Biometrika (2022), **00**, 0, *pp.* 1–16
Printed in Great Britain
doi: 10.1093/biomet/asab048

Asymptotic distribution-free changepoint detection for data with repeated observations

BY HOSEUNG SONG AND HAO CHEN

Department of Statistics, University of California, Davis, One Shields Avenue, Davis, California 95616, U.S.A.

hosong@ucdavis.edu hxchen@ucdavis.edu

SUMMARY

A nonparametric framework for changepoint detection, based on scan statistics utilizing graphs that represent similarities among observations, is gaining attention owing to its flexibility and good performance for high-dimensional and non-Euclidean data sequences. However, this graph-based framework faces challenges when there are repeated observations in the sequence, which is often the case for discrete data such as network data. In this article we extend the graph-based framework to solve this problem by averaging or taking the union of all possible optimal graphs resulting from repeated observations. We consider both the single-changepoint alternative and the changed-interval alternative, and derive analytical formulas to control the Type I error for the new methods, making them readily applicable to large datasets. The extended methods are illustrated on an application in detecting changes in a sequence of dynamic networks over time. All proposed methods are implemented in an `R` package `gSeg` available on CRAN.

Some key words: Categorical data; Discrete data; High-dimensional data; Non-Euclidean data; Nonparametric framework; Scan statistic; Tail probability.

1. INTRODUCTION

1.1. *Background*

Changepoint analysis plays a significant role in many applications where a sequence of observations is collected. In general, the problem concerns testing whether a change has occurred, or whether several changes might have occurred, and identifying the locations of any such changes. In this paper, we consider the offline changepoint detection problem, where a sequence of independent observations $\{y_i\}_{i=1,\ldots,n}$ is completely observed at the time when data analysis is conducted. Here, n is the length of the sequence and i is the time index or other meaningful index depending on the specific application. We consider testing the null hypothesis

$$H_0 : y_i \sim F_0 \quad (i = 1, \ldots, n) \tag{1}$$

against the single-changepoint alternative

$$H_1 : \exists\, 1 \leqslant \tau \leqslant n \text{ such that } y_i \sim \begin{cases} F_1, & i > \tau, \\ F_0, & \text{otherwise,} \end{cases}$$

or the changed-interval alternative

$$H_2 : \exists\, 1 \leqslant \tau_1 \leqslant \tau_2 \leqslant n \text{ such that } y_i \sim \begin{cases} F_1, & i = \tau_1 + 1, \ldots, \tau_2, \\ F_0, & \text{otherwise}, \end{cases} \tag{2}$$

where F_0 and F_1 are two different distributions.

This problem has been extensively studied for univariate data; see Chen & Gupta (2011) for a survey. However, in many modern applications, y_i is high-dimensional or even non-Euclidean. For high-dimensional data, most methods are based on parametric models; see, for example, Zhang et al. (2010), Wang et al. (2018) and Wang & Samworth (2018). To apply these methods, the data sequence needs to follow specific parametric models. In the nonparametric domain, Harchaoui et al. (2009) employed kernel methods, Lung-Yut-Fong et al. (2012) utilized marginal rankings, and Matteson & James (2014) made use of distances between observations. These methods can be applied to a wider range of problems than parametric methods. However, it is in general difficult to conduct theoretical analysis on nonparametric methods, and none of the aforementioned nonparametric methods provides analytical formulas for false discovery control, making them difficult to apply to large datasets.

1.2. *Graph-based changepoint methods*

Recently, Chen & Zhang (2015) and Chu & Chen (2019) developed a graph-based framework to detect changepoints for high-dimensional and non-Euclidean data sequences. This framework is based on a similarity graph G, such as a minimum spanning tree, MST, i.e., a spanning tree connecting all observations such that the sum of distances of the edges in the tree is minimized, constructed on the sample space. Based on G, test statistics rely on three basic quantities, $R_{0,G}(t)$, $R_{1,G}(t)$ and $R_{2,G}(t)$, where for each t, $R_{0,G}(t)$ is the number of edges connecting observations before and after t, $R_{1,G}(t)$ is the number of edges connecting observations prior to t, and $R_{2,G}(t)$ is the number of edges connecting observations after t. Then, four scan statistics were studied: the original edge-count scan statistic $Z(t)$, the weighted edge-count scan statistic $Z_w(t)$, the generalized edge-count scan statistic $S(t)$ and the max-type edge-count scan statistic $M(t)$, which can be applied to various alternatives. For detailed comparisons, see Chu & Chen (2019).

While the methods proposed by Chen & Zhang (2015) and Chu & Chen (2019) work well for continuous data, they encounter problems when the data contain repeated observations, which is often the case for discrete data, such as network data. The reason is that these methods depend on the similarity graph constructed on observations. When there are repeated observations, the similarity graph is in general not uniquely defined, which leads to trouble in applying the methods. For example, Chen & Zhang (2015) analysed a phone-call network dataset with the aim of testing whether there are any changes in the dynamics of the networks over time. In this dataset, a few networks in the sequence are exactly the same, and Chen & Zhang (2015) used the MST as the similarity graph. More specifically, in the dataset there are a total of 330 networks $\{y_1, \ldots, y_{330}\}$ in the sequence, and among them are 290 distinct networks. For example, y_1, y_6 and y_{16} are exactly the same; all repeated observations are listed in the Supplementary Material. Because of these repeated observations, there are numerous ways of assigning edges in the MST. Hence, the MST is not uniquely defined and existing graph-based methods are not reliable, since they are formulated from the unique similarity graph on pooled observations.

Table 1 lists the test statistics and their corresponding p-values for the four testing procedures proposed in Chen & Zhang (2015) and Chu & Chen (2019) on three randomly chosen MSTs. We see that the p-value depends heavily on the choice of MST: it could be very small for one

Table 1. *The p-values and corresponding test statistics (in parentheses) for the four testing procedures proposed in Chen & Zhang (2015) and Chu & Chen (2019): an original edge-count scan statistic* $\max_{n_0 \leqslant t \leqslant n_1} Z_0(t)$*, a generalized edge-count scan statistic* $\max_{n_0 \leqslant t \leqslant n_1} S(t)$*, a weighted edge-count scan statistic* $\max_{n_0 \leqslant t \leqslant n_1} Z_w(t)$*, and a max-type edge-count scan statistic* $\max_{n_0 \leqslant t \leqslant n_1} M(t)$*; here* n_0 *is set to* $\lceil 0.05n \rceil = 17$ *and* n_1 *to* $330 - n_0$*, where* $\lceil x \rceil$ *denotes the smallest integer greater than or equal to* x

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

Table 2. *Notation at time t*

Index of distinct values	1	2	$\cdots$	K	Total
Group 1	$n_{11}(t)$	$n_{12}(t)$	$\cdots$	$n_{1K}(t)$	t
Group 2	$n_{21}(t)$	$n_{22}(t)$	$\cdots$	$n_{2K}(t)$	$n-t$
Total	m_1	m_2	$\cdots$	m_K	n

MST, but very large for another MST, leading to completely different conclusions on whether the sequence is homogeneous or not. Moreover, since the number of possible MSTs is huge, it is impractical to obtain a summary by directly computing the test statistics on all possible MSTs.

1.3. *Contribution of this work*

We extend the methods of Chen & Zhang (2013) and Zhang & Chen (2019) to the changepoint detection setting, and propose new graph-based testing procedures that can deal with repeated observations effectively. This work fills the gap in the graph-based framework in dealing with discrete data. We show that the new tests are asymptotically distribution-free under the null hypothesis of no change, and find that the limiting distributions for two approaches in Chen & Zhang (2013) are the same, even for continuous data. We also derive analytical formulas to approximate permutation p-values for those modified test statistics, making them readily applicable to real datasets. To improve the analytical p-value approximations for finite sample sizes, skewness correction is also performed. We show that the proposed tests work well in detecting changes when the data contain repeated observations. We illustrate the new testing procedures through analysis of a phone-call network dataset. The proposed methods are implemented in an `R` (R Development Core Team, 2021) package `gSeg` available on CRAN.

2. Notation and related existing work

We represent data with repeated observations using a contingency table for each t, as shown in Table 2. Suppose that there are a total of n observations and K distinct values, which we also refer to as categories in the following. Each t divides the sequence into two groups, before or at time t (Group 1) and after time t (Group 2). Let $n_{ik}(t)$ be the number of observations in group i ($i = 1, 2$) and category k ($k = 1, \ldots, K$), and let m_k ($k = 1, \ldots, K$) be the number of observations in category k. Notice that $m_k = n_{1k}(t) + n_{2k}(t)$ ($k = 1, \ldots, K$), $\sum_{k=1}^{K} m_k = n$, $\sum_{k=1}^{K} n_{1k}(t) = t$ and $\sum_{k=1}^{K} n_{2k}(t) = n - t$.

Chen & Zhang (2013) and Zhang & Chen (2019) studied ways of extending the underlying graph-based two-sample tests to accommodate data with repeated observations under the two-sample testing framework. When the data contain repeated observations, the similarity graph is not uniquely defined according to common optimization criteria, such as the MST, leading to multiple optimal graphs. The authors considered two ways of incorporating information from these graphs: averaging and union. Specifically, they first constructed the similarity graph on the distinct values, denoted by C_0. Here, C_0 could be the MST on all distinct values, the nearest-neighbour link, the union of all possible MSTs on the distinct values when the MST on the distinct values is not unique, or some other meaningful graph. Then, the optimal graph initiated from C_0 is defined in the following way: for each pair of edges $(k_1, k_2) \in C_0$, randomly choose an observation with value indexed by k_1 and an observation with value indexed by k_2, and connect these two observations; then, for each k_i $(i = 1, 2)$, if there is more than one observation with value indexed by k_i, connect these observations by a spanning tree, so that any edges in this spanning tree will have distance 0. More detailed explanations of constructing C_0 and the associated optimal graph are provided in the Supplementary Material. Based on these optimal graphs, the averaging statistic is defined by averaging the test statistic over all optimal graphs, and the union statistic is defined by calculating the test statistic on the union of all optimal graphs.

3. Proposed tests

3.1. *Extended test statistics for data with repeated observations*

Here we focus on extending the weighted, generalized and max-type test statistics for repeated observations, which will turn out to be asymptotic distribution-free tests. Details of extending the original edge-count test are given in the Supplementary Material. Based on the two-sample test statistics in Chen & Zhang (2013) and Zhang & Chen (2019), we define the extended basic quantities at time t under the averaging approach as

$$R_{1,(a)}(t) = \sum_{k=1}^{K} \frac{n_{1k}(t)\{n_{1k}(t)-1\}}{m_k} + \sum_{(u,v)\in C_0} \frac{n_{1u}(t)n_{1v}(t)}{m_u m_v}, \tag{3}$$

$$R_{2,(a)}(t) = \sum_{k=1}^{K} \frac{n_{2k}(t)\{n_{2k}(t)-1\}}{m_k} + \sum_{(u,v)\in C_0} \frac{n_{2u}(t)n_{2v}(t)}{m_u m_v}, \tag{4}$$

and under the union approach as

$$R_{1,(u)}(t) = \sum_{k=1}^{K} \frac{n_{1k}(t)\{n_{1k}(t)-1\}}{2} + \sum_{(u,v)\in C_0} n_{1u}(t)n_{1v}(t), \tag{5}$$

$$R_{2,(u)}(t) = \sum_{k=1}^{K} \frac{n_{2k}(t)\{n_{2k}(t)-1\}}{2} + \sum_{(u,v)\in C_0} n_{2u}(t)n_{2v}(t). \tag{6}$$

These are discrete-data versions of $R_{1,G}(t)$ and $R_{2,G}(t)$ for addressing the infeasibility of computing test statistics on data with repeated observations. Hence, a relatively large value of the sum of $R_{1,(a)}(t)$ and $R_{2,(a)}(t)$, or of $R_{1,(u)}(t)$ and $R_{2,(u)}(t)$, can be taken as evidence against the null hypothesis.

Under the null hypothesis H_0 in (1), the null distribution is defined to be the permutation distribution, which places $1/n!$ probability on each of the $n!$ permutations of $\{y_i\}_{i=1,\ldots,n}$. Let pr, E, var and cov denote the probability, expectation, variance and covariance, respectively, under the permutation null distribution. Analytical formulas for the expectation and variance of extended basic quantities can then be calculated through combinatorial analysis; the detailed expressions and proofs are given in the Supplementary Material.

For any candidate value t of τ, the extended test statistics can be defined as

$$Z_{w,(a)}(t) = \frac{R_{w,(a)}(t) - E\{R_{w,(a)}(t)\}}{[\mathrm{var}\{R_{w,(a)}(t)\}]^{1/2}}, \qquad Z_{w,(u)}(t) = \frac{R_{w,(u)}(t) - E\{R_{w,(u)}(t)\}}{[\mathrm{var}\{R_{w,(u)}(t)\}]^{1/2}},$$

$$S_{(a)}(t) = Z^2_{w,(a)}(t) + Z^2_{d,(a)}(t), \qquad S_{(u)}(t) = Z^2_{w,(u)}(t) + Z^2_{d,(u)}(t),$$

$$M_{(a)}(t) = \max\big\{(Z_{w,(a)}(t), |Z_{d,(a)}(t)|\big\}, \qquad M_{(u)}(t) = \max\big\{Z_{w,(u)}(t), |Z_{d,(u)}(t)|\big\},$$

where

$$R_{w,(a)}(t) = \{1 - w(t)\}R_{1,(a)}(t) + w(t)R_{2,(a)}(t),$$

$$R_{w,(u)}(t) = \{1 - w(t)\}R_{1,(u)}(t) + w(t)R_{2,(u)}(t),$$

$$Z_{d,(a)}(t) = \frac{R_{d,(a)}(t) - E\{R_{d,(a)}(t)\}}{[\mathrm{var}\{R_{d,(a)}(t)\}]^{1/2}}, \quad R_{d,(a)}(t) = R_{1,(a)}(t) - R_{2,(a)}(t),$$

$$Z_{d,(u)}(t) = \frac{R_{d,(u)}(t) - E\{R_{d,(u)}(t)\}}{[\mathrm{var}\{R_{d,(u)}(t)\}]^{1/2}}, \quad R_{d,(u)}(t) = R_{1,(u)}(t) - R_{2,(u)}(t),$$

with $w(t) = (t-1)/(n-2)$. Relatively large values of the test statistics are evidence against the null.

3.2. *New scan statistics*

Based on the extended statistics, we can define scan statistics for the single-changepoint alternative to handle data with repeated observations as follows:

(i) extended weighted edge-count scan statistic, $\max_{n_0 \leqslant t \leqslant n_1} Z_{w,(a)}(t)$ and $\max_{n_0 \leqslant t \leqslant n_1} Z_{w,(u)}(t)$;
(ii) extended generalized edge-count scan statistic, $\max_{n_0 \leqslant t \leqslant n_1} S_{(a)}(t)$ and $\max_{n_0 \leqslant t \leqslant n_1} S_{(u)}(t)$;
(iii) extended max-type edge-count scan statistic, $\max_{n_0 \leqslant t \leqslant n_1} M_{(a)}(t)$ and $\max_{n_0 \leqslant t \leqslant n_1} M_{(u)}(t)$.

Here n_0 and n_1 are set to prespecified values. For example, we can set $n_0 = \lceil 0.05n \rceil$ and $n_1 = n - n_0$ so as to include enough observations to represent the distribution.

Each scan statistic has its own characteristics, suited to different situations; see § 5 for a comparison. For example, the extended weighted edge-count test is useful when a changepoint occurs away from the middle of the sequence; the extended generalized edge-count test is effective under both location and scale alternatives; the extended max-type edge-count test is similar, but gives more accurate p-value approximation. The null hypothesis is rejected if the maximum is greater than a certain threshold. In § 4 we describe how to choose the threshold to control the Type I error rate.

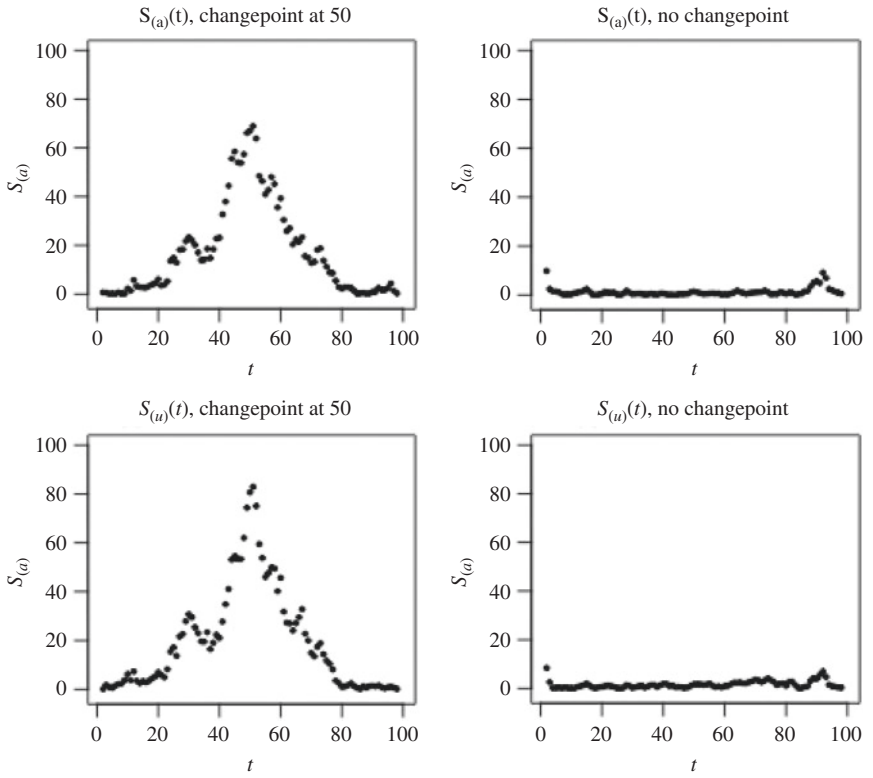

Fig. 1. Plots of $S_{(a)}(t)$ (top row) and $S_{(u)}(t)$ (bottom row) against t for a typical observation from $\mathrm{Mu}(10, \mathrm{prob}_1)$ and the second 50 observations from $\mathrm{Mu}(10, \mathrm{prob}_2)$, where $\mathrm{prob}_1 = (0.2, 0.3, 0.3, 0.2)^{\mathrm{T}}$ and $\mathrm{prob}_2 = (0.4, 0.3, 0.2, 0.1)^{\mathrm{T}}$ (left column), and all 100 observations from $\mathrm{Mu}(10, \mathrm{prob}_1)$ (right column). Here C_0 is the nearest-neighbour link on Euclidean distance.

For illustration, Fig. 1 plots the processes of $S_{(a)}(t)$ and $S_{(u)}(t)$ for the first 50 observations generated from $\mathrm{Mu}\{10, (0.2, 0.3, 0.3, 0.2)^{\mathrm{T}}\}$ and the second 50 observations generated from $\mathrm{Mu}\{10, (0.4, 0.3, 0.2, 0.1)^{\mathrm{T}}\}$, with C_0 being the nearest-neighbour link constructed on Euclidean distance. We see that both $S_{(a)}(t)$ and $S_{(u)}(t)$ peak at the true changepoint $\tau = 50$, as seen in the left panels. On the other hand, when there is no changepoint, $S_{(a)}(t)$ and $S_{(u)}(t)$ have random fluctuations with smaller maximum values, as seen in the right panels. Illustrations of the other test statistics are given in the Supplementary Material.

For testing the null hypothesis H_0 in (1) against the changed-interval alternative H_2 in (2), test statistics can be derived in a similar way to the single-changepoint case. For example, the extended generalized edge-count scan statistics are

$$\max_{\substack{1<t_1<t_2\leqslant n \\ n_0\leqslant t_2-t_1\leqslant n_1}} S_{(a)}(t_1, t_2), \qquad \max_{\substack{1<t_1<t_2\leqslant n \\ n_0\leqslant t_2-t_1\leqslant n_1}} S_{(u)}(t_1, t_2)$$

for the averaging and union approaches, respectively, where $S_{(a)}(t_1, t_2)$ and $S_{(u)}(t_1, t_2)$ are the extended generalized edge-count statistics on the two samples, consisting of observations within $[t_1, t_2)$ and observations outside $[t_1, t_2)$. The details of all statistics for the changed-interval alternative can be found in the Supplementary Material.

4. Analytical p-value approximation

4.1. *Asymptotic distributions of the stochastic processes*

We first consider the averaging approach. We are interested in the tail distribution of the scan statistics under H_0. Taking the extended generalized edge-count scan statistic as an example, we want to compute

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} S_{(a)}(t)\Big\}, \quad \mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} S_{(u)}(t)\Big\}$$

for the single-changepoint alternative and compute

$$\mathrm{pr}\left\{\max_{\substack{1<t_1<t_2\leqslant n\\ n_0\leqslant t_2-t_1\leqslant n_1}} S_{(a)}(t_1,t_2)\right\}, \quad \mathrm{pr}\left\{\max_{\substack{1<t_1<t_2\leqslant n\\ n_0\leqslant t_2-t_1\leqslant n_1}} S_{(u)}(t_1,t_2)\right\}$$

for the changed-interval alternative.

Under the null hypothesis, the scan statistics are defined as the permutation distribution. For a small sample size n, we can directly sample from the permutation distribution to compute the permutation p-value. However, when n is large, one needs to draw a large number of random permutations to get a good estimate of the p-value, which is very time-consuming. Therefore, we seek to derive analytical approximations to these tail probabilities.

By the definition of $Z_{w,(a)}(t)$, $S_{(a)}(t)$ and $M_{(a)}(t)$, the stochastic processes $\{Z_{w,(a)}(t)\}$, $\{S_{(a)}(t)\}$ and $\{M_{(a)}(t)\}$ boil down to two pairs of basic processes: $\{Z_{w,(a)}(t)\}$ and $\{Z_{d,(a)}(t)\}$ for the single-changepoint case and, analogously, $\{Z_{w,(a)}(t_1,t_2)\}$ and $\{Z_{d,(a)}(t_1,t_2)\}$ for the changed-interval case. Therefore, we first study the properties of these basic stochastic processes. Let $\mathcal{E}_u^{C_0}$ be the subgraph of C_0 containing all edges that connect to node u, let $\mathcal{E}_{u,2}$ be the set of edges in C_0 that contain at least one node in $\mathcal{E}_u^{C_0}$, and let $|\mathcal{E}_u^{C_0}|$ and $|\mathcal{E}_{u,2}^{C_0}|$ denote the numbers of edges in the sets. To derive the asymptotic behaviour of the stochastic processes in the averaging approach, we make the following assumptions.

Condition 1. We have that $|C_0|, \sum_{(u,v)\in C_0}(m_u m_v)^{-1} = O(n)$.

Condition 2. We have that $\sum_{u=1}^{K} m_u(m_u + |\mathcal{E}_u^{C_0}|)\big(\sum_{v\in\mathcal{V}_u^{C_0}} m_v + |\mathcal{E}_{u,2}^{C_0}|\big) = o(n^{3/2})$.

Condition 3. We have that $\sum_{(u,v)\in C_0}\Big(m_u + m_v + |\mathcal{E}_u^{C_0}| + |\mathcal{E}_v^{C_0}|\Big)\big(\sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0}} m_w + |\mathcal{E}_{u,2}^{C_0}| + |\mathcal{E}_{v,2}^{C_0}|\big) = o(n^{3/2})$.

Condition 4. We have that $\sum_{u=1}^{K}(|\mathcal{E}_u^{C_0}| - 2)^2/(4m_u) - (|C_0| - K)^2/n = O(n)$.

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

Theorem 1. *Under Conditions 1–4, as $n \to \infty$ the following hold:*

(i) $\{Z_{w,(a)}([nw]) : 0 < w < 1\}$ *and* $\{Z_{d,(a)}([nw]) : 0 < w < 1\}$ *converge to independent Gaussian processes in finite-dimensional distributions, which we denote by* $\{Z^*_{w,(a)}(w) : 0 < w < 1\}$ *and* $\{Z^*_{d,(a)}(w) : 0 < w < 1\}$*, respectively;*

(ii) $\{Z_{w,(a)}([nw_1],[nw_2]) : 0 < w_1 < w_2 < 1\}$ *and* $\{Z_{d,(a)}([nw_1],[nw_2]) : 0 < w_1 < w_2 < 1\}$ *converge to independent Gaussian random fields in finite-dimensional distributions, which*

we denote by $\{Z^*_{w,(a)}(w_1,w_2) : 0<w_1<w_2<1\}$ *and* $\{Z^*_{d,(a)}(w_1,w_2) : 0<w_1<w_2<1\}$, *respectively.*

The proof of this theorem uses a technique developed in Chen & Zhang (2015), which utilizes Stein's method (Chen & Shao, 2005). The details of the proof are provided in the Supplementary Material.

Let $\rho^*_{w,(a)}(u,v) = \mathrm{cov}\{Z^*_{w,(a)}(u), Z^*_{w,(a)}(v)\}$ and $\rho^*_{d,(a)}(u,v) = \mathrm{cov}\{Z^*_{d,(a)}(u), Z^*_{d,(a)}(v)\}$. The next theorem gives explicit expressions for the covariance functions of the limiting Gaussian processes $\{Z^*_{w,(a)}(w) : 0<w<1\}$ and $\{Z^*_{d,(a)}(w) : 0<w<1\}$. It is proved through combinatorial analysis, and details can be found in the Supplementary Material.

THEOREM 2. *The covariance functions of the Gaussian processes* $Z^*_{w,(a)}(w)$ *and* $Z^*_{d,(a)}(w)$ *have the following expressions:*

$$\rho^*_{w,(a)}(u,v) = \frac{(u\wedge v)\{1-(u\vee v)\}}{(u\vee v)\{1-(u\wedge v)\}}, \quad \rho^*_{d,(a)}(u,v) = \left[\frac{(u\wedge v)\{1-(u\vee v)\}}{(u\vee v)\{1-(u\wedge v)\}}\right]^{1/2},$$

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

For the union approach, let $\bar{G}$ be the set of graphs that is the union of all possible optimal graphs between observations $\{y_i\}$, let $\mathcal{E}_i^{\bar{G}}$ be the subgraph of $\bar{G}$ containing all edges that connect to node y_i, and let $|\mathcal{E}_i^{\bar{G}}|$ be the number of edges in the subgraph. We assume the following conditions.

Condition 5. We have that $|\bar{G}| = O(n)$.

Condition 6. We have that $\sum_{u=1}^K m_u^3 \sum_{v\in\mathcal{V}_u^{C_0}} m_v \sum_{v\in\mathcal{V}_u^{C_0}} m_v\big(m_v+\sum_{w\in\mathcal{V}_v^{C_0}\setminus\{v\}} m_w\big) = o(n^{3/2})$.

Condition 7. We have that $\sum_{(u,v)\in C_0} m_u m_v\big(m_u \sum_{w\in\mathcal{V}_u^{C_0}} m_w + m_v \sum_{w\in\mathcal{V}_v^{C_0}} m_w\big) \times \big\{\sum_{w\in\mathcal{V}_u^{C_0}\cup\mathcal{V}_v^{C_0},\, y\in\mathcal{V}_w^{C_0}\setminus\{w\}} m_w(m_w+m_y)\big\} = o(n^{3/2})$.

Condition 8. We have that $\sum_{i=1}^n |\mathcal{E}_i^{\bar{G}}|^2 - 4|\bar{G}|^2/n = O(n)$.

THEOREM 3. *Under Conditions 5–8, as* $n\to\infty$ *the following hold:*

(i) $\{Z_{w,(u)}([nw]) : 0<w<1\}$ *and* $\{Z_{d,(u)}([nw]) : 0<w<1\}$ *converge to independent Gaussian processes in finite-dimensional distributions, which we denote by* $\{Z^*_{w,(u)}(w) : 0<w<1\}$ *and* $\{Z^*_{d,(u)}(w) : 0<w<1\}$, *respectively;*

(ii) $\{Z_{w,(u)}([nw_1],[nw_2]) : 0<w_1<w_2<1\}$ *and* $\{Z_{d,(u)}([nw_1],[nw_2]) : 0<w_1<w_2<1\}$ *converge to independent Gaussian random fields in finite-dimensional distributions, which we denote by* $\{Z^*_{w,(u)}(w_1,w_2) : 0<w_1<w_2<1\}$ *and* $\{Z^*_{d,(u)}(w_1,w_2) : 0<w_1<w_2<1\}$, *respectively.*

See the Supplementary Material for the proof.

Let $\rho^*_{w,(u)}(u,v) = \mathrm{cov}\{Z^*_{w,(u)}(u), Z^*_{w,(u)}(v)\}$ and $\rho^*_{d,(u)}(u,v) = \mathrm{cov}\{Z^*_{d,(u)}(u), Z^*_{d,(u)}(v)\}$. The next theorem provides explicit expressions for the covariance functions of the limiting Gaussian processes $\{Z^*_{w,(u)}(w) : 0<w<1\}$ and $\{Z^*_{d,(u)}(w) : 0<w<1\}$. Its proof is given in the Supplementary Material.

THEOREM 4. *The covariance functions of the Gaussian processes $Z^*_{w,(u)}(w)$ and $Z^*_{d,(u)}(w)$ have the following expressions:*

$$\rho^*_{w,(u)}(u,v) = \frac{(u\wedge v)\{1-(u\vee v)\}}{(u\vee v)\{1-(u\wedge v)\}}, \quad \rho^*_{d,(u)}(u,v) = \left[\frac{(u\wedge v)\{1-(u\vee v)\}}{(u\vee v)\{1-(u\wedge v)\}}\right]^{1/2}.$$

Remark 1. By Theorems 2 and 4, we see that the limiting distributions of $\{Z_{w,(a)}([nw]) : 0 < w < 1\}$ and $\{Z_{w,(u)}([nw]) : 0 < w < 1\}$ are identical and do not depend on the graph at all. The same is true for the Z_d. In addition, the covariance functions in Theorems 2 and 4 are the same as in Theorem 4.3 of Chu & Chen (2019). Hence, limiting distributions of the extended graph-based tests based on Z_w, S and M are exactly the same as their corresponding versions for continuous data. On the other hand, the limiting distributions of the extended original edge-count scan statistics differ from their corresponding versions in the continuous setting; for details see the Supplementary Material.

Remark 2. Conditions 1–8 all constrain the number of repeated observations. Conditions 1 and 5 can usually be satisfied with an appropriate choice of C_0. Conditions 2, 3, 6 and 7 constrain the degrees of nodes in the graph C_0 such that they cannot be too large. Condition 4 ensures that $\{R_{1,(a)}(t), R_{2,(a)}(t)\}^{\mathrm{T}}$ does not degenerate asymptotically so that $S_{(a)}(t)$ is well-defined; similarly for Condition 8.

We have checked these conditions through simulation studies, shown in the Supplementary Material, and find that some of them can be violated even when the p-value approximation still works well. Zhu & Chen (2021) recently studied graph-based two-sample tests for continuous data and proposed much more relaxed conditions than those in Chu & Chen (2019). They checked their conditions under both sparse and dense graphs, and the conditions were found to hold well. We believe that conditions for data with repeated observations in the changepoint setting can also be relaxed. However, this requires a substantial amount of work, which we leave to future research.

4.2. *Asymptotic p-value approximation*

We now examine the asymptotic behaviour of tail probabilities. Using arguments similar to those in the proof of Proposition 3.4 in Chen & Zhang (2015), we can derive analytical approximations to the probabilities. Assume that $n_0, n_1, n, b \to \infty$ in such a way that for some $0 < x_0 < x_1 < 1$ and $b_1 > 0$, we have $n_0/n \to x_0$, $n_1/n \to x_1$ and $b/\sqrt{n} \to b_1$.

Based on Theorems 1 and 3, as $n \to \infty$, for both the averaging and the union approaches, we have

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} Z^*_w(t/n) > b\Big\} \sim b\phi(b)\int_{x_0}^{x_1} h^*_w(x)\nu\big[b_1\{2h^*_w(x)\}^{1/2}\big]\,\mathrm{d}x,$$

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} |Z^*_d(t/n)| > b\Big\} \sim 2b\phi(b)\int_{x_0}^{x_1} h^*_d(x)\nu\big[b_1\{2h^*_d(x)\}^{1/2}\big]\,\mathrm{d}x,$$

where $\nu(s) \approx (2/s)\{\Phi(s/2) - 0.5\}/\{(s/2)\Phi(s/2) + \phi(s/2)\}$ (Siegmund & Yakir, 2007) with $\Phi(\cdot)$, and $\phi(\cdot)$ being the standard normal cumulative density function and probability density function, respectively, and

$$h_w^*(x) = \lim_{s\nearrow x} \frac{\partial \rho_{w,(a)}^*(s,x)}{\partial s} = -\lim_{s\searrow x} \frac{\partial \rho_{w,(a)}^*(s,x)}{\partial s} = \lim_{s\nearrow x} \frac{\partial \rho_{w,(u)}^*(s,x)}{\partial s} = -\lim_{s\searrow x} \frac{\partial \rho_{w,(u)}^*(s,x)}{\partial s},$$

$$h_d^*(x) = \lim_{s\nearrow x} \frac{\partial \rho_{d,(a)}^*(s,x)}{\partial s} = -\lim_{s\searrow x} \frac{\partial \rho_{d,(a)}^*(s,x)}{\partial s} = \lim_{s\nearrow x} \frac{\partial \rho_{d,(u)}^*(s,x)}{\partial s} = -\lim_{s\searrow x} \frac{\partial \rho_{d,(u)}^*(s,x)}{\partial s}.$$

It can be shown that $h_w^*(x) = \{x(1-x)\}^{-1}$ and $h_d^*(x) = \{2x(1-x)\}^{-1}$.

Since $Z_{w,(a)}^*(t)$ and $Z_{d,(a)}^*(t)$ are independent, and $Z_{w,(u)}^*(t)$ and $Z_{d,(u)}^*(t)$ are independent, for both the averaging and the union approaches, we have

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} M^*(t/n) > b\Big\} = 1 - \mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} |Z_d^*(t/n)| < b\Big\}\,\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} Z_w^*(t/n) < b\Big\}.$$

In addition, following arguments similar to those in the proof of Proposition 4.4 in Chu & Chen (2019), we obtain analytical p-value approximations for the extended generalized edge-count test. Assume that $n_0, n_1, n, b_S \to \infty$ in such a way that for some $0 < x_0 < x_1 < 1$ and $b_2 > 0$, we have $n_0/n \to x_0$, $n_1/n \to x_1$ and $b_S/n \to b_2$. Then, as $n \to \infty$, for both the averaging and the union approaches, we have

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} S^*(t/n) > b_S\Big\} \sim \frac{b_S \exp(-b_S/2)}{2\pi} \int_0^{2\pi}\int_{x_0}^{x_1} h_s^*(x,w)\, v\big[\{2b_2 h_s^*(x,w)\}^{1/2}\big]\,\mathrm{d}x\,\mathrm{d}w,$$

where $h_s^*(x,w) = h_d^*(x)\cos^2(w) + h_w^*(x)\sin^2(w)$. The analytical p-value approximations for the changed-interval setting are presented in the Supplementary Material.

Remark 3. In practice, we use $h_w(n,x)$ in place of $h_w^*(x)$, where $h_w(n,x)$ is the finite-sample equivalent of $h_w^*(x)$, that is,

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

with $\rho_w(s,t) = \mathrm{cov}\{Z_w(s), Z_w(t)\}$. The explicit expression for $h_w(n,x)$ is

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

It is clear from the above expression that $h_w(n,x)$ does not depend on the graph C_0, and it is easy to show that $\lim_{n\to\infty} h_w(n,x) = h_w^*(x)$. The finite-sample equivalent of $h_d^*(x)$ is exactly the same as $h_d^*(x)$, i.e.,

$$h_d(n,x) = n \lim_{s\nearrow nx} \frac{\partial \rho_d(s,nx)}{\partial s} = \frac{1}{2x(1-x)},$$

where $\rho_d(s,t) = \mathrm{cov}\{Z_d(s), Z_d(t)\}$.

4.3. *Skewness correction*

The analytical p-value approximations based on the asymptotic results give ballpark estimates of the p-values. However, they are in general not accurate enough if n_0 and n_1 are taken close to the two ends, and when the dimension is high. The inaccuracy can largely be attributed to the

fact that convergence of $Z_{w,(a)}(t)$, $Z_{w,(u)}(t)$, $Z_{d,(a)}(t)$ and $Z_{d,(u)}(t)$ to the Gaussian distribution is slow when t/n is close to 0 or 1.

To improve the analytical p-value approximations, we add extra terms in the formulas to correct for skewness. In our situation, the extent of the skewness depends on the value of t. Hence, we adopt a skewness correction approach discussed in Chen & Zhang (2015), where different amounts of correction are applied for different t values; in particular, this approach leverages better approximation of the marginal probability by using the third moment, γ.

After skewness correction, the analytical p-value approximations for the averaging approach are

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} Z_{w,(a)}(t) > b\Big\} \sim b\phi(b)\int_{n_0/n}^{n_1/n} H_{w,(a)}(nx)h_w(n,x)v\big[b\{2h_w(n,x)/n\}^{1/2}\big]\,\mathrm{d}x,$$

where

$$H_{w,(a)}(t) = \frac{\exp\big[\frac{1}{2}\{b-\hat\theta_{b,w,(a)}(t)\}^2 + \frac{1}{6}\gamma_{w,(a)}(t)\hat\theta^3_{b,w,(a)}(t)\big]}{\{1+\gamma_{w,(a)}(t)\hat\theta_{b,w,(a)}(t)\}^{1/2}},$$

$$\hat\theta_{b,w,(a)}(t) = \frac{\{1+2\gamma_{w,(a)}(t)b\}^{1/2}-1}{\gamma_{w,(a)}(t)},$$

and

$$\mathrm{pr}\Big\{\max_{n_0\leqslant t\leqslant n_1} Z_{d,(a)}(t) > b\Big\} \sim b\phi(b)\int_{n_0/n}^{n_1/n} H_{d,(a)}(nx)h_d(n,x)v\big[b\{2h_d(n,x)/n\}^{1/2}\big]\,\mathrm{d}x,$$

where

$$H_{d,(a)}(t) = \frac{\exp\big[\frac{1}{2}\{b-\hat\theta_{b,d,(a)}(t)\}^2 + \frac{1}{6}\gamma_{d,(a)}(t)\hat\theta^3_{b,d,(a)}(t)\big]}{\{1+\gamma_{d,(a)}(t)\hat\theta_{b,d,(a)}(t)\}^{1/2}},$$

$$\hat\theta_{b,d,(a)}(t) = \frac{\{1+2\gamma_{d,(a)}(t)b\}^{1/2}-1}{\gamma_{d,(a)}(t)},$$

with $\gamma_{w,(a)}(t) = E\{Z^3_w(t)\}$ and $\gamma_{d,(a)}(t) = E\{Z^3_d(t)\}$, whose analytical expressions are provided in the Supplementary Material. The skewness-corrected analytical p-value approximations for the union approach and the changed-interval case can be derived in a similar manner; for details see the Supplementary Material.

Remark 4. By jointly correcting for the marginal probabilities of $Z_w(t)$ and $Z_d(t)$, we can derive skewness-corrected p-value approximations for $\max_{n_0\leqslant t\leqslant n_1} S(t) = \max_{0\leqslant w\leqslant 2\pi} \max_{n_0\leqslant t\leqslant n_1}\{Z_w(t)\sin(w) + Z_d(t)\cos(w)\}$ (Chu & Chen, 2019). However, the integrand could easily be nonfinite, so the method relies heavily on extrapolation. We therefore do not perform skewness correction on $S_{(a)}(t)$ and $S_{(u)}(t)$.

4.4. *Checking analytical p-value approximations under finite n*

We assess the performance of the analytical p-value approximations obtained in § 4.2 and § 4.3. In particular, we compare the critical values for the 0.05 p-value threshold through analytical p-value approximations based on asymptotic results, and skewness correction applied to values

obtained from performing 10 000 permutations under various simulation settings, to check how well the analytical approximation works for finite samples. Here, we focus on the extended max-type scan statistic for the single-changepoint alternative. Results for the other scan statistics and the changed-interval alternative are reported in the Supplementary Material.

We consider three distributions with different dimensions: C1, multinomial $d = 10$ with equal probabilities; C2, Gaussian $d = 100$ with repeated observations; and C3, multinomial $d = 1000$ with equal probabilities. We take C_0 to be the nearest-neighbour link constructed on Euclidean distance. The analytical approximations depend on constraints, n_0 and n_1, on the region in which one searches for the changepoint. For simplicity we set $n_1 = n - n_0$.

Since analytical p-value approximations without skewness correction do not depend on C_0 in the extended weighted, generalized and max-type tests, the critical value is determined only by n, n_0 and n_1. Analytical p-value approximations without skewness correction yield the same result in both the averaging and the union approaches. On the other hand, the skewness-corrected approximated p-values depend on certain characteristics of the graph structure. The structure of the nearest-neighbour link depends on the underlying dataset, and so the critical values vary across simulation runs.

Table 3 shows the results for the extended max-type scan statistics. The top part, labelled A1, presents the analytical critical values without skewness correction. The lower portion reports skewness-corrected analytical critical values and permutation critical values in the averaging and union cases. We also show results for two randomly simulated sequences in each setting. It can be seen that the asymptotic p-value approximation performs reasonably well. As the window size decreases, the analytical critical values become less precise. However, the skewness-corrected approximation performs much better than the approximation without skewness correction. When the dimension is not too high, such as in C1, the skewness-corrected analytical approximation performs reasonably well for n_0 as low as 25. When the dimension is high, such as in C2 and C3, the approximation performs well when $n_0 \geqslant 50$.

5. PERFORMANCE OF THE NEW TESTS

We study two aspects of the performance of the new tests: first, whether the test can reject the null hypothesis of homogeneity when there is a change; and second, if the test can reject H_0, whether it can estimate the location of the changepoint accurately. We use the configuration model random graph $G(v, \vec{k})$ to generate networks. Here, v is the number of vertices and $\vec{k} = (k_1, \ldots, k_v)$ is a degree sequence on v vertices, with k_i being the degree of vertex i. To generate configuration model random graphs, given a degree sequence we choose a uniformly random matching on the degree stubs, or half-edges.

We generate a sequence of $n = 200$ networks from the model

$$y_i \sim \begin{cases} G(v, \vec{k}_1), & i = 1, \ldots, \tau, \\ G(v, \vec{k}_2), & i = \tau + 1, \ldots, 200. \end{cases}$$

In this simulation we explore two cases for the location of the changepoint: in the middle, $\tau = 100$, and close to one end, $\tau = 170$, for $v = 6$ vertices. This dataset has repeated networks. Also, we consider two types of changes:

(i) an equal degree of changes in the network, where all elements of $\vec{k}_1$ are 2 and two elements of $\vec{k}_2$ are 4 with the rest being 2;

Table 3. *Critical values for the single-changepoint scan statistics* $\max_{n_0 \leqslant t \leqslant n_1} M_{(a)}(t)$ *and* $\max_{n_0 \leqslant t \leqslant n_1} M_{(u)}(t)$ *based on the nearest-neighbour link at the* 0.05 *significance level, for* $n = 1000$

	$n_0 = 100$	$n_0 = 75$	$n_0 = 50$	$n_0 = 25$
A1	3.24	3.28	3.32	3.38

Critical values (a)

	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2 (a)	Per (a)	A2 (a)	Per (a)	A2 (a)	Per (a)	A2 (a)	Per (a)
(C1)	3.30	3.30	3.36	3.37	3.43	3.43	3.54	3.58
	3.30	3.30	3.35	3.36	3.43	3.46	3.55	3.62
(C2)	3.36	3.34	3.44	3.45	3.56	3.59	3.72	3.98
	3.34	3.36	3.42	3.47	3.53	3.64	3.76	4.03
(C3)	3.30	3.30	3.38	3.41	3.48	3.57	3.67	3.93
	3.30	3.28	3.38	3.39	3.48	3.56	3.67	3.87

Critical values (u)

	$n_0 = 100$		$n_0 = 75$		$n_0 = 50$		$n_0 = 25$	
	A2 (u)	Per (u)	A2 (u)	Per (u)	A2 (u)	Per (u)	A2 (u)	Per (u)
(C1)	3.32	3.30	3.37	3.40	3.44	3.43	3.54	3.59
	3.31	3.32	3.36	3.35	3.43	3.46	3.55	3.63
(C2)	3.35	3.36	3.42	3.43	3.51	3.52	3.62	3.80
	3.34	3.39	3.40	3.46	3.48	3.55	3.67	3.84
(C3)	3.31	3.30	3.39	3.41	3.50	3.57	3.69	3.93
	3.31	3.28	3.39	3.39	3.50	3.56	3.69	3.87

A1, analytical critical values without skewness correction; A2 (a) and A2 (u), skewness-corrected analytical critical values in the averaging and union approaches, respectively; Per (a) and Per (u), permutation critical values in the averaging and union cases, respectively.

Table 4. *Estimated power of the new tests*

	$Z_{w,(a)}(t)$	$Z_{w,(u)}(t)$	$S_{(a)}(t)$	$S_{(u)}(t)$	$M_{(a)}(t)$	$M_{(u)}(t)$
(S1)	0.98 (0.96)	0.96 (0.89)	0.96 (0.94)	0.95 (0.89)	0.96 (0.95)	0.95 (0.88)
(S2)	0.88 (0.83)	0.89 (0.85)	0.90 (0.84)	0.91 (0.85)	0.89 (0.83)	0.90 (0.87)
(S3)	0.86 (0.83)	0.65 (0.59)	0.85 (0.83)	0.85 (0.82)	0.81 (0.80)	0.70 (0.64)
(S4)	0.81 (0.81)	0.73 (0.70)	0.86 (0.84)	0.93 (0.91)	0.84 (0.81)	0.86 (0.82)

(ii) a random degree of changes in the network, where all elements of $\vec{k}_1$ are 2 and two elements of $\vec{k}_2$ are randomly selected from $\{3, 4, 5\}$ while the rest are 2.

Thus, we consider the following four combinations: S1, an equal degree change at $\tau = 100$; S2, a random degree change at $\tau = 100$; S3, an equal degree change at $\tau = 170$; and S4, a random degree change at $\tau = 170$.

We use an adjacency matrix M_t to represent the network at t, such that the (i, j) element is 1 if vertices i and j are connected and 0 otherwise. We consider the dissimilarity defined by the number of different entries normalized by the geometric mean of the total edges in each of the two networks, $\|M_i - M_j\|_{\mathrm{F}}/(\|M_i\|_{\mathrm{F}}\,\|M_j\|_{\mathrm{F}})^{1/2}$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. As a similarity graph for our new methods we take C_0 to be the nearest-neighbour link.

Table 4 shows the number of null rejections out of 100 at the 0.05 significance level for each method. For the accuracy of estimating changepoint location, the count where the estimated changepoint is within 20 of the true changepoint is shown in parentheses when the null hypothesis

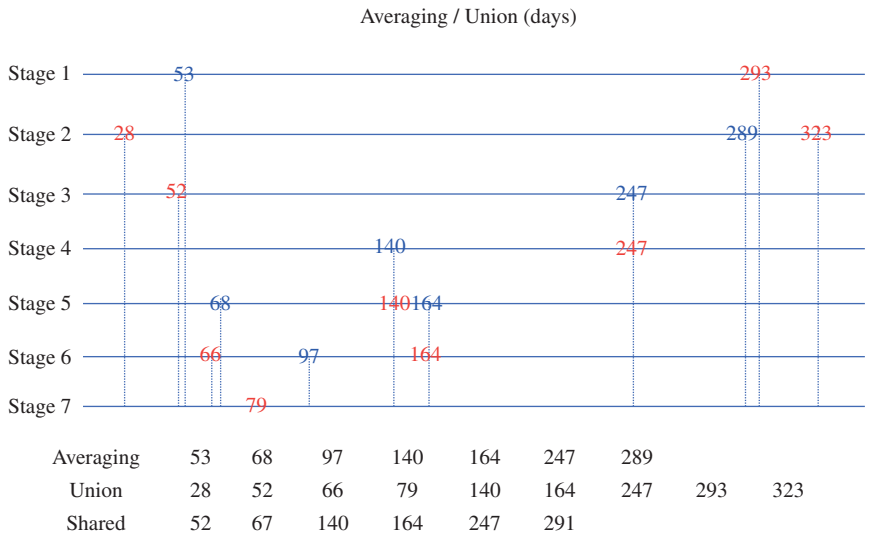

Fig. 2. Estimated changepoints and the order in which changepoints were detected by the averaging (blue) and union (red) approaches.

is rejected. We see that all tests work well in the case of a balanced equal degree of changes, while the extended generalized edge-count test outperforms the other tests in the case of a random degree of changes. In this simulation, equal-degree changes could be considered a mean change, and random-degree changes could be considered a change in both location and scale. Hence, the extended generalized edge-count test and max-type edge-count test perform well in this general scenario. When the changepoint is not at the centre of the sequence, the extended weighted edge-count test performs best, which agrees with what we would expect. It can be seen that the extended generalized edge-count test and max-type edge-count test work well when the change is in both mean and variance in the unbalanced-sample-size case.

6. PHONE-CALL NETWORK DATA ANALYSIS

In this section we apply the new tests to a phone-call network dataset. The MIT Media Laboratory conducted a study following 87 subjects who used mobile phones with a pre-installed device that can record call logs. The study lasted for 330 days from July 2004 to June 2005 (Eagle et al., 2009). One question of interest is whether there is any change in the phone-call patterns between subjects over time; this can be viewed as a change in friendship between the subjects over time.

We discard the phone-calls by day and construct $t = 330$ networks in total with 87 subjects as nodes. We encode each network by the adjacency matrix B_t, whose (i,j) element takes value 1 if subject i called subject j on day t and has value 0 otherwise. We define the distance measure as the number of different entries, i.e., $d(B_i, B_j) = \|B_i - B_j\|_{\mathrm{F}}^2$. Because of repeated observations, many equal distances between distinct values exist. We let C_0 be the nearest-neighbour link in this example.

We use the single-changepoint detection method, applying the extended generalized scan statistic to the phone-call network dataset recursively to detect all possible changepoints. As this dataset contains a lot of noise, we focus on estimated changepoints with a p-value of less than 0.001. Figure 2 shows the estimated changepoints obtained by the averaging approach and the union approach. We see that the two approaches produce quite a number of similar changepoints. We define a changepoint $\hat{\tau}$ as being detected by both approaches if they each find a changepoint

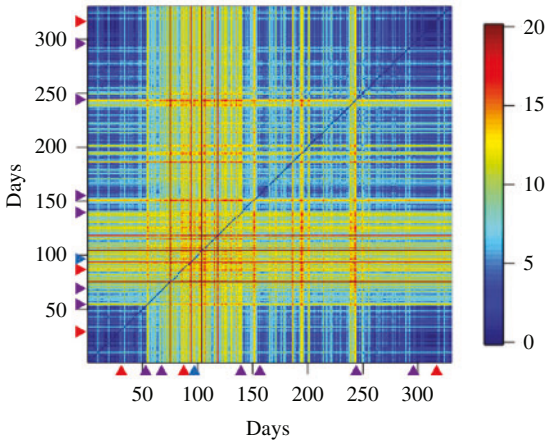

Fig. 3. Heatmap of an L_1-norm distance matrix corresponding to 330 networks. Red and blue triangles indicate estimated changepoints obtained by the union approach and averaging approach, respectively, and purple triangles indicate their shared changepoints.

within the set $[\hat{\tau} - 2,\ \hat{\tau} + 2]$. We then deem the location of the shared changepoint to be the floor of the average of the changepoints detected by the two approaches.

Since we do not know the underlying distribution of the dataset, we perform a check with the distance matrix of the whole period; see Fig. 3. It is evident that there are some signals in this dataset, and they show a reasonably good match with our results from the new tests.

7. Conclusion

In general, the averaging and union approaches work similarly, although inevitably they sometimes produce different results. A brief comparison of the two approaches is provided in the Supplementary Material. The proposed methods detect the most significant single changepoint or the changed interval in the sequence. If the sequence has more than one changepoint, the methods can be applied recursively using techniques such as binary segmentation, circular binary segmentation and wild binary segmentation (Vostrikova, 1981; Olshen et al., 2004; Fryzlewicz, 2014).

Acknowledgement

The authors were supported in part by the U.S. National Science Foundation (DMS-1513653 and DMS-1848579).

Supplementary material

Supplementary material available at *Biometrika* online includes details of the phone-call network dataset, details of constructing C_0, the extended original scan statistic, scan statistics for the changed-interval case, discussions of the two solutions, p-value approximations, and proofs of all the theoretical results.

References

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

CHEN, H. & ZHANG, N. R. (2013). Graph-based tests for two-sample comparisons of categorical data. *Statist. Sinica* **23**, 1479–503.

CHEN, J. & GUPTA, A. K. (2011). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. New York: Springer.

CHEN, L. H. Y. & SHAO, Q.-M. (2005). Stein's method for normal approximation. In *An Introduction to Stein's Method*. Singapore: World Scientific, pp. 1–59.

CHU, L. & CHEN, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *Ann. Statist.* **47**, 382–414.

EAGLE, N., PENTLAND, A. S. & LAZER, D. (2009). Inferring friendship network structure by using mobile phone data. *Proc. Nat. Acad. Sci.* **106**, 15274–8.

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42**, 2243–81.

HARCHAOUI, Z., MOULINES, E. & BACH, F. R. (2009). Kernel change-point analysis. In *Advances in Neural Information Processing Systems (NIPS 2008)*. NeurIPS Proceedings, pp. 609–16.

LUNG-YUT-FONG, A., LÉVY-LEDUC, C. & CAPPÉ, O. (2012). Homogeneity and change-point detection tests for multivariate data using rank statistics. *arXiv:* 1107.1971v3.

MATTESON, D. S. & JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Am. Statist. Assoc.* **109**, 334–45.

OLSHEN, A. B., VENKATRAMAN, E., LUCITO, R. & WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–72.

R DEVELOPMENT CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

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

VOSTRIKOVA, L. Y. (1981). Detecting 'disorder' in multidimensional random processes. *Dokl. Akad. Nauk* **259**, 270–4.

WANG, G., ZOU, C. & YIN, G. (2018). Change-point detection in multinomial data with a large number of categories. *Ann. Statist.* **46**, 2020–44.

WANG, T. & SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Statist. Soc.* B **80**, 57–83.

ZHANG, J. & CHEN, H. (2019). Graph-based two-sample tests for discrete data. *arXiv:* 1711.04349v2.

ZHANG, N. R., SIEGMUND, D. O., JI, H. & LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97**, 631–45.

ZHU, Y. & CHEN, H. (2021). Limiting distributions of graph-based test statistics. *arXiv:* 2108.07446.

[*Received on* 18 *June* 2020. *Editorial decision on* 13 *September* 2021]