplement F [Chu and Chen (2019)]. The overall patterns are similar, but a more careful examination reveals there are some differences. To test whether differences are just by randomness or there is a significant change, we apply the three new approaches together with the method in Chen and Zhang (2015).

Let A_i be the 30 by 30 matrix on day i. We denote 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 MST graph representing similarity between days.

For the period of October 1 through December 31, the edge-count statistic $Z(t_1, t_2)$ reports 11/21/15–12/31/15 (Day 52–92) as the changed interval result. However, the new approaches all report the week right before Christmas, 12/18/15–12/25/15 (Day 79–86), as the changed interval (Table 9). 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 through December 17. During this time period, $Z(t_1, t_2)$ selects 10/27/15–12/17/15 (Day 27–78) as the changed interval. The new test statistics all report the week right before Thanksgiving, 11/20/15–11/27/15 (Day 51–58), as the changed interval. All these tests reject the null hypothesis of no change as well, with p-value < 0.001.

We further continued this process by performing the test on the period October 1 through November 20. The original edge-count test $Z(t_1, t_2)$ reports a changed

TABLE 9
Changed interval results and corresponding p-values (reported in parentheses) for NYC taxi pickups from JFK

Time period	Z	Z_w	S	M
10/1–12/31	11/21–12/31 (< 0.001)	12/18–12/25 (< 0.001)	12/18–12/25 (< 0.001)	12/18–12/25 (< 0.001)
10/1–12/17	10/27–12/17 (0.0011)	11/20–11/27 (< 0.001)	11/20–11/27 (< 0.001)	11/20–11/27 (< 0.001)
10/1–11/20	10/22–11/19 (0.0017)	11/16–11/19 (0.0414)	11/16–11/19 (0.0109)	11/16–11/19 (0.0428)

interval from 10/22/15–11/19/15 (Day 22–50). It rejects the null hypothesis of no change as well, with a small p-value (0.0017). All three new tests report a changed interval of 11/16/15–11/19/15 (Day 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 right before Thanksgiving and the week right before Christmas. To perform a more sanity check, we plot the distance matrix of this whole period (Figure 7, left panel). It is evident that there is some change occurring around Day 60 and Day 80, matching with the results from the new tests. On the other hand, the distance matrix for the first 51 days seems much more uniform (Figure 7, right panel).

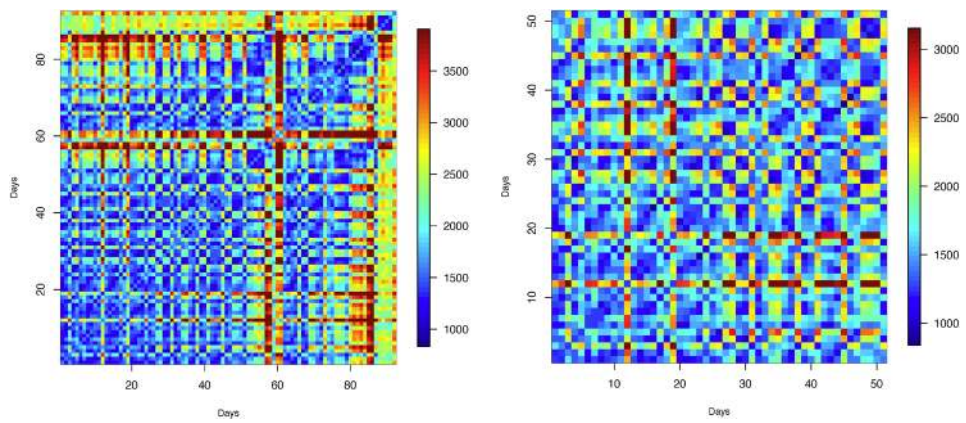


FIG. 7. *Left panel*: *Heatmap of* L_1 *norm distance matrix of vector* v_i *for* $i = 1, \ldots, 92$, *corresponding to dates October* 1, 2015–*December* 31, 2015. *Right panel*: *Heatmap of* L_1 *norm distance matrix of vector* v_i *for* $i = 1, \ldots, 51$, *corresponding to dates October* 1, 2015–*November* 20, 2015.

7. Discussion and conclusion. We propose new graph-based scan statistics for the testing and estimation of change-points that improve upon the framework proposed by Chen and Zhang (2015). Under various common scenarios, the new tests have improved power to detect changes and produce more precise estimates of the location of change-points.

The new scan statistics are based on two basic processes, $Z_w(t)$ and $Z_{\text{diff}}(t)$, with the former sensitive to locational alternatives and the latter sensitive to scale alternatives. These two basic processes rescaled by the length of the sequence—$\{Z_w([nu]) : 0 < u < 1\}$ and $\{Z_{\text{diff}}([nu]) : 0 < u < 1\}$—converge to independent Gaussian processes in finite dimensional distributions under some mild conditions of the graph. The covariance functions of the limiting Gaussian processes do not depend on the graph, so the limiting processes are not affected by the distribution of the observations.

Analytic p-value approximations based on limiting distributions (asymptotic p-value approximation) are derived for all new statistics and the skewness-corrected versions are derived for the weighted edge-count statistic and the max-type edge-count statistic. The asymptotic p-value approximations provides a ballpark estimate of the p-value. The skewness-corrected versions give more accurate approximations. Based on simulation studies, even when the conditions for the graph in deriving the limiting distribution were violated, the analytic p-value formulas still give reasonable approximations. A more detailed discussion on the conditions is in Supplement G [Chu and Chen (2019)].

The performance of the new tests are examined under a number of settings. Simulation results show that the weighted edge-count statistic is extremely useful when the change is locational and the change-point not close to the center of the sequence. When the change in the variance of the distribution is also of interest, the generalized edge-count statistic and the max-type edge-count statistic are recommended. Together with the fact that the skewness-corrected p-value approximations can be easily obtained for the max-type edge-count statistic, the test based on $M(t)$ is preferred to use.

When the independence assumption is violated, instead of using the permutation null, we could do block permutation, that is, the sequence is divided into blocks of size b and the blocks are permuted. In this way, the local structure in the sequence is retained. All these test statistics can be modified accordingly to account for local dependence. The detailed information is in Supplement I [Chu and Chen (2019)].

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data" (DOI: 10.1214/18-AOS1691SUPP; .pdf). The supplementary material contains the new test statistics for the changed-interval alternative, additional technical results and proofs, more illustrations of

the data, additional power and analytical critical value tables and further discussion on the conditions of the graph and the relationship between the new statistics, including an extension of the max-type statistic.

REFERENCES

CARLSTEIN, E., MÜLLER, H.-G. and SIEGMUND, D., eds. (1994). *Change-Point Problems*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23**. IMS, Hayward, CA. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992. MR1477909

CHEN, H., CHEN, X. and SU, Y. (2017). A weighted edge-count two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112**. To appear. DOI:10.1080/01621459.2017.1307757.

CHEN, H. and FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112** 397–409. MR3646580

CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis*: *With Applications to Genetics*, *Medicine*, *and Finance*, 2nd ed. Birkhäuser/Springer, New York. MR3025631

CHEN, L. H. and SHAO, Q.-M. (1994). *Stein's Method for Normal Approximation*. In *An Introduction to Stein's Method*. *Lecture Notes Series* **4** 1–59. World Scientific, Singapore.

CHEN, H. and ZHANG, N. (2015). Graph-based change-point detection. *Ann. Statist.* **43** 139–176. MR3285603

CHU, L. and CHEN, H. (2019). Supplement to "Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data." DOI:10.1214/18-AOS1691SUPP.

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, Chichester. MR2743035

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237

DESOBRY, F., DAVY, M. and DONCARLI, C. (2005). An online kernel change detection algorithm. *IEEE Trans. Signal Process.* **53** 2961–2974. MR2169647

FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. MR0532236

HEARD, N. A., WESTON, D. J., PLATANIOTI, K. and HAND, D. J. (2010). Bayesian anomaly detection methods for social networks. *Ann. Appl. Stat.* **4** 645–662. MR2758643

JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. MR3405600

KOSSINETS, G. and WATTS, D. J. (2006). Empirical analysis of an evolving social network. *Science* **311** 88–90. MR2192483

LUNG-YUT-FONG, A., LÉVY-LEDUC, C. and CAPPÉ, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *J. SFdS* **156** 133–162. MR3436651

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. MR3180567

PARK, Y., WANG, H., NÖBAUER, T., VAZIRI, A. and PRIEBE, C. E. (2015). Anomaly detection on whole-brain functional imaging of neuronal activity using graph scan statistics. In *ACM Conference on Knowledge Discovery and Data Mining* (*KDD*), *Workshop on Outlier Definition*, *Detection*, *and Description* (*ODDx*3).

SIEGMUND, D. and YAKIR, B. (2007). *The Statistics of Gene Mapping*. *Statistics for Biology and Health*. Springer, New York. MR2301277

WANG, H., TANG, M., PARK, Y. and PRIEBE, C. E. (2014). Locality statistics for anomaly detection in time series of graphs. *IEEE Trans. Signal Process.* **62** 703–717. MR3160307

XIE, Y.and SIEGMUND, D. (2013). Sequential multi-sensor change-point detection. *Ann. Statist.* **41** 670–692. MR3023983

ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645. MR2672488

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
ONE SHIELDS AVENUE
DAVIS, CALIFORNIA 95616
USA
E-MAIL: lbchu@ucdavis.edu
hxchen@ucdavis.edu