The Annals of Statistics
2015, Vol. 43, No. 1, 139–176
DOI: 10.1214/14-AOS1269

GRAPH-BASED CHANGE-POINT DETECTION

BY HAO CHEN AND NANCY ZHANG

University of California, Davis and University of Pennsylvania

We consider the testing and estimation of change-points—locations where the distribution abruptly changes—in a data sequence. A new approach, based on scan statistics utilizing graphs representing the similarity between observations, is proposed. The graph-based approach is nonparametric, and can be applied to any data set as long as an informative similarity measure on the sample space can be defined. Accurate analytic approximations to the significance of graph-based scan statistics for both the single change-point and the changed interval alternatives are provided. Simulations reveal that the new approach has better power than existing approaches when the dimension of the data is moderate to high. The new approach is illustrated on two applications: The determination of authorship of a classic novel, and the detection of change in a network over time.

1. Introduction. Change-point models are widely used in various fields for detecting lack of homogeneity in a sequence of observations. In the typical formulation, the observations $\{y_i : i = 1, 2, \ldots, n\}$ are assumed to have distribution F_0 for $i \leq \tau$ and possibly a different distribution F_1 for $i > \tau$. The parameter τ is referred to as the change-point. We consider the case where the total length of the sequence n is fixed. There is a rich literature on theory and applications of this model when y_i are real or integer valued scalars. For example, in a well-known study of the annual flow volume of the Nile River at the city of Aswan, Egypt, from 1871 to 1970, each y_i is a continuous measurement of the annual discharge from the river [Cobb (1978)], and the goal is to detect shifts in flow volume. If the distribution of y_i were assumed to be normal, score- or likelihood-based tests can be applied [James, James and Siegmund (1987)]. Bayesian and nonparametric approaches have also been developed [see Carlstein, Müller and Siegmund (1994) for a survey].

Modern statistical applications are faced with data of increasing richness and dimension. High throughput measurement schemes and digitization in many scientific fields have produced data sequences $\{\mathbf{y}_i : i = 1, 2, \ldots, n\}$, where each $\mathbf{y}_i$ is a high dimensional vector or even a non-Euclidean data object. The dimension of each observation can be larger than the length of the sequence. Testing the homogeneity of such high dimensional sequences is a challenging but important problem. Following are some motivating examples.

Received April 2014.
MSC2010 subject classifications. 62G32.
Key words and phrases. Change-point, graph-based tests, nonparametrics, scan statistic, tail probability, high-dimensional data, complex data, network data, non-Euclidean data.

Network evolution: Data on networks have become increasingly common. For example, email, phone and online chat records can be used to construct a network of social interactions among individuals [Kossinets and Watts (2006), Eagle, Pentland and Lazer (2009)]. High throughput biological experiments have led to the ubiquitous study of protein- or gene-interaction networks. A large part of these studies is characterizing how the network evolves through time. Here, the observation at each time point is a graphical encoding of the network. In a longitudinal study, one might ask whether there is an abrupt shift in network connectivity at any point in time.

Image analysis: Image data collected through time appears in diverse applications, from video surveillance to climatology to neuroscience. The detection of abrupt events, such as security breaches, storms or brain activity, can be formulated as a change-point problem. Here, the observation at each time point is the digital encoding of an image.

Text or sequence analysis: Many classic works in both western and eastern literature have ongoing authorship debates. For example, the debate surrounding both *Tirant lo Blanc*, a Catalan romance, and *Dream of the Red Chamber*, a Chinese masterpiece, is whether there is a change of authorship midway through the novel. In the digital era, an objective approach to these debates is to statistically test for abrupt changes in writing style, which can be reflected by word usage. Similar problems arise in genomic sequence analysis in biology, where it is often of interest to find regions of the genome with different DNA-word compositions [see, e.g., Tsirigos and Rigoutsos (2005)]. In both settings, each observation in the sequence is a vector of word counts over a large dictionary of words.

In all of these examples, the problem can be given the following statistical formulation: We observe a sequence of observations $\{\mathbf{y}_i\}, i = 1, \ldots, n$, indexed by some meaningful ordering, such as time or location. 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

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

or the changed interval alternative

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

where F_0 and F_1 are two probability measures that differ on a set of nonzero measure. Scenarios with multiple change-points can be decomposed into these two types of simple alternatives.

We study this change-point problem under the assumption that $\{\mathbf{y}_i\}$ are *independent*. Independence is an ideal assumption that may be violated in some settings.

However, this assumption allows us to conduct theoretical analysis, which also produce results that are useful when the assumption is slightly violated. We later discuss modifications to our approach when the independence assumption is violated.

In the multivariate setting, existing approaches are limited in many ways. Most methods are based on parametric models that are highly context specific. For example, Zhang et al. (2010) and Siegmund, Yakir and Zhang (2011) studied the problem of detecting common shifts in mean in sequences of independent multivariate Gaussian variables with identity covariance. Under the same setting, Srivastava and Worsley (1986) and James, James and Siegmund (1992) discussed general likelihood ratio tests for a change in mean, which requires that the dimension of the observations be smaller than the number of observations. As we will show in simulations, parametric change-point tests for multivariate data work under very specific assumptions, and are sensitive to violation of these assumptions. The existing parametric tests also cannot be applied in very high dimensions, unless strong assumptions are made to avoid the estimation of the large number of nuisance parameters that are a by-product of increasing dimension.

In the nonparametric context, Desobry, Davy and Doncarli (2005) and Harchaoui, Moulines and Bach (2009) used kernel-based methods. A common drawback for kernel-based methods is that they rely heavily on the choice of the kernel function and its parameters, and the problem becomes more severe when the data is in moderate to high dimensions. Also, none of these methods offer a fast analytical formula for false positive control, thus making them difficult to apply for large data sets. Lung-Yut-Fong, Lévy-Leduc and Cappé (2011) proposed a nonparametric approach based on marginal rank statistics, which is useful if there is a clear ranking mechanism, but also requires the restriction that the number of observations be larger than the dimension of the data.

In this paper, we describe a nonparametric approach to change-point detection and estimation. The approach can be applied to data in arbitrary dimension and even to non-Euclidean data, with a general, analytic formula for type I error control. We illustrate the approach on two applications: Testing for a change in author of a classic European novel, and testing the temporal homogeneity of a social network. We show, via simulations, that as dimension increases this nonparametric method gains power over parametric methods in cases where the parametric methods can be applied. The generality of the new approach and the availability of analytic formulas for type I error make it an easy off-the-shelf tool for homogeneity testing in multivariate settings. The method is implemented in an R package "gSeg," which is available in CRAN.

This paper is organized as follows: In Section 2, we describe the proposed method. The underlying idea is graph-based two-sample tests adapted to the scan-statistic setting. Two-sample tests based on various types of graphs representing the similarity between observations were first proposed in Friedman and Rafsky (1979) and Rosenbaum (2005). We review these previous works in Section 2.1.

Once the graph has been constructed, theoretical analysis of the scan statistic can be decoupled from the modeling of the high dimensional data. We describe the test statistic in the detection of a single change-point in Section 2.2, and that in the detection of a changed interval in Section 2.3. Section 3 gives analytic formulas for approximating the significance of the tests, and evaluates their accuracy in numerical studies. Section 4 evaluates the power of the test via simulations. In Section 5, the new method is applied to the analysis of the text of *Tirant lo Blanc*, and the analysis of the Friendship Network data set collected by the MIT Media Laboratory [Eagle, Pentland and Lazer (2009)]. In Section 6, we discuss some extensions to the approach to deal with local dependency in the sequence and to construct a confidence interval to the change-point. Finally, we conclude with a discussion in Section 7.

2. A graph-based framework for change-point detection. In both the single change-point (1.2) and the changed interval alternatives (1.3), the observations are partitioned into two groups. We allow each group to have a minimum number of observations: $1 < n_0 \le \tau \le n_1 < n$ for the single change-point scenario and $1 < l_0 \le \tau_2 - \tau_1 \le l_1 < n$ for the changed interval scenario, where n_0, n_1, l_0, l_1 are prespecified. Sometimes, these values can be better chosen using domain knowledge. We may also have some further constrains on the locations of τ_1 and τ_2.

We do not impose any restrictions on the sample space or distribution of $\mathbf{y}_i$. Our approach requires that the similarity between $\mathbf{y}_i$ can be represented by a graph, with edges in the graph connecting observations that are "close" in some sense. For the proposed method to have good power, data points drawn from F_0 need to be closer to each other than to data points drawn from F_1, in a global sense, and vice versa. We describe this in more detail next, and briefly review graph-based two-sample tests.

2.1. *Graph-based two-sample tests.* By graph-based tests, we refer to tests that are based on graphs with the observations $\{\mathbf{y}_i\}$ as nodes. The graph is usually derived from a distance or a generalized dissimilarity on the sample space, with edges connecting observations that are close in distance. For example, Friedman and Rafsky (1979) proposed the first graph-based test for testing the null hypothesis that subjects from two groups are equal in distribution against an omnibus alternative. Their method relies on the minimum spanning tree (MST), which is a tree connecting all observations minimizing the total distance across edges. Their test statistic is the number of edges in the tree connecting observations from different groups, rejecting the null hypothesis when this count is low compared to its distribution under permutation. The rationale is that, if the two groups come from different distributions, data points from the same group should be closer to each other, and thus edges in the tree should be more likely to connect subjects within a group.

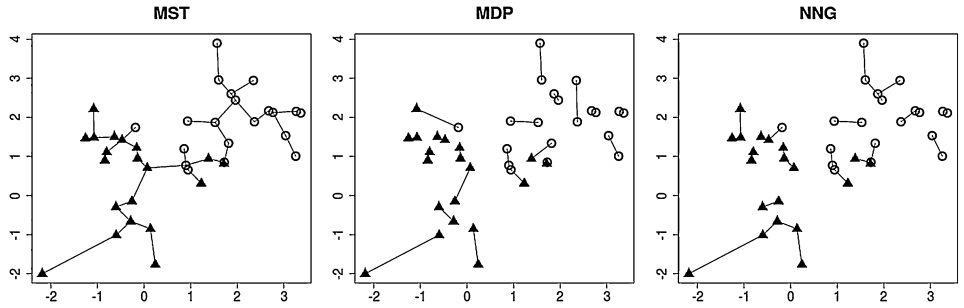


FIG. 1. *The MST, MDP and NNG graphs on an example two-dimensional data set.* 20 *points were drawn from* $\mathcal{N}(\mathbf{0}, I_2)$ (*shown in triangles*) *and* 20 *points were drawn from* $\mathcal{N}((2,2)', I_2)$ (*shown in circles*).

There are many other ways to construct the graph. Rosenbaum (2005) proposed minimum distance pairing (MDP), which divides the n subjects into $n/2$ (assuming n is even) nonoverlapping pairs in such a way as to minimize the total of $n/2$ distances between pairs. For odd N, Rosenbaum suggested creating a pseudo data point that has distance 0 with all other subjects, and later discarding the pair containing this pseudo point. This method has the desirable property of being truly distribution-free.

The nearest neighbor graph (NNG), which connects each data point to its nearest neighbor, can also be used to define a statistic in similar style to Friedman and Rafsky (1979) and Rosenbaum (2005).

Figure 1 illustrates the MST, MDP and NNG on 40 points in $\mathbb{R}^2$. Ways to construct the graph are not limited to these three. In some applications, the graph may be given at the start of the analysis without alluding to an underlying distance measure; see the Haplotype example in Chen and Zhang (2013). The proposed method does not depend on how the graph was constructed. The test statistic and its properties under the permutation null rely only on the graph and not on the underlying distance measure nor on the original data. However, the quality of the graph in separating F_0 and F_1 is integral to the power of the test.

2.2. *Test statistic for a single change-point alternative.* Here, we derive the test statistic for testing the null H_0 (1.1) versus the single change-point alternative H_1 (1.2). Each possible value of τ divides the observations into two groups: Observations come before τ and observations that come after τ. Let G be the similarity graph on $\{\mathbf{y}_i\}$, as described in Section 2.1. We use G to refer to both the graph and its set of edges when the vertex set is implicitly obvious. For any event x let I_x be the indicator function that takes value 1 if x is true, and 0 otherwise. Then, for any candidate value t of τ, the number of edges connecting points from different groups is

$$R_G(t) = \sum_{(i,j)\in G} I_{g_i(t)\neq g_j(t)}, \qquad g_i(t) = I_{i>t}.$$

Here, $g_i(t)$ is an indicator function for the event that $\mathbf{y}_i$ is observed after t. So $R_G(t)$ is the number of edges in the graph G that connect observations from the "past" ($\leq t$) to the "future" ($> t$). Relatively small values of $R_G(t)$ are evidence against the null hypothesis.

Figure 2 illustrates the computation of $R_G(t)$ on a small artificial data set of length $n = 40$ with the first 20 points drawn from $\mathcal{N}(\mathbf{0}, I_2)$ and the second 20

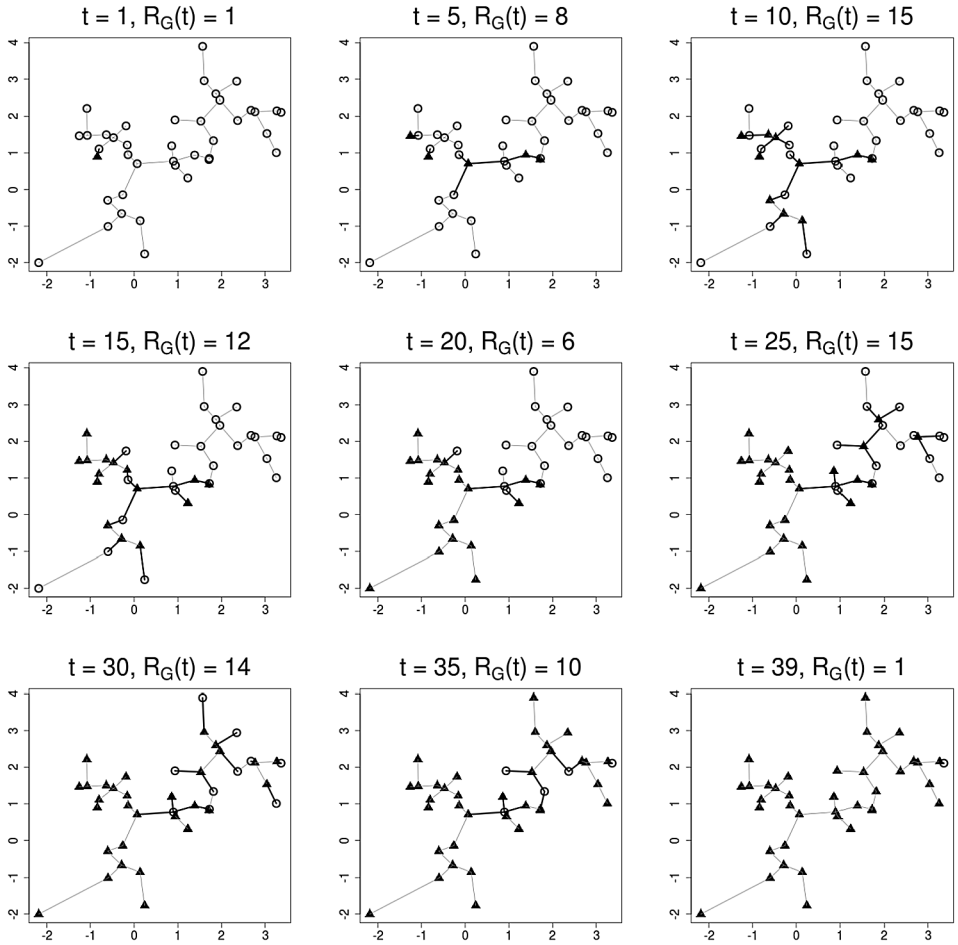


FIG. 2. *The computation of $R_G(t)$ for nine different values of t. The data is a sequence of length $n = 40$, with the first 20 points drawn from $\mathcal{N}(\mathbf{0}, I_2)$ and the second 20 points drawn from $\mathcal{N}((2, 2)', I_2)$. The similarity graph G shown in the plots is the MST on Euclidean distance. Each t divides the observations into two groups, one group for observations before and at t (shown as triangles) and the other group for observations after t (shown as circles). Edges that connect observations from the two different groups (i.e., edges connecting a triangle and a circle) are bold in the graph. Notice that G does not change as t changes, but the group identities of some observations change, causing $R_G(t)$ to change.*

points drawn from $\mathcal{N}((2,2)', I_2)$. The similarity graph G is the MST constructed using Euclidean distance.

Under the null hypothesis H_0 (1.1) and the independence assumption, the joint distribution of $\{\mathbf{y}_i : i = 1, \ldots, n\}$ is the same under the permutation distribution. We define the null distribution of $R_G(t)$ to be the permutation distribution, which places $1/n!$ probability on each of the $n!$ permutations of $\{\mathbf{y}_i : i = 1, \ldots, n\}$. Let $\pi(i)$ be the time of observing $\mathbf{y}_i$ after permutation, then for the permuted sequence, $g_i(t)$ becomes $I_{\pi(i)>t}$. Notice that the graph G is determined by the values of $\mathbf{y}_i$'s, not their order of appearance, and thus remains constant under permutation. When there is no further specification, we denote by $\mathbf{P}$, $\mathbf{E}$, $\mathbf{Var}$ probability, expectation and variance, respectively, under the permutation null distribution.

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

$$Z_G(t) = -\frac{R_G(t) - \mathbf{E}[R_G(t)]}{\sqrt{\mathbf{Var}[R_G(t)]}}. \tag{2.1}$$

In the standardization, we also invert the sign, so that *large* values of $Z_G(t)$ are evidence against the null.

Lemma 2.1 below gives analytic formulas for $\mathbf{E}[R_G(t)]$ and $\mathbf{Var}[R_G(t)]$. Before we state the lemma, we introduce some new notation: Let G_i be the subgraph of G containing all edges that connect to node $\mathbf{y}_i$. As before, we recycle notation and use G_i to denote the set of edges in G_i. $|G_i|$ denotes the number of edges in G_i, which is apparently also the degree of node $\mathbf{y}_i$ in G.

LEMMA 2.1. *Under the permutation null, the expectation and variance of* $R_G(t)$ *are*

$$\mathbf{E}\big(R_G(t)\big) = p_1(t)|G|,$$

$$\mathbf{Var}\big(R_G(t)\big) = p_2(t)|G| + \left(\frac{1}{2}p_1(t) - p_2(t)\right)\sum_i |G_i|^2 + \big(p_2(t) - p_1^2(t)\big)|G|^2,$$

where

$$p_1(t) = \frac{2t(n-t)}{n(n-1)}, \qquad p_2(t) = \frac{4t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)}.$$

The expressions for the expectation and variance are obtained by combinatorial analysis and the details are in Supplement A.1 [Chen and Zhang (2014)].

REMARK 2.2. The expectation and variance of $R_G(t)$ under the permutation null depend only on t, n and two characteristics of the graph—the number of edges ($|G|$) and the sum of squares of node degrees ($\sum_{i=1}^n |G_i|^2$).

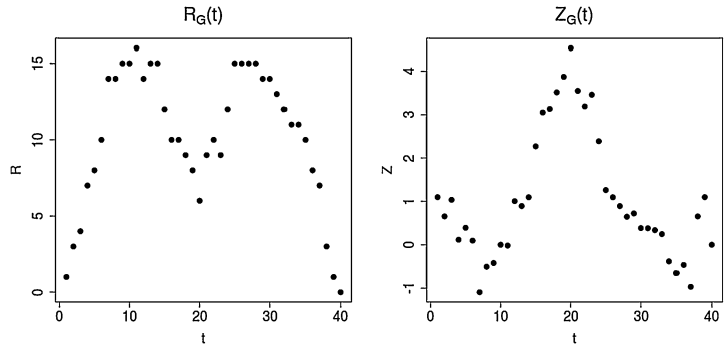


FIG. 3. *The profile of $R_G(t)$ and $Z_G(t)$ against t for the same data set as in Figure 2. There is a change-point at $t = 20$.*

Figure 3 shows the $R_G(t)$ and $Z_G(t)$ processes for the same illustration data set in Figure 2. We see that $Z_G(t)$ peaks at the true change-point 20. For contrast, Figure 4 shows $R_G(t)$ and $Z_G(t)$ for a sequence of 40 points all drawn from $\mathcal{N}(\mathbf{0}, I_2)$. Note that for the latter data set, with no change-point, $Z_G(t)$ exhibits random fluctuation and attains a maximum value much smaller than that of Figure 3.

To test H_0 versus H_a, we use the scan statistic

$$\max_{n_0 \le t \le n_1} Z_G(t), \tag{2.2}$$

where n_0 and n_1 are prespecified constraints for the range of τ as described earlier. The null hypothesis is rejected if the maxima is greater than some threshold. Section 3 describes how to choose the threshold to control the family wise error rate.

2.3. *Test statistic for a changed interval alternative.* Next, we derive the test statistic for testing H_0 (1.1) versus the changed interval alternative H_2 (1.3). Simi-

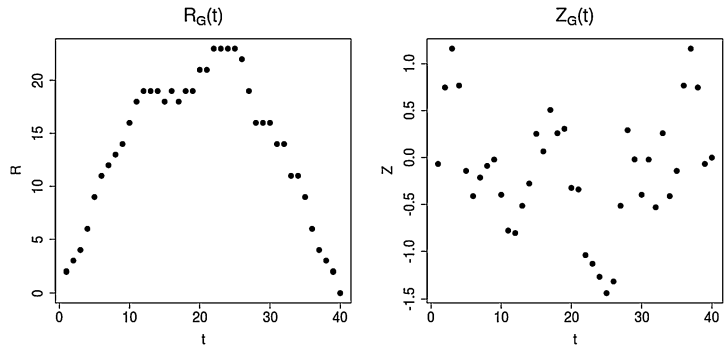


FIG. 4. *The profile of $R_G(t)$ and $Z_G(t)$ against t on a sequence of points all randomly drawn from $\mathcal{N}(\mathbf{0}, I_2)$. There is no change-point in the sequence.*

lar to the single change-point case, any specific alternative (t_1, t_2) divides the data into two groups, one group containing all points observed during $(t_1, t_2]$, and the other group containing all points observed outside of this interval. Then the number of edges in G connecting data points from different groups is

$$R_G(t_1, t_2) = \sum_{(i,j)\in G} I_{g_i(t_1,t_2)\neq g_j(t_1,t_2)}, \qquad g_i(t_1, t_2) = I_{t_1<i\leq t_2}.$$

We standardize $R_G(t_1, t_2)$ as before,

$$Z_G(t_1, t_2) = -\frac{R_G(t_1, t_2) - \mathbf{E}(R_G(t_1, t_2))}{\sqrt{\mathbf{Var}(R_G(t_1, t_2))}}.$$

Lemma 2.3 below gives explicit expressions for the expectation and variance of $R_G(t_1, t_2)$ under the permutation null. The scan statistic involves a maximization over t_1 and t_2,

$$\max_{\substack{1\leq t_1<t_2\leq n\\ l_0\leq t_2-t_1\leq l_1}} Z_G(t_1, t_2), \tag{2.3}$$

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 number of observations in either group is larger than l_0 are considered.

We can further constrain t_1 and t_2 to prefixed sets based on domain knowledge. If we do so, the p-value approximations in Section 3.2 will have minor but obvious modifications, which can be followed straightforwardly by steps given in Section 3.2.

LEMMA 2.3. *Under the permutation null, the expectation and variance of* $R_G(t_1, t_2)$ *are*

$$\mathbf{E}\big(R_G(t_1, t_2)\big) = p_1(t_2 - t_1)|G|,$$

$$\mathbf{Var}\big(R_G(t_1, t_2)\big) = p_2(t_2 - t_1)|G| + \left(\frac{1}{2}p_1(t_2 - t_1) - p_2(t_2 - t_1)\right)\sum_i |G_i|^2$$

$$+ \big(p_2(t_2 - t_1) - p_1^2(t_2 - t_1)\big)|G|^2,$$

where $p_1(\cdot)$ *and* $p_2(\cdot)$ *are defined in Lemma* (2.1).

The proof for this lemma is very similar to the proof of Lemma 2.1 and is omitted here.

3. Analytic approximations to significance levels. How large do the values of the scan statistics (2.2) and (2.3) need to be to constitute sufficient evidence against the null hypothesis of homogeneity? In other words, we are concerned with the tail distribution of the scan statistics under H_0, that is,

$$\mathbf{P}\Big(\max_{n_0\leq t\leq n_1} Z_G(t) > b\Big) \tag{3.1}$$

for the single change-point alternative, and

$$\mathbf{P}\Big(\max_{\substack{1\le t_1<t_2\le n\\ l_0\le t_2-t_1\le l_1}} Z_G(t_1,t_2) > b\Big) \tag{3.2}$$

for the changed interval alternative. In the rest of the paper, we omit the implicitly obvious constraint $1 \le t_1 < t_2 \le n$ for simplicity.

The null distributions of $\max Z_G(t)$ and $\max Z_G(t_1,t_2)$ are defined as the permutation distribution. For small n, we can directly sample from the permutation distribution to approximate (3.1) and (3.2). However, when n is large, permutation is computationally prohibitive, especially for (3.2) where each scan is of order $\mathcal{O}(n^2)$ if $l_1 - l_0 \sim \mathcal{O}(n)$. Therefore, we derive analytic expressions for both tail probabilities to make the method instantly applicable. Treating $\{Z_G(t)\}$ and $\{Z_G(t_1,t_2)\}$ as families of tests, the two probabilities are their family-wise error rates. The tests are dependent since they are all based on the same sequence. The marginal distributions of $Z_G(t)$ and $Z_G(t_1,t_2)$, under permutation, are also quite complicated. Therefore, it is impossible to obtain exact expressions for the two probabilities for finite n. In the rest of this chapter, we give analytic approximations to the two probabilities. We first show that, under mild conditions on G, $\{Z_G([nu]^{1}) : 0 < u < 1\}$ converges to a Gaussian process and $\{Z_G([nu],[nv]) : 0 < u < v < 1\}$ converges to a Gaussian random field as $n \to \infty$ (Section 3.1). We then derive analytic approximations to the two probabilities under Gaussian field approximation (Section 3.2). To achieve better accuracy for the case of small n and for the case where the conditions for Gaussian convergence are questionable, we refine our approximations by correcting the skewness in the marginal distributions (Section 3.3). All of these approximations are checked by numerical studies under a set of representative scenarios (Section 3.5).

3.1. *Asymptotic properties of the processes.* In this section, we derive the limiting distributions of $\{Z_G([nu]) : 0 < u < 1\}$ and $\{Z_G([nu],[nv]) : 0 < u < v < 1\}$ under permutation. 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

$$A_e = G_{e_-} \cup G_{e_+}, \tag{3.3}$$

be the set of edges that connect to either node e_- or node e_+, and

$$B_e = \cup\{A_{e'} : e' \in A_e\}, \tag{3.4}$$

be the set of edges that connect to nodes in G_{e_-} and G_{e_+}.

We define two asymptotic conditions on the graph.

CONDITION 1. $|G| \sim \mathcal{O}(n^{\alpha})$, $0 < \alpha < 1.125$.

[1] $[x]$ is the largest integer that is no larger than x.

CONDITION 2. $\sum_{e\in G}|A_e||B_e| \sim o(n^{1.5(\alpha\wedge 1)})$.

THEOREM 3.1. *Under Conditions 1 and 2, as* $n \to \infty$,

1. $\{Z_G([nu]) : 0 < u < 1\}$ *converges to a Gaussian process, which we denote as* $\{Z_G^\star(u) : 0 < u < 1\}$,
2. $\{Z_G([nu], [nv]) : 0 < u < v < 1\}$ *converges to a two-dimensional Gaussian random field, which we denote as* $\{Z_G^\star(u, v) : 0 < u < v < 1\}$,

under the permutation distribution.

The proof for this theorem utilizes the Stein's method Chen and Shao (2005). The whole proof is in Supplement A.2 [Chen and Zhang (2014)].

REMARK 3.2. Condition 2 restricts both the size and number of hubs, which are nodes with a large degree. The largest hub in the graph must have degree smaller than $n^{0.75(\alpha\wedge 1)}$ to satisfy the condition. On the other hand, if we increase the number of edges in the graph (increase α), the densest graph we could achieve under Condition 2 has the number of edges of order less than $n^{1.125}$. This is because when $|G| \sim \mathcal{O}(n^\alpha)$, $\alpha > 1$, $\sum_{e\in G}|A_e||B_e| \geq n^\alpha n^{\alpha-1} n^{2(\alpha-1)} = n^{4\alpha-3}$.

LEMMA 3.3. *The covariance function of the Gaussian process* $Z_G^\star(u), 0 < u < 1$, *defined as* $\rho_G^\star(u, v) \triangleq \mathbf{cov}(Z_G^\star(u), Z_G^\star(v))$ *has the following expression*:

$$\begin{aligned}\rho_G^\star(u,v) &= \frac{2(u\wedge v)^2(1-(u\vee v))^2|G|}{\sigma_G^\star(u)\sigma_G^\star(v)} \\ &\quad + \frac{(u\wedge v)(1-(u\vee v))(1-2u)(1-2v)\sum_i |G_i|^2}{\sigma_G^\star(u)\sigma_G^\star(v)},\end{aligned} \tag{3.5}$$

where

$$\sigma_G^\star(u) = \sqrt{2u^2(1-u)^2|G| + u(1-u)(1-2u)^2\sum_i |G_i|^2}.$$

The lemma is proved through combinatorial analysis and the details are in Supplement A.3 [Chen and Zhang (2014)].

$\rho_G^\star(u, v)$ is partially differentiable in $u(\neq v)$ to all orders. So, fixing v, the kth order left- and right-derivatives in u at $u = v$ are well defined for all k. We denote the kth left- and right-derivative by $f_{v,-}^{(k)}(0)(\equiv \lim_{u\nearrow v}\frac{\partial\rho_G^\star(u,v)}{\partial u})$ and $f_{v,+}^{(k)}(0)$, respectively. One important property, which can be checked by tedious algebra, is that $f_{v,-}'(0) = -f_{v,+}'(0)$.

3.2. *Asymptotic approximations to p-values.* We now examine the asymptotic behavior of the two probabilities (3.1) and (3.2). Our approximations will involve the function $\nu(x)$ defined as

$$\nu(x) = 2x^{-2}\exp\left\{-2\sum_{m=1}^{\infty} m^{-1}\Phi\left(-\frac{1}{2}xm^{1/2}\right)\right\}, \qquad x > 0. \tag{3.6}$$

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:

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2) - 0.5)}{(x/2)\Phi(x/2) + \phi(x/2)}. \tag{3.7}$$

The following proposition is the foundation for obtaining analytic approximations to the probabilities.

PROPOSITION 3.4. *Assume that* $n_0 \to \infty$, $n_1 \to \infty$, $b \to \infty$, *and* $n \to \infty$ *in a way such that for some* $0 < x_0 < x_1 < 1$ *and* $b_0 > 0$

$$n_i/n \to x_i \qquad (i = 0, 1) \quad \text{and} \quad b/\sqrt{n} \to b_0.$$

Then as $n \to \infty$,

$$P\Big(\max_{n_0 \le t \le n_1} Z_G^\star(t/n) > b\Big) \sim b\phi(b)\int_{x_0}^{x_1} h^\star_{r_0,r_1}(x)\nu\Big(b_0\sqrt{2h^\star_{r_0,r_1}(x)}\Big)\,dx, \tag{3.8}$$

$$\begin{aligned} &P\Big(\max_{n_0 \le t_2 - t_1 \le n_1} Z_G^\star(t_1/n, t_2/n) > b\Big) \\ &\quad\sim b^3\phi(b)\int_{x_0}^{x_1}\Big(h^\star_{r_0,r_1}(x)\nu\Big(b_0\sqrt{2h^\star_{r_0,r_1}(x)}\Big)\Big)^2(1-x)\,dx, \end{aligned} \tag{3.9}$$

where

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

with $r_0 \triangleq \lim_{n\to\infty} |G|/n$, *and* $r_1 \triangleq \lim_{n\to\infty} \sum_i |G_i|^2/|G|$.

The proof of this proposition utilizes Woodroofe's method [Woodroofe (1976, 1978)] and Siegmund's method [Siegmund (1988, 1992)]. The whole proof is in Supplement A.4 [Chen and Zhang (2014)].

REMARK 3.5. Since $n\sum_i |G_i|^2 \ge (\sum_i |G_i|)^2 = 4|G|^2$, $r_1 - 4r_0$ is always nonnegative, and

$$h^\star_{r_0,r_1}(x) \in \left[\frac{1}{2x(1-x)}, \frac{1}{x(1-x)}\right].$$

Based on Proposition 3.4, when $\sum_{e\in G}|A_e||B_e| \sim o(n^{3/2})$, $|G| \sim \mathcal{O}(n)$, we approximate (3.1) and (3.2) by

$$
\begin{aligned}
&P\Big(\max_{n_0\le t\le n_1} Z_G(t) > b\Big)\\
&\quad\sim b\phi(b)\int_{n_0/n}^{n_1/n} h^{\star}_{\hat r_0,\hat r_1}(x)\nu\Big(b_0\sqrt{2h^{\star}_{\hat r_0,\hat r_1}(x)}\Big)\,dx,
\end{aligned}
\tag{3.10}
$$

$$
\begin{aligned}
&P\Big(\max_{n_0\le t_2-t_1\le n_1} Z_G(t_1,t_2) > b\Big)\\
&\quad\sim b^3\phi(b)\int_{n_0/n}^{n_1/n} \Big(h^{\star}_{\hat r_0,\hat r_1}(x)\nu\Big(b_0\sqrt{2h^{\star}_{\hat r_0,\hat r_1}(x)}\Big)\Big)^2(1-x)\,dx,
\end{aligned}
\tag{3.11}
$$

where $\hat r_0 = |G|/n$, $\hat r_1 = \sum_i |G_i|^2/|G|$.

REMARK 3.6. In practice, when using (3.10) and (3.11) to approximate the tail probabilities, we use $h_G(n,x)$ in place of $h^{\star}_{\hat r_0,\hat r_1}(x)$, where $h_G(n,x)$ is the finite-sample equivalent of $h^{\star}_{\hat r_0,\hat r_1}(x)$ for the stochastic process $Z_G([nu])$. That is,

$$
\begin{aligned}
h^{\star}_{\hat r_0,\hat r_1}(x) &= \lim_{u\nearrow x}\frac{\partial\rho^{\star}_G(u,x)}{\partial u},\\
h_G(n,x) &= \frac{1}{n}\lim_{s\nearrow nx}\frac{\partial\rho_G(s,nx)}{\partial s},
\end{aligned}
$$

where $\rho_G(s,t) \triangleq \mathbf{cov}(Z_G(s), Z_G(t))$. The explicit expression for $h_G(n,x)$ is

$$
\begin{aligned}
&h_G(n,x)\\
&\quad= \frac{(n-1)[h_1(n,x)|G| + h_2(n,x)\sum_{i=1}^n |G_i|^2 - h_3(n,x)|G|^2]}{2x(1-x)[h_4(n,x)|G| + h_5(n,x)\sum_{i=1}^n |G_i|^2 - h_6(n,x)|G|^2]},
\end{aligned}
\tag{3.12}
$$

where

$$
\begin{aligned}
h_1(n,x) &= 4n(n-1)\big(-2nx^2+2nx-1\big),\\
h_2(n,x) &= n\big[n(n+1)(1-2x)^2 - 2(n-1)\big],\\
h_3(n,x) &= 4n\big[n(1-2x)^2 - 1\big],\\
h_4(n,x) &= 4n(n-1)(nx-1)(n-nx-1),\\
h_5(n,x) &= n(n-1)\big[n^2(1-2x)^2 - n + 2\big],\\
h_6(n,x) &= 4n\big[n^2(1-2x)^2 - 2n\big(1-3x+3x^2\big) + 1\big].
\end{aligned}
$$

It is easy to show that $\lim_{n\to\infty} h_G(n,x) = h^{\star}_{\hat r_0,\hat r_1}(x)$.

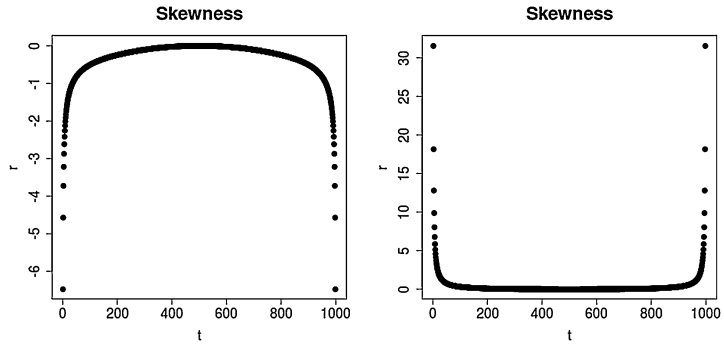


FIG. 5. *Plots of skewness* $\gamma_G(t)$ $(= \mathbf{E}(Z_G(t)))$ *against* t *with* G *being MST* (*left panel*) *and MDP* (*right panel*) *constructed on Euclidean distance on a sequence of* 1000 *points randomly generated from* $\mathcal{N}(\mathbf{0}, I_{100})$.

3.3. *Skewness correction.* Convergence of $Z_G(t)$ to normal is slow if t/n is close to 0 or 1. Also, we may doubt the validity of Conditions 1 and 2 if the graph contains large hubs. For instance, as we show via simulation in Section 3.5 and as detailed in Radovanović, Nanopoulos and Ivanović (2010), MST and NNG constructed on high dimensional data can have large hubs under standard distance measures, such as L_2 and L_1. Then the statistic $Z_G(t)$ is left-skewed (see Figure 5, left panel), and the p-value approximations (3.10) and (3.13) overestimate the tail probabilities. The other extreme is the MDP, where each node has degree 1 and the graph is completely "flat." The statistic $Z_G(t)$ and the two ends is right-skewed (see Figure 5, right panel), and the p-value approximations (3.10) and (3.13) underestimate the true tail probabilities.

Skewness correction in tail probability approximation of change-point tests was first carried out in Tu and Siegmund (1999) and later modified in Tang and Siegmund (2001). Both of these papers applied a universal third moment correction. In our problem, the extent of the skewness of $Z_G(t)$ depends on the value of t. This can be seen clearly in Figure 5, as $Z_G(t)$ is more skewed toward the two ends. Since universal corrections are too crude, we adopt a different approach where the skewness correction adapts to the skewness of $Z_G(t)$ at each t. In particular, we give a better approximation to the marginal probability, $\mathbf{P}(Z_G(t) \in b + dx/b)$ in the single change-point case and $\mathbf{P}(Z_G(t_1, t_2) \in b + dx/b)$ in the changed interval case, for which normal approximation was used in producing the approximations (3.10) and (3.13).

Consider first the approximation of the marginal probability $\mathbf{P}(Z \in b + dx/b)$, suppressing in our notation the dependence on t. Since Z has been properly standardized, $\mathbf{E}(Z) = 0, \mathbf{E}(Z^2) = 1$. Let $\gamma = \mathbf{E}(Z^3)$ be the skewness term, which can be calculated explicitly by a combinatorial analysis described in Section 3.4 below. We make use of the cumulant generating function $\psi(\theta) = \log \mathbf{E}_P(e^{\theta Z})$. By change

of measure $dQ_\theta = e^{\theta Z - \psi(\theta)}\,dP$, we can approximate $\mathbf{P}(Z \in b + dx/b)$ 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

$$\theta_b \approx (-1+\sqrt{1+2\gamma b})/\gamma.$$

More details are given in Supplement B.1 [Chen and Zhang (2014)].

The p-value approximations, after correcting for the skewness of the marginal distribution of the two processes, become

$$\begin{aligned}&\mathbf{P}\Bigl(\max_{n_0\le t\le n_1} Z_G(t) > b\Bigr)\\ &\quad\approx b\phi(b)\int_{n_0/n}^{n_1/n} S_G(nx)h_G(n,x)\nu\Bigl(\sqrt{2b_0^2 h_G(n,x)}\Bigr)\,dx,\end{aligned}\tag{3.13}$$

where

$$S_G(t) = \frac{\exp((1/2)(b-\hat\theta_{b,G}(t))^2 + (1/6)\gamma_G(t)\hat\theta_{b,G}(t)^3)}{\sqrt{1+\gamma_G(t)\hat\theta_{b,G}(t)}},\tag{3.14}$$

with $\gamma_G(t) = \mathbf{E}[Z_G^3(t)]$ and $\hat\theta_{b,G}(t) = (-1+\sqrt{1+2\gamma_G(t)b})/\gamma_G(t)$.

$$\begin{aligned}&\mathbf{P}\Bigl(\max_{n_0\le t_2-t_1\le n_1} Z_G(t_1,t_2) > b\Bigr)\\ &\quad\approx \frac{\phi(b)}{b}\sum_{n_0\le t_2-t_1\le n_1} S_G(t_1,t_2)\bigl(b_0^2 h_G\bigl(n,(t_2-t_1)/n\bigr)\\ &\qquad\qquad\times\nu\bigl(b_0\sqrt{2h_G\bigl(n,(t_2-t_1)/n\bigr)}\bigr)\bigr)^2,\end{aligned}\tag{3.15}$$

where

$$\begin{aligned}&S_G(t_1,t_2)\\ &\quad= \frac{\exp((1/2)(b-\hat\theta_{b,G}(t_1,t_2))^2 + (1/6)\gamma_G(t_1,t_2)\hat\theta_{b,G}(t_1,t_2)^3)}{\sqrt{1+\gamma_G(t_1,t_2)\hat\theta_{b,G}(t_1,t_2)}},\end{aligned}\tag{3.16}$$

with $\gamma_G(t_1,t_2) = \mathbf{E}[Z_G^3(t_1,t_2)]$ and

$$\hat\theta_{b,G}(t_1,t_2) = \bigl(-1+\sqrt{1+2\gamma_G(t_1,t_2)b}\bigr)/\gamma_G(t_1,t_2).$$

REMARK 3.7. When the marginal distribution is highly left-skewed, it is possible that $\gamma(t)$ is too small for $1+2\gamma(t)b$ to be positive. This does not mean that the solution to $\dot\psi_t(\theta) = b$ does not exist, but that higher moments are needed to get a good approximation. In this paper, we apply an easy heuristic fix to this prob-

lem: Since $1+2\gamma(t)b<0$ usually happens when t/n is close to 0 or 1, within this problematic region $\theta_b(t)$ can be extrapolated using its values outside the region. The details of the extrapolation method are given in Supplement B.2 [Chen and Zhang (2014)].

3.4. *Explicit expressions for skewness.* We now derive an explicit expression for the skewness terms $\gamma_G(t)$ and $\gamma_G(t_1,t_2)$ that are used in (3.13) and (3.15). We have

$$\mathbf{E}\big(Z_G^3(t)\big)=\frac{\mathbf{E}^3(R_G(t))+3\mathbf{E}(R_G(t))\,\mathbf{Var}(R_G(t))-\mathbf{E}(R^3(t))}{(\mathbf{Var}(R_G(t)))^{3/2}},$$

$$\mathbf{E}\big(Z_G^3(t_1,t_2)\big)=\frac{\mathbf{E}^3(R_G(t_1,t_2))+3\mathbf{E}(R_G(t_1,t_2))\,\mathbf{Var}(R_G(t_1,t_2))-\mathbf{E}(R^3(t_1,t_2))}{(\mathbf{Var}(R_G(t_1,t_2)))^{3/2}}.$$

The explicit expressions of $\mathbf{E}(R_G(t))$, $\mathbf{Var}(R_G(t))$, $\mathbf{E}(R_G(t_1,t_2))$, and $\mathbf{Var}(R_G(t_1,t_2))$ are given in Lemmas 2.1 and 2.3. The explicit expressions of $\mathbf{E}^3(R_G(t))$ and $\mathbf{E}^3(R_G(t_1,t_2))$ are given in the following lemma.

LEMMA 3.8.

$$\begin{aligned}\mathbf{E}\big(R_G^3(t)\big)&=p_1(t)|G|+\frac{3}{2}p_1(t)\sum_i|G_i|\big(|G_i|-1\big)\\&\quad+3p_2(t)\Big(|G|\big(|G|-1\big)+\frac{1}{2}\sum_i|G_i|\big(|G_i|-1\big)\big(|G|-|G_i|\big)\Big)\\&\quad-3p_2(t)\Big(\sum_i|G_i|\big(|G_i|-1\big)+\sum_{(i,j)\in G}\big(|G_i|-1\big)\big(|G_j|-1\big)\Big)\\&\quad+p_3(t)\sum_i|G_i|\big(|G_i|-1\big)\big(|G_i|-2\big)\\&\quad+p_4(t)\Big(|G|\big(|G|-1\big)\big(|G|-2\big)+6\sum_{(i,j)\in G}\big(|G_i|-1\big)\big(|G_j|-1\big)\Big)\\&\quad-2p_4(t)\sum_{(i,j)\in G}\big|\{k:(i,k),(j,k)\in G\}\big|\\&\quad-p_4(t)\Big(\sum_i|G_i|\big(|G_i|-1\big)\big(3|G|-2|G_i|-2\big)\Big).\end{aligned}$$

The functions $p_1(t)$ and $p_2(t)$ are given in Lemma 2.1, and

$$p_3(t):=\frac{t(n-t)((n-t-1)(n-t-2)+(t-1)(t-2))}{n(n-1)(n-2)(n-3)},$$

$$p_4(t):=\frac{8t(t-1)(t-2)(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)(n-3)(n-4)(n-5)}.$$

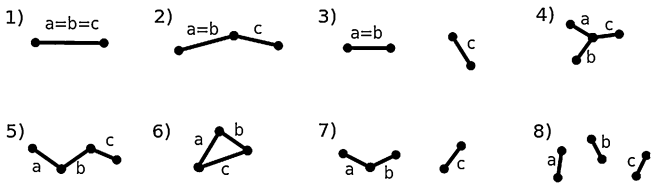


FIG. 6. *Eight configurations of three edges* (a, b, c) *randomly chosen, with replacement, from the graph.*

Also

$$\mathbf{E}^3\big(R_G(t_1, t_2)\big) = \mathbf{E}^3\big(R_G(t_2 - t_1)\big). \tag{3.17}$$

PROOF. For the uncentered process $R_G(t)$,

$$\mathbf{E}\big(R_G^3(t)\big) = \sum_{(i,j),(k,l),(u,v)\in G} \mathbf{P}\big(g_i(t) \neq g_j(t), g_k(t) \neq g_l(t), g_u(t) \neq g_v(t)\big).$$

There are in total eight different configurations for three edges randomly chosen (with replacement) from the graph (see Figure 6 for illustrations). We derive $\mathbf{P}(g_i(t) \neq g_j(t), g_k(t) \neq g_l(t), g_u(t) \neq g_v(t)) \stackrel{\Delta}{=} P_3$ separately for each configuration:

(1) The three edges are actually the same edge

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t)\big) = \frac{2t(n-t)}{n(n-1)}.$$

(2) Two edges are the same and share one node with the third edge

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t), g_i(t) \neq g_k(t)\big) = \frac{t(n-t)}{n(n-1)}.$$

(3) Two edges are the same and do not share any node with the third edge

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t), g_k(t) \neq g_l(t)\big) = \frac{4t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)}.$$

(4) The three edges share one node, and neither of them share the other node (star-shaped)

$$\begin{aligned} P_3 &= \mathbf{P}\big(g_i(t) \neq g_j(t), g_i(t) \neq g_k(t), g_i(t) \neq g_l(t)\big) \\ &= \frac{t(n-t)((n-t-1)(n-t-2) + (t-1)(t-2))}{n(n-1)(n-2)(n-3)}. \end{aligned}$$

(5) One edge share one node with another edge and share the other node with the third edge. No node sharing between the second and the third edge (linear

chain)

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t), g_i(t) \neq g_k(t), g_j(t) \neq g_l(t)\big) = \frac{2t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)}.$$

(6) The three edges form a triangle

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t), g_j(t) \neq g_k(t), g_k(t) \neq g_i(t)\big) = 0.$$

(7) Two edges share one node, and share no node with the third edge

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t), g_i(t) \neq g_k(t), g_u(t) \neq g_v(t)\big) = \frac{2t(t-1)(n-t)(n-t-1)}{n(n-1)(n-2)(n-3)}.$$

(8) No pair of the three edges share any node

$$P_3 = \mathbf{P}\big(g_i(t) \neq g_j(t), g_k(t) \neq g_l(t), g_u(t) \neq g_v(t)\big) = \frac{8t(t-1)(t-2)(n-t)(n-t-1)(n-t-2)}{n(n-1)(n-2)(n-3)(n-4)(n-5)}.$$

Among all $|G|^3$ possible ways of randomly selecting the three edges, the number of occurrences for each of the configuration are:

(1) $|G|$;
(2) $3\sum_i |G_i|(|G_i|-1)$;
(3) $3|G|(|G|-1) - 3\sum_i |G_i|(|G_i|-1)$;
(4) $\sum_i |G_i|(|G_i|-1)(|G_i|-2)$;
(5) $6\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1) - 6\sum_{(i,j)\in G}|\{k:(i,k),(j,k)\in G\}|$;
(6) $2\sum_{(i,j)\in G}|\{k:(i,k),(j,k)\in G\}|$;
(7) $3\sum_i |G_i|(|G_i|-1)(|G|-|G_i|) + 6\sum_{(i,j)\in G}|\{k:(i,k),(j,k)\in G\}| - 12\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1)$;
(8) $|G|(|G|-1)(|G|-2) + 6\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1) - 2\sum_{(i,j)\in G}|\{k:(i,k),(j,k)\in G\}| - \sum_i |G_i|(|G_i|-1)(3|G|-2|G_i|-2)$.

The lemma follows by summing up all of the probabilities as enumerated above.

It is not hard to observe that the number of occurrences only depends on the sizes of the two groups, so $\mathbf{E}^3(R_G(t_1,t_2)) = \mathbf{E}^3(R_G(t_2-t_1))$. □

The terms in $\mathbf{E}(R_G^3(t))$ can be rearranged and written in other forms. The expansion shown in Lemma 3.8 makes it easier to understand the origin of each term in the context of the proof. If we examine the expression, we would find $\gamma_G(t)$ are fully determined by t, n, $|G|$, $\sum_i |G_i|^2$, $\sum_i |G_i|^3$, $\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1)$ and the number of triangles in G.

For MDP, only configurations (1), (3) and (8) are possible, and the number of occurrences of each case is

(1) $|G| = n$;
(3) $3|G|(|G|-1) = 3n(n-1)$;
(8) $|G|(|G|-1)(|G|-2) = n(n-1)(n-2)$;

and its $\mathbf{E}(R_G^3(t))$ has a much simpler expression:

$$\begin{aligned}\mathbf{E}\big(R_G^3(t)\big) &= p_1(t)n + p_2(t)3n(n-1) + p_4(t)n(n-1)(n-2)\\ &= \big(p_1(t) - 3p_2(t) + 2p_4(t)\big)n + 3\big(p_2(t) - p_4(t)\big)n^2 + p_4(t)n^3.\end{aligned}$$

3.5. *Numerical studies*. In this section, we check the analytic approximations to p-values, both assuming Gaussianity and after skewness correction, through numerical studies. We examine both the accuracy of the critical value and the coverage probability.

3.5.1. *Critical value*. We compare the critical values obtained from (3.10), (3.13), (3.11) and (3.15) to those obtained from doing 10,000 permutations, under various simulation settings. In each simulation, i.i.d. sequences of length 1000 were generated from a given distribution F_0 in $\mathbb{R}^d$. MST, MDP and NNG were constructed on the data based on Euclidean distance. For each graph, analytic and permutation critical values were computed for both 0.05 and 0.01 p-value thresholds.

We first check the single change-point alternative. Tables 1–3 show the results for the single change-point alternative with the underlying graph being MST or MDP. Results for when the underlying graph is NNG, shown in Supplement C.1.1 [Chen and Zhang (2014)], are similar to those for when the graph is MST. In the column headers, "A1" denotes critical values obtained assuming Gaussianity (3.10), "A2" denotes critical values obtained after correcting for skewness (3.13), and "Per" denotes critical values obtained by 10,000 permutations (can be viewed as the true p-value).

Six different choices for F_0 are shown, for two different distributions (standard normal and exponential with mean 1), each in three different dimensions ($d =$ 1, 10, or 100). For $d = 10$ or 100, each element of the data vector is generated independently from the given distribution. The analytic approximations depend also on constraints on the region in which the change-point is searched. These are reflected in the choice of n_0 and n_1 (l_0 and l_1 for the changed interval alternative). To make things simple, we set $n_1 = n - n_0$. In general, the analytic approximations become less precise when the minimum segment length decreases. This is mainly because the Gaussian approximation (and skewness correction) to the distribution of $Z(t)$ degrades for small samples.

Both the analytic and permutation p-values depend on certain characteristics of the graph's structure. The structures of MST (for $d \geq 2$) and NNG depend on the underlying data set, and thus the critical values vary by simulation run. In such cases, we show results for 5 randomly simulated sequences. Two characteristics of the graph are also shown for each simulated sequence: The sum of squared node

TABLE 1

Critical values for the single change-point scan statistic based on MST at 0.05 *significance level.* $n = 1000$. *"A1" denotes critical values obtained assuming Gaussianity* (3.10), *"A2" denotes critical values obtained after correcting for skewness* (3.13), *and "Per" denotes critical values obtained by* 10,000 *permutations*

	Critical values										
	$n_0 = 100$			$n_0 = 50$			$n_0 = 25$			**Graph**	
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 1$	2.98	3.05	3.04	3.08	3.22	3.23	3.14	3.39	3.49	4994	2
$d = 10$	2.92	2.90	2.90	3.00	2.95	2.95	3.05	2.98	2.96	5430	8
$N(0, 1)$	2.92	2.89	2.89	3.00	2.95	2.92	3.05	2.97	2.95	5438	7
	2.92	2.90	2.87	3.00	2.95	2.94	3.05	2.98	2.96	5394	7
	2.92	2.89	2.86	3.00	2.94	2.90	3.05	2.97	2.92	5534	8
	2.92	2.89	2.89	3.00	2.95	2.92	3.05	2.97	2.95	5460	7
$d = 10$	2.93	2.91	2.89	3.01	2.97	2.96	3.06	3.00	2.97	5064	7
Exp(1)	2.93	2.91	2.88	3.01	2.97	2.92	3.06	3.00	2.95	5082	7
	2.93	2.91	2.91	3.01	2.98	2.97	3.06	3.01	3.00	5028	5
	2.93	2.91	2.87	3.01	2.98	2.93	3.06	3.01	2.97	5028	6
	2.93	2.91	2.88	3.01	2.96	2.92	3.06	2.98	2.94	5180	9
$d = 100$	2.86	2.69	2.68	2.94	2.70	2.68	3.00	2.70	2.68	12,454	38
$N(0, 1)$	2.86	2.72	2.72	2.95	2.74	2.72	3.00	2.74	2.72	10,904	38
	2.86	2.70	2.66	2.94	2.71	2.66	3.00	2.71	2.66	11,294	42
	2.87	2.72	2.68	2.95	2.74	2.68	3.00	2.74	2.68	10,690	40
	2.86	2.69	2.65	2.94	2.70	2.65	3.00	2.70	2.65	11,722	40
$d = 100$	2.85	2.64	2.60	2.93	2.65	2.60	2.99	2.65	2.60	14,706	56
Exp(1)	2.87	2.77	2.76	2.95	2.80	2.77	3.01	2.81	2.77	9608	25
	2.84	2.62	2.53	2.93	2.62	2.53	2.99	2.62	2.53	15,536	77
	2.86	2.74	2.69	2.95	2.76	2.69	3.00	2.76	2.69	10,890	30
	2.86	2.72	2.66	2.94	2.73	2.66	3.00	2.73	2.66	12,018	39

degrees ($\sum_i |G_i|^2$) and the maximum node degree ($d_{\max}$). These quantities give some intuition on the size and density of hubs in the graph. Since the MST for any one-dimensional data set is a chain, in this case the critical values do not change with simulation run for each setting of the parameters.

The structure of the MDP graph is always the same for all data sets. Therefore, the critical values for MDP-based scan depend only on n, n_0, n_1 (l_0 and l_1 for the changed interval alternative). The critical values for MDP-based scan do not depend on the dimension or the underlying distribution of the data. As emphasized in Rosenbaum (2005), statistics based on the MDP is truly a distribution-free method, which can sometimes be desirable.

We can see from the tables that the analytic approximations after skewness correction perform much better than the analytic approximations under Gaussian assumption, especially when dimension increases. The accuracy of the skew-

TABLE 2
Critical values for the single change-point scan statistic based on MST at 0.01 *significance level.* $n = 1000$

	Critical values										
	$n_0 = 100$			$n_0 = 50$			$n_0 = 25$			**Graph**	
	A1	**A2**	**Per**	**A1**	**A2**	**Per**	**A1**	**A2**	**Per**	$\sum \lvert G_i \rvert^2$	$d_{\max}$
$d = 1$	3.52	3.62	3.67	3.60	3.81	3.85	3.65	4.05	4.31	4994	2
$d = 10$	3.47	3.43	3.46	3.53	3.46	3.48	3.57	3.48	3.48	5430	8
$N(0, 1)$	3.47	3.43	3.44	3.53	3.46	3.46	3.57	3.47	3.46	5438	7
	3.47	3.43	3.44	3.53	3.46	3.47	3.58	3.48	3.48	5394	7
	3.47	3.42	3.38	3.53	3.46	3.40	3.57	3.47	3.41	5534	8
	3.47	3.43	3.44	3.53	3.46	3.46	3.57	3.47	3.46	5460	7
$d = 10$	3.48	3.45	3.40	3.54	3.49	3.44	3.58	3.50	3.45	5064	7
Exp(1)	3.48	3.44	3.40	3.54	3.48	3.42	3.58	3.50	3.44	5082	7
	3.48	3.45	3.47	3.54	3.49	3.49	3.58	3.51	3.52	5028	5
	3.48	3.45	3.41	3.54	3.49	3.44	3.58	3.51	3.46	5028	6
	3.48	3.44	3.49	3.54	3.47	3.53	3.58	3.48	3.54	5180	9
$d = 100$	3.42	3.17	3.19	3.48	3.17	3.19	3.53	3.17	3.19	12,454	38
$N(0, 1)$	3.42	3.21	3.24	3.49	3.21	3.24	3.53	3.21	3.24	10,904	38
	3.42	3.19	3.17	3.49	3.19	3.17	3.53	3.19	3.17	11,294	42
	3.42	3.22	3.18	3.49	3.22	3.18	3.53	3.22	3.18	10,690	40
	3.42	3.18	3.21	3.49	3.18	3.21	3.53	3.18	3.21	11,722	40
$d = 100$	3.41	3.14	3.12	3.48	3.14	3.12	3.52	3.14	3.12	14,706	56
Exp(1)	3.43	3.28	3.26	3.49	3.28	3.26	3.54	3.28	3.26	9608	25
	3.41	3.15	3.10	3.48	3.15	3.10	3.52	3.15	3.10	15,536	77
	3.42	3.24	3.21	3.49	3.24	3.21	3.53	3.24	3.21	10,890	30
	3.42	3.22	3.13	3.48	3.22	3.13	3.53	3.22	3.13	12,018	39

TABLE 3
Critical values for the single change-point scan statistic based on MDP. $n = 1000$

			$d = 1$		$d = 10$		$d = 100$	
n_0	**A1**	**A2**	$N(0, 1)$	**Exp(1)**	$N(0, 1)$	**Exp(1)**	$N(0, 1)$	**Exp(1)**
				Significance level = 0.05				
200	2.82	2.84	2.83	2.81	2.85	2.85	2.85	2.83
100	2.98	3.07	3.06	3.04	3.08	3.08	3.07	3.05
50	3.08	3.27	3.30	3.29	3.35	3.36	3.35	3.31
25	3.14	3.48	3.54	3.58	3.57	3.66	3.60	3.60
				Significance level = 0.01				
200	3.38	3.43	3.39	3.38	3.44	3.46	3.45	3.44
100	3.52	3.66	3.66	3.64	3.67	3.75	3.67	3.59
50	3.60	3.90	3.99	3.99	3.94	4.05	3.95	3.99
25	3.65	4.21	4.61	4.65	4.78	4.72	4.59	4.81

corrected approximation does not degrade significantly with dimension. For the statistics based on MST and NNG, the skew-corrected approximations remain accurate for window sizes as small as 25 at both 0.05 and 0.01 significance levels. For the statistics based on MDP, the skew-corrected approximations work well when the minimum window size is as small as 25 at 0.05 significance level, and 50 at 0.01 significance level.

There is not much difference between results for simulations based on normal and those based on exponential distributions. The main factor influencing approximation accuracy, other than the minimum window size, is the dimension (d). As dimension increases, the graph becomes more "star-shaped" as reflected by the increase in both $\sum |G_i|^2$ and $d_{\max}$. As shown in Section 3.3, skewness and other higher order moments of $Z_G(t)$ are a function of polynomials of the node degrees. Thus, the increase in the number and density of hubs makes skewness correction important in high dimensions. This also indicates that a different distance measure other than Euclidean distance in high dimension to better distinguish different distributions.

For the changed interval alternative, the results are similar, with details in Supplement C.1.2 [Chen and Zhang (2014)].

3.5.2. *Coverage probability.* For both widely used significance levels, 0.05 and 0.01, we also check the coverage probability of the p-value approximations. From the previous section on checking critical values, we see that the underlying distribution of the data does not affect the result, so we generate data only from the multivariate Gaussian distribution. We now expand our study to a denser graphs. In each simulation run, a sequence of length 1000 were generated from $\mathcal{N}(\mathbf{0}, I_d)$. 1,3,5-MST/MDP/NNG were constructed on the data based on Euclidean distance. 1-MST is the same as MST, which we also call the 1st MST. The 2nd MST is defined as a spanning tree that is orthogonal to the 1st MST (not using any edge in the 1st MST) minimizing the total distance over the edges, and the 2-MST is defined as the union of 1st and 2nd MST. Recursively, the kth MST is the spanning tree that is orthogonal to all i MSTs ($i < k$) minimizing the total distance over edges, and the k-MST is defined as the union of all of the ith MSTs, $i = 1, \ldots, k$. Similar definitions apply to k-MDP and k-NNG.

In each simulation run, we calculated the critical value based on the p-value approximation for a given significance level (0.05 or 0.01), and used this critical value as the actual threshold. Then we did 10,000 permutations and calculated the percentage of the permutations with the scan statistic larger than the threshold. This percentage is viewed as the coverage probability. We checked the coverage probability for data in low dimension ($d = 10$) and high dimension ($d = 100$), with 100 simulation runs for each. Figures 7 and 8 show boxplots of the coverage probability for the single change-point alternative with the smallest window size (n_0) being 50. The results for n_0 being 25 or 100, other settings unchanged, are shown

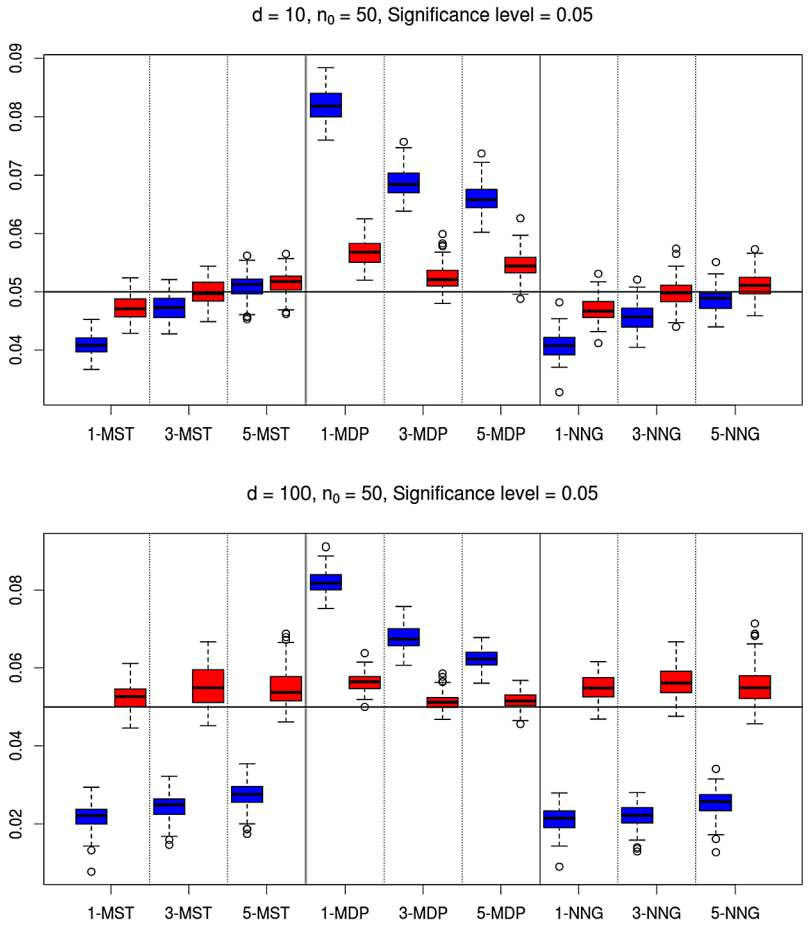


FIG. 7. *Boxplots for coverage probability with significance level* 0.05 *under the single change-point alternative. The smallest windows size is* 50. *The dimension of each observation in the sequence is* 10 *in the upper panel and* 100 *in the lower panel. For each type of graph, the result from the* p*-value approximation assuming Gaussianity is shown in blue and that after skewness correction is shown in red.*

in Supplementary material C.2 [Chen and Zhang (2014)]. The coverage probabilities for the p-value approximation assuming Gaussianity (3.10) are shown in blue and those after skewness correction (3.13) are shown in red. We see that coverage probabilities based on the skewness-corrected p-value approximation are closer to the designed significance level, with the improvement being very significant for MDP in all scenarios and for MST/NNG when the data dimension is high.

Base on results in both the critical values and coverage probabilities, the skew-corrected approximations are quite safe to use.

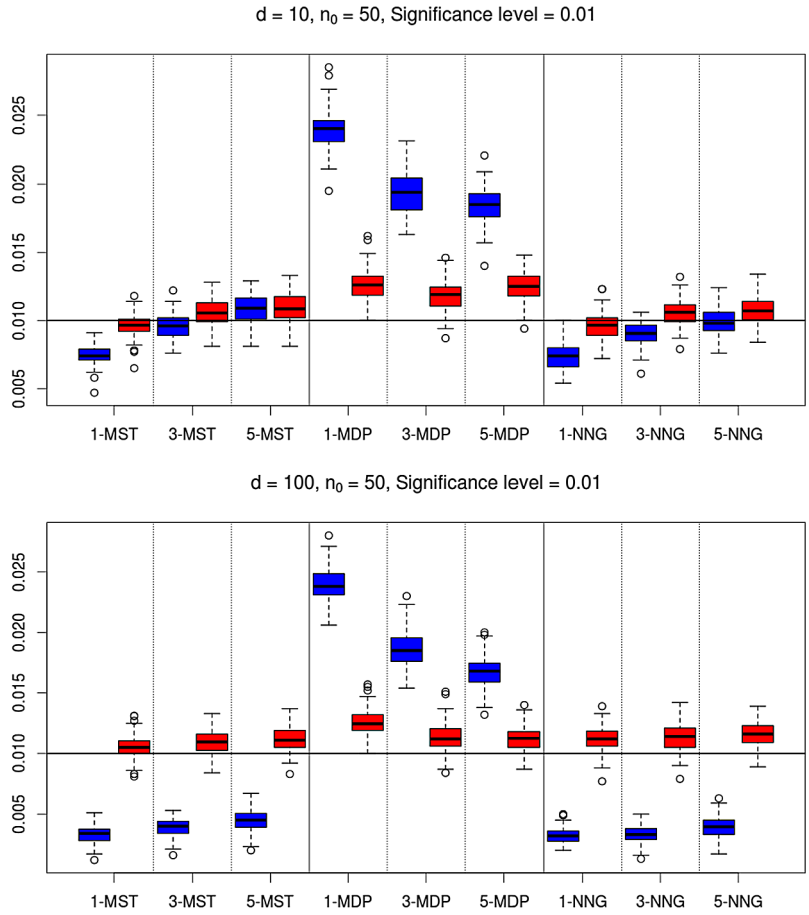


FIG. 8. *Boxplots for coverage probability with significance level* 0.01. *Other parameters remain unchanged from Figure* 7.

4. Power comparisons. To examine the power of our proposed method, we consider cases that parametric methods are applicable. In particular, we consider cases where normal theory can apply. In the first simulation set-up, we generated a sequence of 200 observations from the following model:

$$\mathbf{y}_t \sim \begin{cases} N(\mathbf{0}, I_d), & t = 1, \ldots, 100; \\ N(\boldsymbol{\mu}, \Sigma), & t = 101, \ldots, 200. \end{cases}$$

As before, d is the dimension of each observation. There is a change-point at 100. The mean μ of the second half of the data is shifted from 0 by amount Δ in Euclidean distance. We considered cases where the covariance matrix remains constant ($\Sigma = I_d$), as well as cases where the covariance matrix also changes. When the covariance matrix changes, we set Σ to a diagonal matrix with $\Sigma[1, 1] = d^{1/3}$

and $\Sigma[i,i] = 1$ for $i = 2, \ldots, d$. We chose Δ for each value of d so that most methods have moderate power.

Hotelling's T^2 is a parametric test designed specifically for detecting a change in multivariate normal mean when there is no change in variance. When there is a change in both mean and variance, the generalized likelihood ratio test (GLR) can be used. We compare the graph-based scan statistics to scan statistics based on these two existing methods. For any candidate change-point t, the Hotelling's T^2 is

$$T^2(t) = \frac{t(n-t)}{n}(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}_t^*)^T \widetilde{\Sigma}^{-1}(\bar{\mathbf{y}}_t - \bar{\mathbf{y}}_t^*),$$

where

$$\bar{\mathbf{y}}_t = \sum_{i=1}^{t} \mathbf{y}_i/t, \qquad \bar{\mathbf{y}}_t^* = \sum_{i=t+1}^{n} \mathbf{y}_i/(n-t),$$

$$\widetilde{\Sigma} = (n-2)^{-1}\left[\sum_{i=1}^{t}(\mathbf{y}_i - \bar{\mathbf{y}}_t)(\mathbf{y}_i - \bar{\mathbf{y}}_t)^T + \sum_{i=t+1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}}_t^*)(\mathbf{y}_i - \bar{\mathbf{y}}_t^*)^T\right].$$

The GLR is

$$\mathrm{GLR}(t) = n\log|\hat{\Sigma}_n| - t\log|\hat{\Sigma}_t| - (n-t)\log|\hat{\Sigma}_t^*|,$$

where

$$\hat{\Sigma}_t = \frac{\sum_{i=1}^{t}(\mathbf{y}_i - \bar{\mathbf{y}}_t)(\mathbf{y}_i - \bar{\mathbf{y}}_t)^T}{t}, \qquad \hat{\Sigma}_t^* = \frac{\sum_{i=t+1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}}_t^*)(\mathbf{y}_i - \bar{\mathbf{y}}_t^*)^T}{n-t}.$$

$T^2(t)$ and $\mathrm{GLR}(t)$ both have some constraints on the dimension of the data. For T^2, the number of observations n needs to be larger than the dimension of the data d so that $\widetilde{\Sigma}$ can be inverted. For GLR, both t and $n-t$ need to be larger than the dimension of the data so that the determinants of $\hat{\Sigma}_t$, $\hat{\Sigma}_t^*$ are not zero. Thus, when $d \le 20$, we set $n_0 = d + 10$ and $n_1 = n - n_0$. When $d > 20$, we set $n_0 = 50$ and $n_1 = 150$. (An exception for GLR is that when $d = 50$, n_0 and n_1 are set to 60 and 140, resp., so that the test statistic can be calculated.)

Scan statistics based on the three ways of constructing the graph—MST, MDP and NNG—using Euclidean distance are compared to scan statistics based on maximization of $T^2(t)$ and $\mathrm{GLR}(t)$. We also examined the power of denser graphs: 3-MST, 3-MDP and 3-NNG. The significance level is determined through 10,000 permutation runs (for Hotelling T^2 and GLR) or skew-corrected approximations (for graph-based methods). Table 4 shows the number of trials, out of 100, that the null hypothesis is rejected at 0.05 level for each of these methods. 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 given in parentheses. In the table, bold numbers are cases when the graph-based method outperforms both tests

based on normal theory. In general, they appear when the dimension is relatively high.

First, compare the graph-based methods to Hotelling's T^2: When the variance does not change, T^2 outperforms all other methods in low to moderate dimension ($d < 150$). This is expected, as T^2 was designed specifically for this scenario. Remarkably, graph-based methods surpass T^2 at its own game when dimension is high ($d \geq 150$). If we increase the dimension further, our proposed method is still working while the standard Hotelling's T^2 is no longer applicable in the case where the variance also changes. By assuming an incorrect alternative, the power of T^2 is quickly surpassed by graph-based methods, for d as low as 5.

Comparing graph-based methods to the GLR-based scan statistic, we see a similar pattern: When dimension is low ($d = 1, 5, 10$), GLR-based scans dominate in power when both the mean and variance changes. Graph-based methods exceed GLR in power when d increases, already performing much better by $d = 20$, which is considered quite low in today's applications. The low power of GLR at even moderate dimension is due to its requirement that the covariance matrix be estimated for both segments.

We also considered a case where the normality assumption is violated by generating data from the log-normal distribution ($\Sigma = I_d$). Then graph-based methods outperform T^2 by $d = 10$, and GLR even when $d = 1$ (3-MST and 3-NNG).

Comparing among the graph-based scan statistics, we see that MST and NNG have comparable power, and dominate MDP in all scenarios. An explanation is that, of these three types of graphs, the MDP retains the least information from the data, having half as many edges as the other two graphs. The fact that denser graphs lead to higher power is also evident as we compare the performance of 3-MST/MDP/NNG to the (1-) versions. Also, 3-MST/MDP/NNG have similar power, indicating that power is not sensitive to the method of graph construction, so long as the graph and distance function effectively separates F_0 from F_1.

Another interesting fact on the graph-based tests is that their power mainly depends on the size of the change and not decrease much as the dimension increases. This can be seen clearly in the first table in Table 4. As we increase the change (Δ) from 1.2 ($d = 100$) to 2 ($d = 175$), there is an increasing trend in power for each of the graph-based tests. On the other hand, there is a slightly decreasing trend for the Hotelling's T^2 test. Also, as we jump from $d = 175$ to $d = 500$, we only increase the change a little (2 to 2.5) to have all the graph-based tests remain similar power. These results show that the graph-based tests are powerful in high dimension despite the hubbing phenomenon.

For all scenarios that the null is rejected, we also tally whether the estimated change-points are within [80, 120] to check their accuracy (numbers in parentheses, Table 4). We see that, in terms of the accuracy, the graph-based methods are comparable to, if not better than, that based on normal theory.

TABLE 4

Number of simulated sequences (out of 100) with significance less than 5%, and the numbers in parentheses are those having the estimated change-point within [80, 120]

Normal data, $\Sigma = I$

d Δ	**1** **0.5**	**10** **0.8**	**50** **1**	**100** **1.2**	**125** **1.4**	**150** **1.6**	**175** **2**	**500** **2.5**
T^2	*85* (68)	*97* (83)	*80* (64)	*69* (58)	*69* (58)	*66* (54)	*53* (42)	– –
GLR	*74* (60)	*26* (14)	*12* (0)	– –	– –	– –	– –	– –
1-, 3-MST	*15*, *30* (4, 16)	*20*, *52* (13, 37)	*14*, *42* (11, 37)	*17*, *38* (13, 34)	*27*, *48* (18, 44)	*38*, *65* (33, **59**)	***60***, ***86*** (**54, 85**)	***58, 87*** (**51, 85**)
1-, 3-MDP	*13*, *19* (0, 6)	*16*, *34* (6, 23)	*14*, *29* (7, 18)	*15*, *24* (8, 15)	*30*, *42* (19, 24)	*26*, *48* (19, 37)	*40*, ***77*** (30, **63**)	***49, 72*** (**29, 56**)
1-, 3-NNG	*11*, *28* (3, 17)	*20*, *51* (14, 39)	*18*, *40* (14, 32)	*17*, *32* (12, 28)	*27*, *51* (19, 47)	*32*, ***67*** (27, **61**)	*53*, ***87*** (**49, 85**)	***57, 88*** (**50, 85**)

Normal data, Σ is diagonal with $\Sigma[1,1] = d^{1/3}$, $\Sigma[i,i] = 1, i = 2, \ldots, d$

d Δ	**1** **0.5**	**5** **0.4**	**10** **0.1**	**20** **0.2**
T^2	*78* (60)	*16* (11)	*7* (1)	*7* (1)
GLR	*65* (45)	*80* (70)	*69* (59)	*23* (10)
1-, 3-MST	*14*, *33* (5, 17)	*29*, *52* (15, 30)	*35*, *61* (24, 52)	***62***, ***85*** (**44**, **77**)
1-, 3-MDP	*14*, *17* (1, 6)	*12*, *29* (4, 17)	*17*, *42* (7, 28)	***44***, ***76*** (**24**, **61**)
1-, 3-NNG	*8*, *31* (3, 11)	*28*, *45* (12, 28)	*30*, *64* (18, 51)	***58***, ***86*** (**42**, **74**)

Log-normal data, $\Sigma = I$

d Δ	**1** **0.7**	**5** **0.9**	**10** **1**	**20** **1**	**50** **1.2**	**75** **1.4**	**100** **1.4**
T^2	*78* (57)	*84* (69)	*78* (61)	*54* (38)	*57* (43)	*43* (34)	*28* (20)
GLR	*30* (21)	*15* (9)	*16* (8)	*14* (3)	11 (0)	– –	– –
1-, 3-MST	*24*, *59* (11, 43)	*41*, *67* (29, 54)	*43*, ***89*** (38, **79**)	*33*, ***64*** (26, **57**)	*45*, ***66*** (32, **60**)	***54***, ***84*** (**48, 81**)	***52***, ***75*** (**44, 72**)
1-, 3-MDP	*18*, *28* (5, 17)	*23*, *53* (7, 35)	*24*, *52* (15, 40)	*13*, *32* (3, 19)	*30*, *51* (19, 37)	*23*, ***65*** (19, **50**)	*24*, ***58*** (15, **41**)
1-, 3-NNG	*18*, *51* (10, 33)	*38*, *67* (27, 53)	*32*, *77* (27, **66**)	*27*, ***60*** (20, **52**)	*46*, ***70*** (32, **61**)	***49***, ***85*** (**43**, **82**)	***46***, ***75*** (**38**, **71**)

5. Real data examples. We illustrate the new approach on two different applications. The first is a statistical analysis of the text of *Tirant lo Blanc*. The second is a longitudinal study of a network through time.

5.1. *Authorship debate*. *Tirant lo Blanc*, a chivalry novel published in 1490, is considered to be one of the best known medieval works of literature in Catalan, and is well recognized to be a major influence to *Don Quixote*. For such an important work in western literature, there is a long lasting debate regarding its authorship originated from conflicting information provided in its first published version. The dedicatory letter at the beginning of the book states,

> ...So that no one else can be blamed if any faults are found in this work, I, Joanot Martorell, knight, take sole responsibility for it, as I have carried out the task single-handedly...

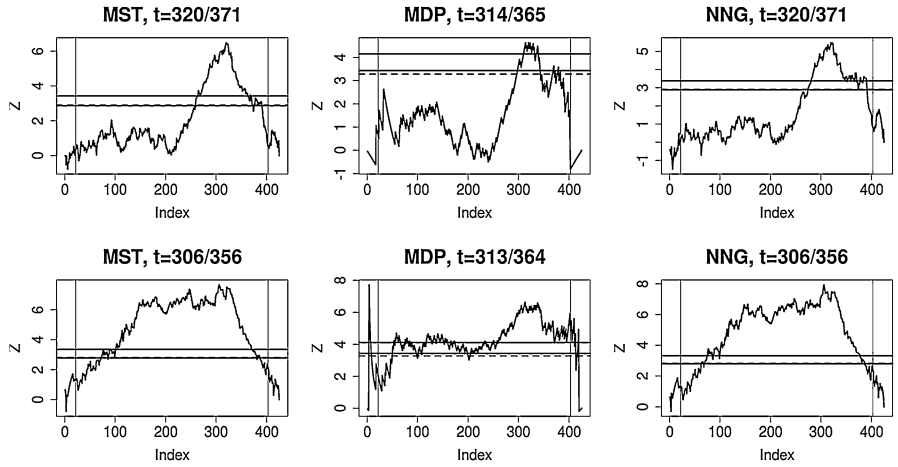
However, the colophon at the end of the book states something different,

> ...by the magnificent and virtuous knight, Sir Joanot Martorell, who because of his death, could only finish writing three parts of it. The fourth part, which is the end of the book, was written by the illustrious knight Sir Marti Joan de Galba. If faults are found in that part, let them be attributed to his ignorance...

This inconsistency sparked a debate, still ongoing, about the authorship of *Tirant lo Blanc* since its publication. Opinions have mainly fallen into two camps, one favoring single authorship by Joanot Martorell and the other favoring a change of author somewhere between Chapters 350 and 400 with 487 chapters in total. One objective way to settle this debate is through the statistical analysis of word usage, which reflects the unique writing style of different people.

Girón, Ginebra and Riba (2005) analyzed two sets of word usage statistics extracted from the book. The first, which we call the word length data set, categorizes the words in each chapter by its length, with a single category for all words with length greater than nine letters. Thus, this data set represents each chapter by a vector of length 10. The second, which we call the context-free word frequency data set, counts the occurrence of the 25 most frequent context-free words in each chapter. Girón, Ginebra and Riba (2005) analyzed the two data sets using a Bayesian multinomial change-point model and a Bayesian clustering method, and concluded in favor of the change of author hypothesis with the estimated change-point between Chapters 371 and 382.

Here, we apply the graph based change-point method to the two data sets, treating each chapter as a time-point. There are in total 487 chapters, and we use the 425 chapters that have more than 200 words. For both data sets, we normalized the count vector for each chapter by dividing the total number of words in the chapter. Thus, our data is a sequence of 425 normalized proportions, of dimension 10 for the word length data and dimension 25 for the context-free word frequency data. The L_2 norm is used to construct the MST, MDP and NNG graphs representing similarity between chapters. $Z_G(t)$ and the estimated change-points, computed for

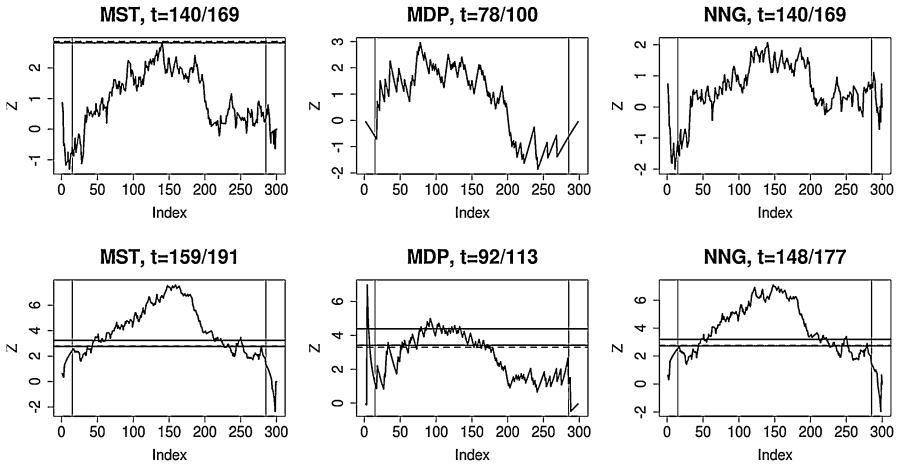


Data	MST	MDP	NNG
Word length	0.0000 (1.5e–9)	0.0042 (0.0018)	0.0000 (7.5e–7)
Context-free word frequency	0.0000 (2.7e–13)	0.0000 (6.1e–6)	0.0000 (3.0e–14)

FIG. 9. *Results of graph-based scans of chapter-wise word usage frequencies of* Tirant lo Blanc, *based on the word length data* (*first row*) *and context-free word frequency data* (*second row*). *The three columns show scans based on three different graphs*: *MST*, *MDP and NNG from left to right. In each plot*, $Z_G(t)$ *is plotted along* t (*chapter*). *The estimated change-point is shown in the caption above the plot in the form* A/B, *where* A *is the index of the change-point within the* 425 *chapters used for analysis*, *and* B *is the chapter number in the novel. The two vertical lines show* n_0 *and* n_1; *we excluded the first* 5% *and the last* 5% *of the points. The horizontal lines show critical values at* 0.05 *and* 0.01 *significance levels*, *with the solid lines showing critical values computed from* 10,000 *permutations and the dashed lines showing those computed from the analytic approximation with skewness correction. The table lists the* p*-values for the tests through* 10,000 *permutations with the skew-corrected approximations in parentheses.*

each type of graph, are shown in Figure 9. Test results using the three different graphs and the two data sets support the change of author hypothesis, with the estimated change-point around Chapter 360, which is consistent with the view that there is a change of author somewhere between Chapters 350 and 400. The p-values are shown in the table in Figure 9.

To check the robustness of our analysis, we also applied the scan on data for the first 350 chapters to see if it rejects the null there. Opinions seem to be quite uniform that the first 350 chapters were all written by Joanot Martorell. The results are shown in Figure 10. The word length data does not reject the null for the 350 chapters at 0.05 significance level. However, the context-free word frequency data supports a change-point, although different graphs favor different locations for the change-point. The p-values of the tests are shown in the table in Figure 10. One explanation is that the context-free word frequency is still affected by the context, and thus less robust than the word length in reflecting writing styles. It



Data	MST	MDP	NNG
Word length	0.0485 (0.0562)	0.1079 (0.1040)	0.3053 (0.3527)
Context-free word frequency	0.0000 (2.9e–13)	0.0018 (0.0009)	0.0000 (1.3e–11)

FIG. 10. *Results from the first* 350 *chapters. The setting of the figure is the same as in Figure* 9. *The table lists the p-values for the tests through* 10,000 *permutations with the skew-corrected approximations in parentheses.*

is also possible that the first author's writing style evolves as he proceeded in the dimension of context-free word frequency.

5.2. *Friendship network.* The MIT Media Laboratory conducted a study following 90 subjects, consisting of students and staff at the university, using mobile phones with preinstalled software recording call logs from July 2004 to June 2005 [Eagle, Pentland and Lazer (2009)]. In this analysis, we extract the information on the caller, callee and time for every call that was made during the study period. The question of interest is whether phone call patterns changed during this time, which may reflect a change in relationship among these subjects. We bin the calls by day and, for each day, construct a network with the 90 subjects as nodes and a link between two subjects if they had at lease one call on that day. We encode the network of each day by an adjacency matrix, with 1 for element $[i, j]$ if there is an edge between subject i and subject j, and 0 otherwise. Thus, the processed data are adjacency matrices, one for each day from 2004/7/20 to 2005/6/14.

We show results for graphs constructed using two different dissimilarity measures. Let A_i be the 90 by 90 adjacency matrix on day i. We denote v_i to be the vector form of A_i. The dissimilarities are:

(1) the number of different edges: $\|v_i - v_j\|_1 = \|v_i - v_j\|_2^2$,

(2) the number of different edges, normalized by the geometric mean of the total for each day: $\frac{\|v_i - v_j\|_1}{\sqrt{\|v_i\|_1 \|v_j\|_1}}$.

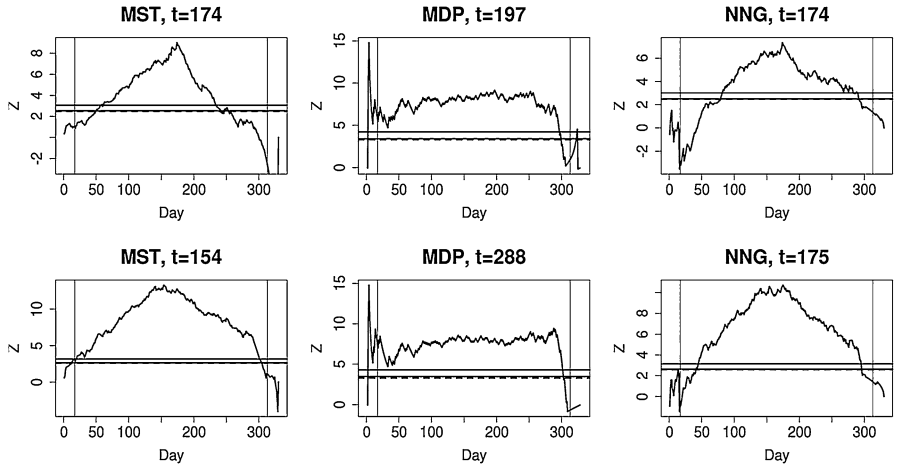


FIG. 11. *Results of graph-based scans of the MIT phone call network. Top row shows results from using number of different edges as the dissimilarity measure and bottom row shows results from using the normalized number of different edges. The three columns show three different ways of constructing the graph*: *MST, MDP and NNG from left to right. The content in each plot is the same as in Figure* 9.

Results based on different dissimilarities and different ways of constructing the graph are shown in Figure 11. We see that statistics based on MST and NNG give similar results under both dissimilarities. Based on the scans using MST and NNG, a change-point occurred at around December 19, 2004 ($t = 154$) or January 9/10, 2005 ($t = 174/175$), which are almost the two ends of the winter break. So these results suggest a change of phone call pattern as they move from the fall quarter to the spring quarter. The statistic based on MDP is quite horizontal for a long range of time. One reason is that for this network data, the change is relatively gradual rather than abrupt and the constructing of MDP then tend to connect observation t to observations $t - 1$ or $t + 1$, which makes the resulting graph not informative in determine the location of a "big" change. The p-values for the scan based on MST and NNG under both dissimilarity measures are all <0.0001, by both 10,000 permutations and skew-corrected approximations.

6. Extensions. In this section, we discuss some extensions to the approach to deal with local dependency in the sequence (Section 6.1) and to construct a confidence interval for the change-point (Section 6.2).

6.1. *Block permutation for local dependency.* In both applications, independence is a useful but idealized assumption for the data. One way to deal with local dependency is to define the null distribution as the distribution under block permutation rather than permutation. In block permutation, the sequence is divided into blocks of size b and the blocks are permuted.[2] The standardized count is then

TABLE 5
p-values from 10,000 *block permutations for the authorship data set*

	Word length			Context-free word frequency		
Block size	**MST**	**MDP**	**NNG**	**MST**	**MDP**	**NNG**
1*	0	0.0042	0	0	0	0
2	0	0.0029	0	0	0	0
5	0	0.0041	0	0	0.0001	0
10	0	0.0057	0	0	0.0006	0

*: a block size of 1 is equivalent to permutation under independence assumption.

defined as

$$Z_{G,\mathrm{bp}}(t) = -\frac{R_G(t) - \mathbf{E}_{\mathrm{bp}}(R_G(t))}{\sqrt{\mathbf{Var}_{\mathrm{bp}}(R_G(t))}}, \tag{6.1}$$

where $\mathbf{E}_{\mathrm{bp}}(R_G(t))$ and $\mathbf{Var}_{\mathrm{bp}}(R_G(t))$ are the expectation and variance for $R_G(t)$ under block permutation.[3] The test statistic is now defined as

$$\max_{n_0 \le t \le n_1} Z_{G,\mathrm{bp}}(t), \tag{6.2}$$

and the p-value for the above statistic can be obtained by block permutation. While analytical formulas for the null moments and the family-wise error rate under the block permutation model are too complicated to be practical, for medium to small data sets these quantities can be obtained by brute force computation.

We used the block permutation model to analyze both the *Tirant lo Blanc* authorship and the friendship network data. The p-values for the authorship data under different block sizes (2, 5, 10) are summarized in Table 5, and plots of the $Z_{G,\mathrm{bp}}$ values are shown in Figure 12 (block size 5) and Supplement D.1.1 [Chen and Zhang (2014)] (block size 2 and 10). Results for the authorship data with the first 350 chapters and the phone call network data are in Supplement D [Chen and Zhang (2014)].

In all cases, block permutation gives the same conclusion as permutation. Block permutation tends to increase the p-value when the block size is large. Simulation studies show that this is the case even when the sequence is made up of independent observations. It is due to the fact that block permutation with large blocks produces a less homogeneously mixed sequence. Despite the slight decrease in significance, the fact that p-values remain in the same regime even under block permutation boosts our confidence in our conclusions for both applications.

[2]There are b ways to divide the sequence into blocks of size b, with the first block of size $1, 2, \ldots, b$. For each block permuted sequence, we first randomly chosen one way from the b ways, and then randomly permute the blocks.

[3]$\mathbf{E}_{\mathrm{bp}}(\cdot)$ and $\mathbf{Var}_{\mathrm{bp}}(\cdot)$ can be calculated by doing, for example, 10,000 block permutations.

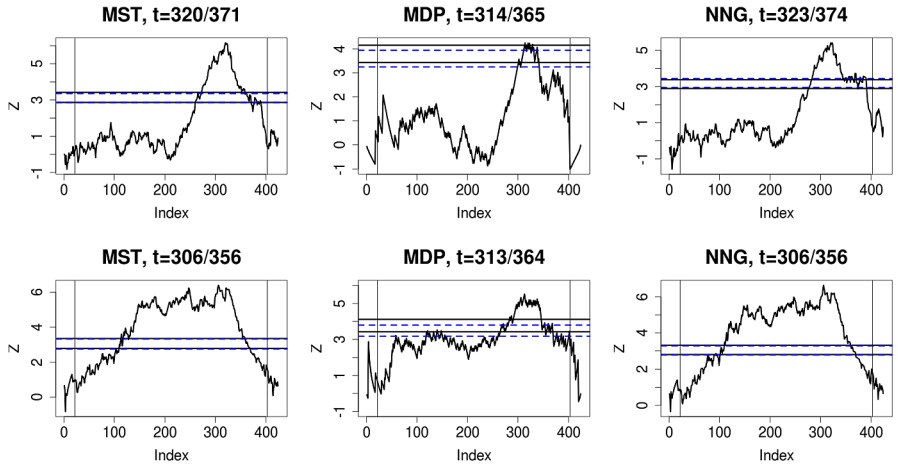


FIG. 12. *Results of graph-based scans of chapter-wise word usage frequencies of* Tiran lo Blanc, *based on the word length data* (*first row*) *and context-free word frequency data* (*second row*), *under block permutation with block size* 5. *The critical values at* 0.05 *and* 0.01 *levels*, *obtained from* 10,000 *block permutation runs are shown in these blue dash lines*. *The solid black lines are critical values from permutation*.

6.2. *Confidence interval for estimated change-point*. Upon rejection of the null one is often concerned with the accuracy of the estimate of the change-point location. To some extent, we explored this in Section 4 by tallying whether the estimated change-point is within a fixed window centered at the true change-point. Here, we describe a procedure for constructing a confidence region for the change-point, which is motivated by the approach studied by Worsley (1986), where it was called a Cox–Spjøtvoll-type confidence region citing the original paper Cox and Spjøtvoll (1982).

The Cox–Spjøtvoll type confidence region is based on the duality relationship where the α level confidence region for a change-point contains all values k that partition the sequence into two subsequences (before and after k), where within the subsequences the hypothesis of homogeneity cannot be rejected at level α. That is, let p_k^L and p_k^R be the p-values for testing the null hypothesis of homogeneity in respectively the left and right subsequences when partitioned at k. A $1-\alpha$ confidence region can be expressed as

$$D_\alpha = \{k : p_k^L, p_k^R \geq 1 - \sqrt{1-\alpha}\}.$$

On the word length data, the 0.01 confidence region $D_{0.01}$ for the chapter where the author changes from Joanot Martorell to Marti Joan de Galba is $\{296\} \cup [298, 355]$. While this region is informative, the break between Chapters 296 and 298 is hard to interpret. Note that for a value $k < \hat{\tau}$ to belong to a $1-\alpha$ level Cox–Spjøtvoll region, both the subsequence to the left of k and the subsequence to the right of k must test negative for a change-point. The subsequence to

the right of k, which contains $\hat{\tau}$, usually tests positive if there are enough points between k and $\hat{\tau}$. In this way, the confidence region has the desirable tendency of including points close to $\hat{\tau}$. The subsequence to the left of k, which does not include $\hat{\tau}$, may test positive for a change-point for two reasons: Inhomogeneity in the left subsequence, for example, existence of another change-point before τ, or a false positive due to random chance. Neither of these reasons seems to have much to do with our precision of estimating τ with $\hat{\tau}$.

For example, 297 is excluded in the confidence region for the change of author in *Tirant lo Blanc* because of what happened in Chapters 1 to 297, not because of what happened in Chapter 298 onward (the right subsequence actually test negative). There is no historical evidence pointing to a third author, and thus we are willing to believe that there is either one author (Joan Martorell) or two authors (Joan Martorell and Marti Joan de Galba). Thus, when we compute our confidence region for the change-point, we are doing so under the premise that there is a single change in author. Hence, not including 297 due to possible inhomogeneity prior to Chapter 297, when our best estimate of the change-point is 320, seems a bit silly. We would much rather include 297, claim to have a conservative interval, and forego the exact coverage property of the Cox–Spjøtvoll region.

Motivated by these considerations, we modify the Cox–Spjøtvoll type confidence region in the following way: If k comes before the estimated change-point ($\hat{\tau}$), we test whether the right-subsequence contains a change-point; and if k comes after $\hat{\tau}$, we test whether the left-subsequence contains a change-point. In other words, let

$$C_{\alpha,L} = \{k < \hat{\tau} : p_k^R \geq 1 - \sqrt{1-\alpha}\},$$
$$C_{\alpha,R} = \{k < \hat{\tau} : p_k^L \geq 1 - \sqrt{1-\alpha}\},$$

our confidence region is $C_\alpha = C_{\alpha,L} \cup C_{\alpha,R} \cup \{\hat{\tau}\}$. Since $C_\alpha \supseteq D_\alpha$, C_α is a conservative α level confidence region. C_α is more likely than D_α to form an interval, and despite its conservativeness it is more accurate in reflecting the precision of $\hat{\tau}$ in estimating τ when we believe τ to be the sole change-point.

This modified procedure is illustrated on the word length data shown in Figure 13. The 0.01 confidence region for the location of change in author is $[281, 355]$, which correspond to original chapter numbers 330 to 409. Comparing to $D_{0.01}$, we deduce that not only 297 but 281–295 were excluded from $D_{0.01}$ due to possible inhomogeneity in the left subsequence. As for any real data, homogeneity is an ideal and not a completely correct assumption for the *Tirant lo Blanc* word length sequence. In reporting the C_α region, we are choosing a region that is more conservative, but in turn, more robust against slight deviations from the model.

7. Conclusions and discussion. The proposed method for change-point detection can be applied to a wide range of data, requiring only the existence of a

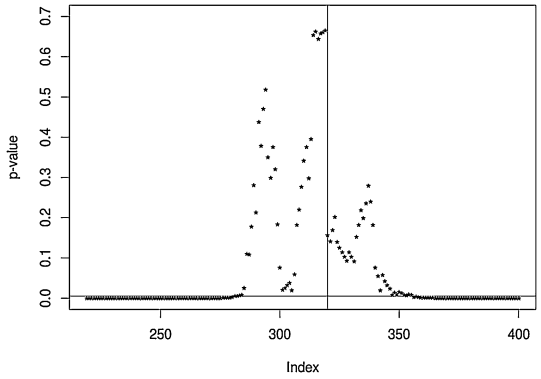


FIG. 13. *Illustration for computing the C_α confidence region for location of change in author in* Tirant lo Blanc *on word length data with G being a MST. The vertical line is the estimated change–point. The x-axis is the indices on the chapters being used and y-axis is the p-value for the right- or left-subsequence (depending on whether the point is resp. before or after $\hat{\tau}$). The horizontal line is at the value $1 - \sqrt{1 - 0.01}$.*

dissimilarity measure on the sample space. In applications, the choice of a good dissimilarity measure is critical, and domain knowledge should be used to design a measure that is sensitive to the signal of interest. The graph-based approach in this paper decouples this modeling choice of dissimilarity measure from the formal test for a change-point. Given the graph, the scan statistics are straightforward to compute, with general off-the-shelf analytic formulas for family-wise error control.

We have shown that the p-value approximations are quite accurate. Our simulations were for a data sequence of length $n = 1000$. The accuracy of the approximations depend on the minimum allowed group size n_0 (l_0 for the changed interval alternative) and not so much on n. Accuracy also depends on the structure of the graph. When the graph is dominated by hubs, skewness correction is critical for the approximations to be accurate. For *extremely* star-shaped graphs, we imagine that adjusting for kurtosis and higher order moments would also be helpful. The strategy would be similar to skewness correction, but more technically complicated. We do not compute these higher order terms in this paper, but if needed they can be computed in a similar fashion as the skewness term with the aid of a symbolic computation software.

If hubs dominate the topology of the graph, perturbation of any hub can change the topology drastically, and $R_G(t)$, which does not take into account the interaction between edges, loses all information regarding the high order structure. Under such circumstances, the particular graph would not be useful for differentiating F_1 from F_0, and one would need to explore other dissimilarity measures and graph construction methods on the data. Radovanović, Nanopoulos and Ivanović (2010) studied the hubbing phenomenon in high dimensional data under several similarity measures, which can serve as a starting point for choosing informative similarity measures for particular problems.

Compared to parametric approaches, the graph-based approach requires far fewer assumptions, but also makes less use of the data. Although this leads to loss of power in low dimensions if the data indeed follow the parametric model, it leads to robustness and wider applicability. An important observation is that the graph-based approach has desirable power, compared to existing parametric tests, in moderate and high dimensions. For high dimensional data, it is often hard to predict the direction and nature of the change. Without such prior knowledge, parametric models would require the estimation of many parameters, most of which would be unrelated to the change. For example, the Hotelling T^2 statistic requires the estimation of the large covariance matrix. If, by prior knowledge or data preprocessing, we can circumvent the covariance estimation, then Hotelling T^2 would be preferable when the data satisfies its assumptions—normality with no change of variance. Otherwise, graph-based approaches gain increasing advantage over Hotelling's T^2 as d increases, even in the problem for which Hotelling's T^2 was explicitly designed.

We explored three different ways of constructing the underlying graph given a dissimilarity measure. From the numerical results and the analysis of the MIT cell phone network, we see that scans based on MST and NNG perform similarly, while scans based on MDP have lower power. We suspect this is due to the fact that MDP is the least dense graph and utilizes the least amount of information from the original data set. This is confirmed as the power increases when we use denser graphs (3-MST/MDP/NNG vs. 1-MST/MDP/NNG). More study is needed to determine what is the optimal choice of graph. One may also consider assigning weights to the edges. As in all problems, building more assumptions into the statistic leads to improved power if the assumptions are true, but sacrifices robustness.

The analytic moment and significance formulas assume independent observations. When there is local dependence, block permutations may be useful in producing more accurate p-values. We illustrated this in Section 6.1. Block permutation is computationally intensive, and in practice one always wrestles with the question of how to choose the block size. When local dependence is weak, as for our data examples, the thresholds given by block permutation are quite close to the analytic thresholds that assume dependence.

A Cox–Spjøtvoll type confidence region, as proposed by Worsley (1986), can be computed under this graph-based framework to assess the uncertainty in the estimation of the change-point. As described in Section 6.2, we find Worsley's approach to be sometimes misleading in practice, and propose a modification that is conservative but more robust. Our discussion focused on the inference for the chapter where authorship changed in *Tirant lo Blanc*, because this seems to be a problem where the space of models is limited, and the interpretation of the change-point parameter is clear.

If more than one change-point or changed interval were of interest, the graph-based scan can be applied recursively in a procedure that is called binary or circular binary segmentation [Olshen et al. (2004), Vostrikova (1981)].

Acknowledgments. We thank David Siegmund, Jerome Friedman and Susan Holmes for helpful discussions. We also thank J. Girón for kindly providing the data for the analysis of Tirant lo Blanc.

SUPPLEMENTARY MATERIAL

Supplement to "Graph-based change-point detection" (DOI: 10.1214/14-AOS1269SUPP; .pdf). *Supplement A*: *Proofs for lemmas, propositions and theorems*. We provide the proofs to the lemmas, propositions and theorems. *Supplement B*: *Skewness correction*. We provide the details to the skewness correction we used. *Supplement C*: *Checking analytic approximations to p-values*. We provide more results in checking analytic approximations to p-values. *Supplement D*: *Block permutation results.* We provide more results on block permutation in analyzing the two real data examples.

REFERENCES

CARLSTEIN, E. G., MÜLLER, H. G. and SIEGMUND, D. (1994). *Change-Point Problems*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23**. IMS, Hayward, CA. MR1477909

CHEN, L. H. Y. and SHAO, Q. M. (2005). Stein's method for normal approximation. *An Introduction to Stein's Method* **4** 1–59.

CHEN, H. and ZHANG, N. R. (2013). Graph-based tests for two-sample comparisons of categorical data. *Statist*. *Sinica* **23** 1479–1503. MR3222245

CHEN, H. and ZHANG, N. (2014). Supplement to "Graph-based change-point detection." DOI:10.1214/14-AOS1269SUPP.

COBB, G. W. (1978). The problem of the Nile: Conditional solution to a changepoint problem. *Biometrika* **65** 243–251. MR0513930

COX, D. R. and SPJØTVOLL, E. (1982). On partitioning means into groups. *Scand*. *J*. *Stat*. **9** 147–152. MR0680910

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

EAGLE, N., PENTLAND, A. S. and LAZER, D. (2009). Inferring friendship network structure by using mobile phone data. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **106** 15274–15278.

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

GIRÓN, J., GINEBRA, J. and RIBA, A. (2005). Bayesian analysis of a multinomial sequence and homogeneity of literary style. *Amer*. *Statist*. **59** 19–30. MR2109428

HARCHAOUI, Z., MOULINES, E. and BACH, F. R. (2009). Kernel change-point analysis. In *Advances in Neural Information Processing Systems*.

JAMES, B., JAMES, K. L. and SIEGMUND, D. (1987). Tests for a change-point. *Biometrika* **74** 71–83. MR0885920

JAMES, B., JAMES, K. L. and SIEGMUND, D. (1992). Asymptotic approximations for likelihood ratio tests and confidence regions for a change-point in the mean of a multivariate normal distribution. *Statist*. *Sinica* **2** 69–90. MR1152298

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. (2011). Homogeneity and change-point detection tests for multivariate data using rank statistics. Preprint. Available at arXiv:1107.1971.

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572.

RADOVANOVIĆ, M., NANOPOULOS, A. and IVANOVIĆ, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *J. Mach. Learn. Res.* **11** 2487–2531. MR2727772

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 515–530. MR2168202

SIEGMUND, D. (1988). Approximate tail probabilities for the maxima of some random fields. *Ann. Probab.* **16** 487–501. MR0929059

SIEGMUND, D. O. (1992). Tail approximations for maxima of random fields. In *Probability Theory* (*Singapore*, 1989) 147–158. de Gruyter, Berlin. MR1188717

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

SIEGMUND, D., YAKIR, B. and ZHANG, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.* **5** 645–668. MR2840169

SRIVASTAVA, M. S. and WORSLEY, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *J. Amer. Statist. Assoc.* **81** 199–204. MR0830581

TANG, H. K. and SIEGMUND, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2** 147–162.

TSIRIGOS, A. and RIGOUTSOS, I. (2005). A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Res.* **33** 922–933.

TU, I.-P. and SIEGMUND, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis. *Adv. in Appl. Probab.* **31** 510–531. MR1724565

VOSTRIKOVA, L. J. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR* **259** 270–274.

WOODROOFE, M. (1976). Frequentist properties of Bayesian sequential tests. *Biometrika* **63** 101–110. MR0415920

WOODROOFE, M. (1978). Large deviations of likelihood ratio statistics with applications to sequential testing. *Ann. Statist.* **6** 72–84. MR0455183

WORSLEY, K. J. (1986). Confidence regions and test for a change-point in a sequence of exponential family random variables. *Biometrika* **73** 91–104. MR0836437

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: hxchen@ucdavis.edu

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: nzh@wharton.upenn.edu