Electronic Journal of Statistics
Vol. 19 (2025) 1425–1452
ISSN: 1935-7524
https://doi.org/10.1214/25-EJS2367

On the tightness of graph-based statistics

Lynna Chu[1], **Hao Chen**[2]

[1] *Department of Statistics, Iowa State University, Ames, Iowa 50010 U.S.A.,*
e-mail: lchu@iastate.edu

[2] *Department of Statistics, University of California, Davis, Davis, CA 95616 U.S.A.,*
e-mail: hxchen@ucdavis.edu

Abstract: We establish tightness of graph-based stochastic processes in the space $D[0+\epsilon, 1-\epsilon]$ with $\epsilon > 0$ that allows for discontinuities of the first kind. The graph-based stochastic processes are based on statistics constructed from similarity graphs. In this non-parametric setting, the classic characterization of tightness is intractable, making it difficult to obtain convergence of the limiting distributions for graph-based stochastic processes. We take an alternative approach and study the behavior of the higher moments of the graph-based test statistics. We show that, under mild conditions of the graph, tightness of the stochastic process can be established by obtaining upper bounds on the graph-based statistics' higher moments. Explicit analytical expressions for these moments are provided. The results are applicable to generic graphs, including dense graphs where the number of edges can be of higher order than the number of observations. Numerical studies are presented to provide insight as to when tightness holds, and potential extensions to other scenarios are explored.

MSC2020 subject classifications: Primary 60G99; secondary 60C05.
Keywords and phrases: Change-point, graph-based tests, non-parametric, scan statistic, Gaussian process, tightness, non-Euclidean data.

Received December 2024.

1. Introduction

Change-point detection aims to estimate and test for the presence of change-points, locations where the distribution abruptly changes, in a sequence of observations. Research interest in change-point problems has surged in recent years and substantial contributions by the statistics community have resulted in a range of works [1, 21, 12, 13, 20, 23, 19, 17, 16]. In particular, an area of emphasis has been given to handling complex data types such as high-dimensional data or non-Euclidean data objects, including networks and images. Most change-point methods targeting complex data types are non-parametric and aim to make minimal assumptions on the underlying data generating mechanism in order to be widely applicable without restrictive assumptions (see [14, 15, 18, 11] and references therein). An obstacle for non-parametric works is that theoretical guarantees can pose immense challenges. For example, fast type I error control via analytical p-value approximations are generally difficult to work out in the non-parametric setting. The increasing complexity and volume of modern

datasets call for methods that can offer fast ways to assess changes while controlling type I error. However, most non-parametric approaches still depend on re-sampling techniques to obtain p-value approximations.

Recently, a graph-based framework for change-point detection was proposed in [7] and further studied in [9] that aims to address the needs of modern change-point applications by offering power, flexibility, and fast type I error control. The framework is a non-parametric approach that utilizes test statistics constructed from similarity graphs and is applicable to any data type, including high-dimensional and object data, as long as a similarity measure can be defined on the sample space. The similarity graph can be provided by domain knowledge or it can be generated according to some criteria, such as the minimum spanning tree or the nearest neighbor graph. No distributional form or family needs to be specified and the approach is powerful for detecting general changes (mean, variance, covariance, higher moments etc.) without needing to directly estimate the parameters of interest. This flexibility makes the approach applicable to a broad range of problems. Moreover, simulation studies and real data applications demonstrate that the approach is powerful under many settings involving high-dimensional and non-Euclidean data types [7, 9].

The graph-based framework is also equipped with analytical p-value approximations for testing the significance of change-points. This extends the graph-based frameworks applicability to settings where the volume or complexity of the observations make it computationally infeasible to assess significance via re-sampling techniques. A key step in obtaining these analytical p-value approximations is proving, under certain regularity conditions, that the stochastic processes of the graph-based test statistics converge to Gaussian processes in finite dimensional distribution (see Theorem 3.1 in [7] and Theorem 4.1 in [9]). These asymptotic results kick in even for relatively moderate sample sizes (in the hundreds) and do not require the number of observations to grow rapidly with dimensions. The p-value approximations given in [9] are also asymptotically distribution-free, meaning they do not depend on the underlying similarity graph. More details are provided in Section 2.

While the existing asymptotic theory perform well for finite samples, notably, the current theory given in [7] and [9] do not imply *convergence in distribution to Gaussian processes* since tightness of the processes is not established. Since the analytical p-value approximations necessitate that the graph-based processes converge in distribution to a Gaussian process, it is crucial to establish tightness of the graph-based processes. Tightness guarantees the existence of limit points for weak convergence and it ensures that intervals between the time points considered in the finite-dimensional distribution are well-behaved. This is essential for the type of test statistic, the maximum scan statistic, used in this framework (see (6) below).

In this paper, we establish tightness of the stochastic processes for non-parametric graph-based test statistics under mild conditions on the graph. In terms of theoretical work, our proof provides the final piece in establishing the limiting distribution of these graph-based processes, which is distinctly challenging to establish for non-parametric methods. To do so, we derive explicit

expressions for higher product moments of graph-based test statistics; these are obtained by studying configurations of the graph and combinatorial analysis. Importantly, our results hold for any generic graph, including dense graphs, and can be generalized to other graph-based stochastic processes to establish weak convergence. In terms of practical applications, our results provide further confidence in utilizing the asymptotic p-value approximations for modern data applications and the testing of change-points.

The paper is organized as follows: Section 2 provides a brief overview of the graph-based framework. The main results are given in Section 3 and the proof is provided in Section 4, with additional details in the Appendix. Section 5 presents numerical studies and discusses the conditions required in the theorems. Section 6 concludes with remarks on potential applications to other scenarios.

2. Review of the graph-based framework

Let $\{\mathbf{y_i} : i = 1, \ldots, n\}$ be a data sequence indexed by time or some other meaningful ordering, where $\mathbf{y}_t$ could be a high-dimensional observation or non-Euclidean object. In the single change-point setting, there possibly exists a change-point τ such that $\mathbf{y}_t$ follows some unknown distribution for $t \leq \tau$ and follows a different (unknown) distribution for $t > \tau$. Consider that each time t divides the sequence of observations into two samples: those observations before time t and those observations after time t. The graph-based framework utilizes graph-based two-sample test statistics to test whether or not these two samples are from the same distribution. By graph-based two-sample tests we refer to tests that are based on graphs with the observations $\{\mathbf{y}_i\}$ as nodes. The graph, G, is constructed from all observations in the sequence and is usually derived from a distance or a generalized dissimilarity on the sample space, with edges in the graph connecting observations that are "close" in some sense. For example, G could be the minimum spanning tree (MST), which is a tree connecting all observations such that the sum of the distances of edges in the tree is minimized; G could also be the nearest neighbor graph (NNG) where each observation connects to its nearest neighbors. Four statistics are considered in [7] and [9]. These are based on 3 quantities of the graph which we briefly discuss below.

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

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

For any candidate value t of τ, the three quantities are:

$$R_0(t) = \sum_{e \in G} I_{J_e(t)=0}, \quad R_1(t) = \sum_{e \in G} I_{J_e(t)=1}, \quad R_2(t) = \sum_{e \in G} I_{J_e(t)=2}. \tag{1}$$

Then $R_0(t)$ is the number of edges connecting observations before and after t, $R_1(t)$ is the number of edges connecting observations prior to t, and $R_2(t)$ is the number of edges that connect observations after t.

The four statistics considered are the edge-count test statistic (2), generalized edge-count test statistic (3), weighted edge-count test statistic (4), and max-type edge-count test statistic (5):

$$Z_0(t) = -\frac{R_0(t) - \mathbf{E}(R_0(t))}{\sqrt{\mathbf{Var}(R_0(t))}}, \tag{2}$$

$$S(t) = \begin{pmatrix} R_1(t) - \mathbf{E}(R_1(t)) \\ R_2(t) - \mathbf{E}(R_2(t)) \end{pmatrix}^T \mathbf{\Sigma}^{-1}(t) \begin{pmatrix} R_1(t) - \mathbf{E}(R_1(t)) \\ R_2(t) - \mathbf{E}(R_2(t)) \end{pmatrix}, \tag{3}$$

$$Z_w(t) = \frac{R_w(t) - \mathbf{E}(R_w(t))}{\sqrt{\mathbf{Var}(R_w(t))}}, \tag{4}$$

with $R_w(t) = p(t)R_1(t) + q(t)R_2(t), \qquad p(t) = \frac{n-t-1}{n-2}, \quad q(t) = \frac{t-1}{n-2},$

$$M(t) = \max\left(|Z_{\text{diff}}(t)|, Z_w(t)\right), \tag{5}$$

where $Z_{\text{diff}}(t) = \frac{R_{\text{diff}}(t) - \mathbf{E}(R_{\text{diff}}(t))}{\sqrt{\mathbf{Var}(R_{\text{diff}}(t))}}$, with $R_{\text{diff}}(t) = R_1(t) - R_2(t)$.

The expected value and variance of the four test statistics are computed under the permutation null distribution and their explicit expressions can be found in [7, 6, 4, 9]. Each of the test statistics has its own niche where it dominates; a detailed discussion can be found in [9, 5].

The null hypothesis of no change-point is rejected when the maximum scan statistic

$$\max_{n_0 \le t \le n_1} Z_0(t), \quad \max_{n_0 \le t \le n_1} Z_w(t), \quad \max_{n_0 \le t \le n_1} S(t), \quad \max_{n_0 \le t \le n_1} M(t) \tag{6}$$

is greater than a threshold with n_0 and n_1 being pre-specified constraints controlling where we search for the change-point. When n is small, this threshold can be obtained from permutation directly. However, this becomes computationally expensive for large n and instead, [7] and [9] provide accurate analytical formulas to approximate the p-values for these scan statistics.

To illustrate the accuracy of these p-value approximations, we compare the critical values based on asymptotic theory (labeled 'Asy') to the permutation critical values, obtained from implementing 10,000 permutations directly (labeled 'Perm'). The results are shown in Tables 1 - 3. Sequences of length $n = 500$ were generated from multivariate normal (Table 1), multivariate t with 5 degrees of freedom (Table 2), or multivariate log-normal distributions (Table 3). We can see even for finite sample sizes, the critical value approximations are performing reasonably in the high-dimension setting ($d > n$) relative to the permutation critical values. Our asymptotic results are reasonable even in the presence of heavy tails or skewness. Additional tables comparing critical values for different values of n, n_0 and n_1 can be found in Supplement A [10].

TABLE 1
Critical values for the graph-based scan statistics at 0.05 significance level. Observations are simulated from d-dimensional normal distribution. $n = 500$, $n_0 = 50$, $n_1 = n - n_0$.

	$\max_t Z_0(t)$		$\max_t Z_w(t)$		$\max_t S(t)$		$\max_t M(t)$	
	Asy	Perm	Asy	Perm	Asy	Perm	Asy	Perm
$d = 500$	2.69	2.72	3.00	2.99	12.90	13.20	3.25	3.27
$d = 1000$	2.68	2.67	2.99	2.97	12.90	12.92	3.24	3.22
$d = 2000$	2.66	2.66	2.99	2.97	12.90	13.28	3.24	3.25

TABLE 2
Critical values for the graph-based scan statistics at 0.05 significance level. Observations are simulated from d-dimensional t_5 *distribution.* $n = 500$, $n_0 = 50$, $n_1 = n - n_0$.

	$\max_t Z_0(t)$		$\max_t Z_w(t)$		$\max_t S(t)$		$\max_t M(t)$	
	Asy	Perm	Asy	Perm	Asy	Perm	Asy	Perm
$d = 500$	2.56	2.55	2.99	3.00	12.90	13.33	3.26	3.27
$d = 1000$	2.52	2.50	2.99	3.03	12.90	13.28	3.25	3.27
$d = 2000$	2.51	2.49	2.99	3.05	12.90	13.30	3.28	3.21

TABLE 3
Critical values for the graph-based scan statistics at 0.05 significance level. Observations are simulated from d-dimensional log-normal ($\mu = 0, \sigma = 1$) *distribution.* $n = 500$, $n_0 = 50$, $n_1 = n - n_0$.

	$\max_t Z_0(t)$		$\max_t Z_w(t)$		$\max_t S(t)$		$\max_t M(t)$	
	Asy	Perm	Asy	Perm	Asy	Perm	Asy	Perm
$d = 500$	2.33	2.43	2.99	3.04	12.90	13.70	3.21	3.35
$d = 1000$	2.74	2.36	2.99	3.04	12.90	13.99	3.21	3.36
$d = 2000$	2.74	2.30	2.99	3.03	12.90	14.00	3.21	3.39

We also report the empirical size at the significance level of 0.05 based on the asymptotic critical values. The setup is as follows: A sequence of length $n = 500$ was generated from one of three distributions: multivariate d-dimensional normal, t with 5 degrees of freedom, or multivariate log-normal distribution. For each sequence, the asymptotic critical value was calculated and used as the threshold. Using this threshold, we performed 10,000 permutations and computed the percentage of permutations with a maximum scan statistic exceeding the threshold. The empirical size, averaged over 100 trials, is presented in Table 4. The results show that the empirical size is close to the nominal size across different distributions.

3. Tightness of basic processes

3.1. Notation

Let $f_n \precsim g_n$ denote that f_n is bounded above by g_n (up to a constant) asymptotically and $f_n = o(g_n)$ denote that f_n is dominated by g_n asymptotically. We

TABLE 4
Empirical size of graph-based scan statistics at 0.05 *significance level.* $n = 500$, $n_0 = 50$, $n_1 = n - n_0$.

	$\max_{n_0 \le t \le n_1} Z_0(t)$	$\max_{n_0 \le t \le n_1} Z_w(t)$	$\max_{n_0 \le t \le n_1} S(t)$	$\max_{n_0 \le t \le n_1} M(t)$
normal, $d = 500$	0.0496	0.0496	0.0505	0.0505
normal, $d = 1000$	0.0504	0.0499	0.0532	0.0511
normal, $d = 2000$	0.0495	0.0504	0.0546	0.0510
t_5, $d = 500$	0.0504	0.0491	0.0552	0.0516
t_5, $d = 1000$	0.0504	0.0494	0.0557	0.0520
t_5, $d = 2000$	0.0503	0.0502	0.0561	0.0536
log-normal $d = 500$	0.0509	0.0492	0.0557	0.0516
log-normal $d = 1000$	0.0496	0.0494	0.0551	0.0518
log-normal, $d = 2000$	0.0477	0.0526	0.0631	0.0613

also write $f_n = O(g_n)$ to denote that f_n is bounded above and below by g_n, asymptotically; this will also be notated as $f_n \asymp g_n$.

3.2. Asymptotic null distributions of the basic processes

Given the scan statistics, we reject the null hypothesis of no change-point if the scan statistic is larger than a threshold. Explicitly, we are interested in the following tail probabilities: $P\left(\max_{n_0 \le t \le n_1} Z_0(t) > b_Z\right)$, $P\left(\max_{n_0 \le t \le n_1} S(t) > b_S\right)$, $P\left(\max_{n_0 \le t \le n_1} Z_w(t) > b_{Z_w}\right)$, and $P\left(\max_{n_0 \le t \le n_1} M(t) > b_M\right)$.

To obtain analytical approximations of these tail probabilities, [7] and [9] studied the properties of the stochastic processes $\{Z_0(t)\}, \{S(t)\}, \{Z_w(t)\}$, and $\{M(t)\}$ under the null hypothesis. Based on Lemma 3.1 in [9], $S(t)$ can be expressed as $S(t) = Z_w^2(t) + Z_{\text{diff}}(t)$, where $Z_w(t)$ and $Z_{\text{diff}}(t)$ are uncorrelated. Furthermore, $Z(t)$ can be expressed as

$$Z_0(t) = \frac{2\sigma_{R_w} Z_w(t)}{\sqrt{4\sigma_{R_w}^2 + (p(t) - q(t))^2 \sigma_{R_{\text{diff}}}^2}} + \frac{(p(t) - q(t))\sigma_{R_{\text{diff}}} Z_{\text{diff}}(t)}{\sqrt{4\sigma_{R_w}^2 + (p(t) - q(t))^2 \sigma_{R_{\text{diff}}}^2}},$$

where $\sigma_{R_w}^2 = \mathbf{Var}(R_w(t))$, $\sigma_{R_{\text{diff}}}^2 = \mathbf{Var}(R_{\text{diff}}(t))$, and $p(t)$ and $q(t)$ are defined as in (4). Therefore, these stochastic processes boil down to the basic processes: $\{Z_{\text{diff}}(t)\}$ and $\{Z_w(t)\}$.

In order to show that the limiting distributions of the basic processes converge to Gaussian processes, the classic approach as presented in [3] is to establish:

1. The convergence of $\{Z_w(\lfloor nu \rfloor) : 0 < u < 1\}$, and $\{Z_{\text{diff}}(\lfloor nu \rfloor) : 0 < u < 1\}$ to multivariate Gaussian in finite dimensional distributions.[1]
2. The tightness of $\{Z_w(\lfloor nu \rfloor) : 0 < u < 1\}$ and $\{Z_{\text{diff}}(\lfloor nu \rfloor) : 0 < u < 1\}$.

The first point has been proven in [7] and [9]. We prove here that the second point, tightness of the graph-based stochastic processes, does indeed hold under mild conditions on the graph.

[1] Throughout the paper, we use $\lfloor x \rfloor$ to denote the largest integer that is no larger than x.

3.3. Main results

We first state our main results and then give an outline of the proof. We use G to denote both the graph and its sets of edges. Let G_i be the subgraph of G containing all the edges that connect to node $\mathbf{y}_i$. Then, $|G_i|$ is the number of edges in G_i or the node degree of $\mathbf{y}_i$ in G. These results hold for generic similarity graphs, including dense graphs. We refer to a graph as dense if the number of edges is of higher order than the number of observations, i.e. if $|G| = O(kn)$ such that $k = O(n^\alpha)$.

Theorem 3.1. *Under the condition that k is at least $O(1)$ and $\sum_{i=1}^{n} |G_i|^2 = o(kn^2)$, the stochastic process $\{Z_w(\lfloor nu \rfloor) : 0 < u < 1\}$ is tight on the space $D[0+\epsilon, 1-\epsilon]$, where ϵ is a positive constant.*

Theorem 3.2. *Under the condition that k is at least $O(1)$ and $\sum_i |G_i|^2 - \frac{4|G|^2}{n}$ is at least $O(k^2)$, the stochastic process $\{Z_{diff}(\lfloor nu \rfloor) : 0 < u < 1\}$ is tight on the space $D[0+\epsilon, 1-\epsilon]$, where ϵ is a positive constant.*

These conditions are more relaxed than the conditions in [7] and [9] when obtaining convergence in finite dimensional distributions. A comparison of the conditions is provided in Section 5.1.

Let $D = D[0,1]$ be the space of real functions x on $[0,1]$ that are right-continuous and have left-hand limits:

(i) For $0 \le t < 1$, $x(t+) = \lim_{s \downarrow t} x(t)$ exists and $x(t+) = x(t)$,
(ii) For $0 \le t < 1$, $x(t-) = \lim_{s \uparrow t} x(t)$.

Functions satisfying these two properties are known as cadlag functions. A function x is said to have a discontinuity of the first kind at t if the left and right limits exist but differ and $x(t)$ lies between them. Any discontinuities of a cadlag function, an element of D, are of the first kind. Since

$$\lim_{u \downarrow c} Z_w(\lfloor nu \rfloor) = Z_w(\lfloor nc \rfloor), \qquad \lim_{u \uparrow c} Z_w(\lfloor nu \rfloor) = Z_w(\lfloor nu \rfloor),$$

$$\lim_{u \downarrow c} Z_{\text{diff}}(\lfloor nu \rfloor) = Z_{\text{diff}}(\lfloor nc \rfloor) \qquad \lim_{u \uparrow c} Z_{\text{diff}}(\lfloor nu \rfloor) = Z_{\text{diff}}(\lfloor nu \rfloor),$$

it follows that $Z_w(\lfloor nu \rfloor)$ and $Z_{\text{diff}}(\lfloor nu \rfloor)$ are right-continuous and have left-hand limits and therefore belong to the space D.

The classical characterization of tightness on the space D is given by Theorem 13.2 in [3], a version of which is presented here:

Definition 3.3. A sequence of stochastic processes $\{X^n(u) : 0 \le u \le 1\}$ in D is tight if and only if:

(i) *The sequence $\{X^n(u) : 0 \le u \le 1\}$ is stochastically bounded in D,*
(ii) *For each $\epsilon > 0$,*

$$\lim_{\delta \to 0} \limsup_n P(\omega'(X^n, \delta) > \epsilon) = 0,$$

where

$$\omega'(x,\delta) = \inf_{t_i} \max_i \sup_{s,t\in[t_{i-1},t_i)} |x(s)-x(t)|$$

and the infimum extends over all sets of $\{t_i\}$ *satisfying* $\min_{1\le i\le\nu}(t_i - t_{i-1}) > \delta$, *with* $0 < \delta < 1$ *and* t_i, $i = 1,\dots,\nu$.

In general these conditions are difficult to verify, since they involve understanding the limit supremum of a sequence. We instead take an alternative approach and use the tightness criterion proposed by Kolmogorov-Chentsov ([8], Theorem 1); a variant can also be found in [3]. The criterion is as follows:

Definition 3.4. A sequence of stochastic processes $X^n(u)$, $n = 1,2,\dots$, right continuous with left-hand limits, is tight if there are positive constants C,β,α not depending on n such that for any $0 \le u \le v \le w \le 1$,

$$\mathbf{E}(|X^n(v) - X^n(u)|^{2\beta}|X^n(w) - X^n(v)|^{2\beta}) \le C(w-u)^{1+\alpha}.$$

We set $\alpha = 1, \beta = 1$ so the criterion becomes:

$$\mathbf{E}\left((Z_w^n(v) - Z_w^n(u))^2(Z_w^n(w) - Z_w^n(v))^2\right) \le C_w(w-u)^2 \tag{7}$$

$$\mathbf{E}\left((Z_{\text{diff}}^n(v) - Z_{\text{diff}}^n(u))^2(Z_{\text{diff}}^n(w) - Z_{\text{diff}}^n(v))^2\right) \le C_{\text{diff}}(w-u)^2 \tag{8}$$

where the notation $Z_w^n(u) = Z_w(\lfloor nu\rfloor)$ and $Z_{\text{diff}}^n(u) = Z_{\text{diff}}(\lfloor nu\rfloor)$. Both inequalities automatically hold when $(w-u) \le \frac{1}{n}$ since at least one of the following is true: (i) $\lfloor nu\rfloor = \lfloor nv\rfloor$, (ii) $\lfloor nv\rfloor = \lfloor nw\rfloor$. In what follows, we focus on the case when $(w-u) > \frac{1}{n}$. Observe that $Z_w^n(u)$ and $Z_{\text{diff}}^n(u)$ are not well-defined at the boundaries, when $u = 0$ or $u = 1$. We further assume that $u,v,w = O(1)$ and therefore, cannot be too close to the boundaries. As such, we establish tightness on the domain $[0+\epsilon, 1-\epsilon]$, where ϵ is a positive constant.

An outline of our proof for Theorems 3.1 and 3.2 is as follows: we obtain explicit expressions for the 4th moments and product moments of Z_w and Z_{diff} using combinatorial analysis. This involves determining the different graph configurations for 4 edges to be randomly selected (with replacement) from the graph and obtaining the probabilities that each configuration will occur for the graph. Focusing on the leading terms of each configuration, we show these are bounded by $C(w-u)^2$.

4. Proof of Theorems 3.1 and 3.2

For simplicity, let $\lfloor nu\rfloor = r$, $\lfloor nv\rfloor = s$, and $\lfloor nw\rfloor = t$ and $r < s < t$. Then, expanding (7), we have

$$\begin{aligned}\mathbf{E}((Z_w^n(v)-Z_w^n(u))^2(Z_w^n(w) - Z_w^n(v))^2) = \\ &\mathbf{E}(Z_w^2(r)Z_w^2(s)) - 2\mathbf{E}(Z_w^2(r)Z_w(s)Z_w(t)) + \mathbf{E}(Z_w^2(r)Z_w^2(t)) \\ &- 2\mathbf{E}(Z_w(r)Z_w^3(s)) + \mathbf{E}(Z_w^2(s)Z_w^2(t)) - 2\mathbf{E}(Z_w(r)Z_w(s)Z_w^2(t)) \\ &+ \mathbf{E}(Z_w^4(s)) - 2\mathbf{E}(Z_w^3(s)Z_w(t)) + 4\mathbf{E}(Z_w(r)Z_w^2(s)Z_w(t)),\end{aligned}$$

and similarly for $\mathbf{E}\left((Z^n_{\text{diff}}(v) - Z^n_{\text{diff}}(u))^2 (Z^n_{\text{diff}}(w) - Z^n_{\text{diff}}(v))^2\right)$ (8).

For the two basic processes, the following analytical expressions are needed for Z_w:

$$\mathbf{E}(Z_w^2(r)Z_w(s)Z_w(t)), \quad (9)$$
$$\mathbf{E}(Z_w(r)Z_w(s)Z_w^2(t)), \quad (10)$$
$$\mathbf{E}(Z_w(r)Z_w^2(s)Z_w(t)), \quad (11)$$
$$\mathbf{E}(Z_w^2(r)Z_w^2(s)), \quad (12)$$
$$\mathbf{E}(Z_w^2(r)Z_w^2(t)), \quad (13)$$
$$\mathbf{E}(Z_w^2(s)Z_w^2(t)), \quad (14)$$
$$\mathbf{E}(Z_w(r)Z_w^3(s)), \quad (15)$$
$$\mathbf{E}(Z_w^3(s)Z_w(t)), \quad (16)$$
$$\mathbf{E}(Z_w^4(s)), \quad (17)$$

and the following analytical expressions are needed for Z_{diff}:

$$\mathbf{E}(Z_{\text{diff}}^2(r)Z_{\text{diff}}(s)Z_{\text{diff}}(t)), \quad (18)$$
$$\mathbf{E}(Z_{\text{diff}}(r)Z_{\text{diff}}(s)Z_{\text{diff}}^2(t)), \quad (19)$$
$$\mathbf{E}(Z_{\text{diff}}(r)Z_{\text{diff}}^2(s)Z_{\text{diff}}(t)), \quad (20)$$
$$\mathbf{E}(Z_{\text{diff}}^2(r)Z_{\text{diff}}^2(s)), \quad (21)$$
$$\mathbf{E}(Z_{\text{diff}}^2(r)Z_{\text{diff}}^2(t)), \quad (22)$$
$$\mathbf{E}(Z_{\text{diff}}^2(s)Z_{\text{diff}}^2(t)), \quad (23)$$
$$\mathbf{E}(Z_{\text{diff}}(r)Z_{\text{diff}}^3(s)), \quad (24)$$
$$\mathbf{E}(Z_{\text{diff}}^3(s)Z_{\text{diff}}(t)), \quad (25)$$
$$\mathbf{E}(Z_{\text{diff}}^4(s)). \quad (26)$$

It is straightforward to see that all the expressions can be decomposed as combinations of R_1 and R_2. Since explicit expressions for the expectation, variance, and third moments of $R_w(\cdot)$, $R_{\text{diff}}(\cdot)$, and $R(\cdot)$ can be found in [7] and [9], the remaining unknown quantities to be derived are the product moments of $R_1(\cdot)$ and $R_2(\cdot)$, which can be expressed as

$$\mathbf{E}(R_1^a(t_1^\star)R_2^b(t_2^\star)R_1^c(t_3^\star)R_2^d(t_4^\star))$$

where $a, b, c, d = 0, 1, 2, 3, 4$ such that $a + b + c + d = 4$ and $t_1^\star, t_2^\star, t_3^\star, t_4^\star = r, s, t$. The full list of product moments can be found in Supplement C [10].

To derive the analytical expressions for the product moments we need to:

1. Determine different configurations for 4 edges to be randomly selected (with replacement) from the graph,
2. Derive probabilities separately for each configuration.

There are in total nineteen different configurations for four edges randomly chosen (with replacement) from the graph; see Figure 1 for an illustration of each configuration. Let G be the similarity graph and G_i be the subgraph of G containing all edges that connect to node $\mathbf{y}_i$. Then $|G_i|$ is the degree of node $\mathbf{y}_i$ in G. Among all $|G|^4$ possible ways of randomly selecting the four edges, the number of occurrences for each of the configuration are:

1) $|G|$
2) $7x_1$

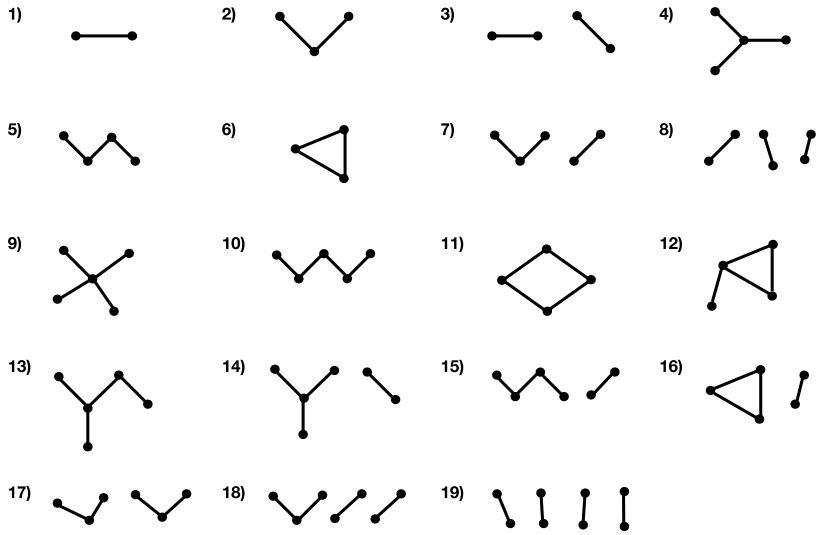

FIG 1. *Nineteen configurations of 4 edges randomly chosen, with replacement, from the graph.*

3) $7|G|(|G|-1)-7x_1$
4) $6x_2$
5) $36x_3$
6) $12x_5$
7) $18x_4-72x_3+36x_5$
8) $6|G|(|G|-1)(|G|-2)-12x_5-18x_4+36x_3-6x_2$
9) x_6
10) $12x_7-24x_8$
11) $6x_8$
12) $24x_9$
13) $12x_{10}-48x_9$
14) $4x_{11}-12x_{10}+24x_9$
15) $24x_{12}-24x_7+24x_8$
16) $8x_{13}-24x_9$
17) $3x_{14}-12x_7+12x_8$
18) $6x_{15}+36x_7-24x_8+72x_9-12x_{10}-48x_{12}-24x_{13}-6x_{14}$
19) $12x_{10}-12x_7-x_6-4x_{11}+24x_{12}+3x_{14}-6x_{15}+6x_8+16x_{13}+|G|(|G|-1)(|G|-2)(|G|-3)$

with $x_1,\ldots,x_{15}$ defined as:

$$x_1=\sum_{i=1}^{n}|G_i|^2-2|G|,$$

$$x_2=\sum_{i=1}^{n}|G_i|^3-3\sum_{i=1}^{n}|G_i|^2+4|G|,$$

$$
\begin{aligned}
x_3 =& \sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1),\\
x_4 =& |G|\sum_{i=1}^{n}|G_i|^2+\sum_{i=1}^{n}|G_i|^2-\sum_{i=1}^{n}|G_i|^3-2|G|^2,\\
x_5 =& \sum_{(i,j)}|\{l:(i,l),(j,l)\in G\},\\
x_6 =& \sum_{i=1}^{n}|G_i|^4-6\sum_{i=1}^{n}|G_i|^3+11\sum_{i=1}^{n}|G_i|^2-12|G|,\\
x_7 =& \sum_{(i,j),(j,l),i\neq l}(|G_i|-1)(|G_l|-1),\\
x_8 =& \sum_{(i,j),(j,l),i\neq l}|\{m:(i,m),(l,m)\in G\}||,\\
x_9 =& \sum_{(i,j)}\sum_{l:(i,l),(j,l)\in G}(|G_l|-2),\\
x_{10} =& \sum_{i}\sum_{j\in G_i;j\neq i}(|G_i|-1)^2(|G_j|-1)-2\sum_{i,j\in G}(|G_i|-1)(|G_j|-1),\\
x_{11} =& 4|G|^2-3|G|\sum_{i=1}^{n}|G_i|^2+|G|\sum_{i=1}^{n}|G_i|^3\\
&-2\sum_{i=1}^{n}|G_i|^2+3\sum_{i=1}^{n}|G_i|^3-\sum_{i=1}^{n}|G_i|^4,\\
x_{12} =& |G|\sum_{(i,j)}(|G_i|-1)(|G_j|-1)-\sum_{i}\sum_{j\in G_i;j\neq i}(|G_i|-1)^2(|G_j|-1)\\
&-\sum_{(i,j)}(|G_i|-1)(|G_j|-1),\\
x_{13} =& \sum_{(i,j)}\sum_{l:(i,l),(j,l)\in G}|G\setminus\{i,j,l\in G_l\}|,\\
x_{14} =& \sum_{i}\sum_{j\neq i}(|G_i\setminus\{j\in G_i\}|)(|G_i\setminus\{j\in G_i\}|-1)(|G_j\setminus\{i\in G_j\}|)\times\\
&(|G_j\setminus\{i\in G_j\}|-1),\\
x_{15} =& \sum_{i=1}^{n}|G_i|^4-2|G|\sum_{i=1}^{n}|G_i|^3+|G|^2\sum_{i=1}^{n}|G_i|^2+|G|\sum_{i=1}^{n}|G_i|^2\\
&-\sum_{i=1}^{n}|G_i|^2-2|G|^3+2|G|^2.
\end{aligned}
$$

Observe that the sum of these occurrences add up to $|G|^4$. Each occurrence represents the number of possible ways that configuration can occur. For example, for configuration 2, there are 7 ways that the 4 edges can selected (with

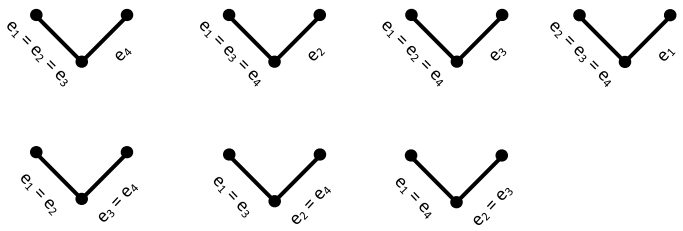

FIG 2. *Illustration of number of occurrences* ($7x_1$) *for configuration 2. The 4 randomly selected edges are labelled* $e_1, e_2, e_3,$ *and* e_4.

replacement) to obtain configuration 2. The multiplication by x_1 accounts for the number of ways two edges can share a node. See Figure 2 for an illustration for configuration 2. Two illustrative examples are presented in Supplement B [10], which provide combinatorial details on how to derive the probability for each configuration.

4.1. Expression for Z_w

The similarity graph G can be a generic graph constructed from a similarity measure, such as the Euclidean distance. Without loss of generality, $|G| = O(kn)$ with $k = O(n^\alpha), 0 \le \alpha < 1$. We assume that $u, v, w = O(1)$. To establish (7), we focus on the leading terms on the left-hand side of the inequality. After extensive simplification, the leading term for the denominator of $\mathbf{E}((Z_w^n(v) - Z_w^n(u))^2(Z_w^n(w) - Z_w^n(v))^2)$ is

$$\mathrm{den}_{Z_w} \triangleq v^2w^2(kn^2 - \sum_{i=1}^{n} |G_i|^2)^2(1-u)^2(1-v)^2. \tag{27}$$

The leading term for the numerator is

$$\begin{aligned}
\mathrm{num}_{Z_w} \triangleq (w-v)(v-u)\Big(&k^2n^4C_{w,1} + x_{14}C_{w,2} + C_{w,3}\sum_{i=1}^{n}|G_i|^4 \\
&+ nC_{w,4}\sum_{i=1}^{n}|G_i|^3 + C_{w,5}\sum_{i}\sum_{j\in G_i; j\neq i}(|G_i|-1)^2(|G_j|-1) \\
&+ kn^2C_{w,6}\sum_{i=1}^{n}|G_i|^2 + n^2C_{w,7}\sum_{i=1}^{n}|G_i|^2 \\
&+ knC_{w,8}\sum_{i,j\in G}(|G_i|-1)(|G_j|-1) + nC_{w,9}\sum_{i,j\in G}(|G_i|-1)(|G_j|-1) \\
&+ nx_7C_{w,10} + nx_8C_{w,11} + nx_9C_{w,12}\Big)
\end{aligned}$$

with

$$C_{w,1} = 4vw(1-v)(1-u) + 2(v-u)(w-v),$$

$$
\begin{aligned}
C_{w,2} =& 8vw(v-u)(1-u)(1-v),\\
C_{w,3} =& -2(v-u)(w-v)+2v(1-u)(1+v)+vw(5u-7)(1-v),\\
C_{w,4} =& 8v(w-v)-8w+2v(2+9w)(1-u)(1-v),\\
C_{w,5} =& 8(w-uv)+(48-56v)(w-v)+16(3v^2+w)(1-u)\\
&-4vw(49-37u)(1-v),\\
C_{w,6} =& -4(v-u)(w-v)-8vw(1-v)(1-u),\\
C_{w,7} =& 2(w-uv)+2(1-2v)(w-v)+vw(9u-11)(1-v)+2v^2(1-u),\\
C_{w,8} =& 16(v-u)(w-v)+32vw(1-v)(1-u),\\
C_{w,9} =& 2(28v-23)(w-v)-2(23v^2+9w)(1-u)\\
&-2vw(72u-95)(1-v)+10(uv-w),\\
C_{w,10} =& -8vw(1-u)(1-v)-4(w-v)(v-u),\\
C_{w,11} =& 4vw(1-u)(1-v)+2(w-v)(v-u),\\
C_{w,12} =& 8v(5v(1-v)-(1-u)(12v^2-7v+2))\\
&-(w-v)(24(1-u)+8(1-v)(12uv-17v+4)).
\end{aligned}
$$

Since $u, v, w = O(1)$, the terms $C_{w,1}, \ldots, C_{w,12}$ can be bounded asymptotically by a constant. Then the expression $\mathrm{num}_{Z_w}/\mathrm{den}_{Z_w}$ can be bounded by $C(w-u)^2$ as long as the ratio of graph configurations in the numerator and denominator can be bounded asymptotically by $O(1)$. The terms in the numerator involving graph configurations that need to be bounded are:

1. k^2n^4,
2. x_{14},
3. $\sum_{i=1}^n |G_i|^4$,
4. $n\sum_{i=1}^n |G_i|^3$,
5. $\sum_i \sum_{j\in G_i; j\neq i}(|G_i|-1)^2(|G_j|-1)$,
6. $kn^2\sum_{i=1}^n |G_i|^2$, $n^2\sum_{i=1}^n |G_i|^2$,
7. $kn\sum_{i,j\in G}(|G_i|-1)(|G_j|-1)$, $n\sum_{i,j\in G}(|G_i|-1)(|G_j|-1)$,
8. nx_7,
9. nx_8,
10. nx_9.

Since $(w-v)(v-u) < (w-u)^2$ for $u < v < w$, if the ratio of each term to the denominator $(kn^2 - \sum_{i=1}^n |G_i|^2)^2$ is bounded by $O(1)$, the entire expression can be asymptotically bounded by a constant $C_w \times (w-u)^2$.

In the following, we assume that $\sum_{i=1}^n |G_i|^2 = o(kn^2)$ and check each configuration in their order of appearance. Technical details are deferred to Appendix A. Key results and properties are summarized as follows:

- $x_{14} \precsim (\sum_{i=1}^n |G_i|^2)^2$,
- $\sum_{1=1}^n |G_i|^4 \precsim k^2n^4$, $n\sum_{1=1}^n |G_i|^3 \precsim k^2n^4$, and $kn^2\sum_{i=1}^n |G_i|^2 \precsim k^2n^4$,
- $\sum_i \sum_{j\in G_i; j\neq i}(|G_i|-1)^2(|G_j|-1) < 2|G|\sum_{i=1}^n |G_i|^3 \precsim 2|G|kn^3 \asymp k^2n^4$,
- $\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1) < \sum_{i=1}^n |G_i|(|G|-|G_i|) = |G|\sum_{i=1}^n |G_i| - \sum_{i=1}^n |G_i|^2 < 2|G|^2 \asymp 2k^2n^2$,

- $nx_7 < n\sum_{i=1} |G_i|(|G| - |G_i|) < 2n|G|^2 \asymp k^2n^3$,
- $nx_8 \precsim kn^4$,
- $nx_9 \precsim k^2n^3 (x_9 \precsim k^2n^2)$.

Based on these inequalities, we can bound the graph configurations with the denominator's $(kn^2 - \sum_{i=1}^n |G_i|^2)^2$. Since we assume $\sum_{i=1}^n |G_i|^2 = o(kn^2)$, it follows that $k^2n^4 \asymp (kn^2 - \sum_{i=1}^n |G_i|^2)^2$, which implies $\frac{k^2n^4}{(kn^2 - \sum_{i=1}^n |G_i|^2)^2} \asymp O(1)$.

Since $x_{14} \precsim k^2n^4$, we have $\frac{x_{14}}{(kn^2 - \sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$.

Similarly, from the key results $\sum_{1=1}^n |G_i|^4 \precsim k^2n^4$ and $n\sum_{1=1}^n |G_i|^3 \precsim k^2n^4$, we have $\frac{\sum_{i=1}^n |G_i|^4}{(kn^2 - \sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$ and $\frac{n\sum_{i=1}^n |G_i|^3}{(kn^2 - \sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$.

It follows that all remaining terms can be asymptotically bounded by k^2n^4. Therefore, the ratio of the graph configurations to $(kn^2 - \sum_{i=1}^n |G_i|^2)^2$ is asymptotically bounded by $O(1)$.

4.2. Expression for Z_{diff}

We adopt a similar approach for Z_{diff} and study the analytical expression for $\mathbf{E}\left((Z^n_{\text{diff}}(v) - Z^n_{\text{diff}}(u))^2 (Z^n_{\text{diff}}(w) - Z^n_{\text{diff}}(v))^2\right)$. This expression can be written as the combination of terms involving u, v, and w and terms involving graph configurations. We first show that the expressions involving u, v, and w $(K_1(u, v, w), \ldots,$
$K_6(u, v, w))$ can be bounded by $C(w-u)^2$ or $C(w-u)$. We then show that the graph-configurations are bounded asymptotically by $O(1)$ or $O(1/n)$. It follows then that the entire expression can be bounded by a constant $C_{\text{diff}} \times (w-u)^2$.

Let $e_v = v(1-v)$, $e_w = w(1-w)$, and $e_u = u(1-u)$. The leading term for the denominator of $\mathbf{E}\left((Z^n_{\text{diff}}(v) - Z^n_{\text{diff}}(u))^2 (Z^n_{\text{diff}}(w) - Z^n_{\text{diff}}(v))^2\right)$ is:

$$\text{den}_{Z_{\text{diff}}} = (nV_G)^2 w(1-u) e_u e_v^3 e_w$$

with $V_G = \sum_i |G_i|^2 - 4|G|^2/n$.

For the numerator of $\mathbf{E}\left((Z^n_{\text{diff}}(v) - Z^n_{\text{diff}}(u))^2 (Z^n_{\text{diff}}(w) - Z^n_{\text{diff}}(v))^2\right)$, we group the leading terms by their graph configurations. The numerator can be expressed as

$$\begin{aligned}
&K_1(u, v, w) \times k^4n^2 + K_2(u, v, w) \times k^2n(\sum_{i=1}^n |G_i|^2) + K_3(u, v, w) \times \sum_{i=1}^n |G_i|^4 \\
&+ K_4(u, v, w) \times k\sum_{i=1}^n |G_i|^3 + K_5(u, v, w) \times x_{14} \\
&+ K_6(u, v, w) \times \sum_i \sum_{j \in G_i; j \neq i} (|G_i| - 1)^2 (|G_j| - 1).
\end{aligned}$$

We can establish that the coefficients $K_1(u, v, w)$, $K_2(u, v, w), K_3(u, v, w)$, $K_4(u, v, w)$, $K_5(u, v, w)$, and $K_6(u, v, w)$ are bounded by $C(w-u)^2$ or $C(w-u)$. The technical details, being long and complex, are deferred to Appendix B.

In order for the entire expression to be bounded by $C(w-u)^2$ we need the graph configurations in the numerator and denominator to be bounded by $O(1)$ or $O(1/n)$. Recall that the leading term is the denominator is $(nV_G)^2$. Let $\tilde{d}_i = |G_i| - \frac{2|G|}{n}$, then $V_G = \sum_{i=1}^n \tilde{d}_i{}^2$. The graph configurations in the numerator involve:

1. k^4n^2,
2. $k^2n\sum_{i=1}^n |G_i|^2$,
3. $\sum_{i=1}^n |G_i|^4$,
4. $k\sum_{i=1}^n |G_i|^3$,
5. x_{14},
6. $\sum_i \sum_{j\in G_i; j\neq i}(|G_i|-1)^2(|G_j|-1)$.

Let $k = O(n^\alpha), 0 \leq \alpha < 1$. Suppose the largest (centered) degree $\tilde{d}_i \precsim O(n^\beta)$, where $0 \leq \beta < 1$.

We first focus on the second configuration 2 in the numerator, we have:

$$\sum_{i=1}^n |G_i|^2 = \sum_{i=1}^n (\tilde{d}_i + \frac{2|G|}{n})^2 \precsim \sum_{i=1}^n (n^\beta + n^\alpha)^2 \precsim n^{2\beta+1} + n^{2\alpha+1}.$$

Since $k^2n \precsim O(n^{2\alpha+1})$, it follows that the entire expression $kn^2\sum_{i=1}^n |G_i|^2 \precsim n^{2\beta+2\alpha+2} + n^{4\alpha+2}$.

In the denominator, if $\alpha \leq \beta$, then $V_G = \sum_{i=1}^n \tilde{d}_i{}^2 \succsim n^{2\beta}$, and $(nV_G)^2 \succsim n^{4\beta+2}$. Then the ratio of the numerator 2 and denominator gives us

$$\frac{n^{2\alpha+2\beta+2} + n^{4\alpha+2}}{n^{4\beta+2}} \precsim O(1).$$

If $\alpha > \beta$, then $k^2n\sum_{i=1}^n |G_i|^2 \precsim n^{4\alpha+2}$. With the assumption that $V_G \succsim k^2 \asymp n^{2\alpha}$, we have $(nV_G)^2 \succsim n^{4\alpha+2}$. Other terms can be done in a similar way. Notice that:

1. $k^4n^2 \precsim O(n^{4\alpha+2})$,
3. $\sum_{i=1}^n |G_i|^4 = \sum_{i=1}^n (\tilde{d}_i + \frac{2|G|}{n})^4 \precsim \sum_{i=1}^n (n^\beta + n^\alpha)^4 \precsim n^{4\beta+1} + n^{4\alpha+1}$,
4. $k\sum_{i=1}^n |G_i|^3 \precsim n^\alpha \sum_{i=1}^n (n^\beta + n^\alpha)^3 \precsim n^{3\beta+\alpha+1} + n^{4\alpha+1}$,
5. $x_{14} = \sum_i \sum_{j\neq i}(|G_i \setminus \{j \in G_i\}|)(|G_i \setminus \{j \in G_i\}| - 1)(|G_j \setminus \{i \in G_j\}|)(|G_j \setminus \{i \in G_j\}| - 1)$
   $\precsim \sum_{i=1} |G_i|^2 \sum_{j=1} |G_j|^2 \precsim \sum_{i,j} (n^\beta + n^\alpha)^4 \precsim n^{4\beta+2} + n^{4\alpha+2}$,
6. $\sum_{i=1} \sum_{j\in G_i; j\neq i}(|G_i|-1)^2(|G_j|-1)$
   $\precsim \sum_{i=1} \sum_{j\in G_i; j\neq i} |G_i|^2|G_j| \precsim n^{3\beta+1+\alpha}$. □

Therefore, the ratio of the first 5 configurations can be bounded by $O(1)$ and the 6th configuration can be bounded by $O(1/n)$. To see that the 6th configuration can be bounded by $O(1/n)$, consider that if $\alpha \leq \beta$, then $(nV_G)^2 \succsim n^{4\beta+2}$ and the ratio of the numerator and denominator is $\frac{1}{n^{(1+\beta-\alpha)}}$. If $\alpha > \beta$, then $(nV_G)^2 \succsim n^{4\alpha+2}$ and the ratio becomes $\frac{1}{n^{(3(\alpha-\beta)+1)}}$. Recall that expression for Z_{diff} can be expressed as the linear combination of the leading coefficients $K_1(u,v,w), \ldots, K_6(u,v,w)$ multiplied by their respective graph configurations.

We have established that $K_1(u,v,w),\ldots,K_5(u,v,w)$ are bounded by $C(w-u)^2$ and $K_6(u,v,w)$ is bounded by $C(w-u)$. Combining these results, and that we are considering the case that $(w-u) > \frac{1}{n}$, it follows that the expression for Z_{diff} can be bounded by $C(w-u)^2$.

5. Discussion

5.1. Comparison of conditions

The conditions for tightness in Theorems 3.1 and 3.2 are much more relaxed than the conditions needed to establish convergence of finite-dimensional distribution in Theorem 4.1 in [9]. To see this, we first define some notations used in [9]. For an edge $e = (e^-, e^+) \in G$, define $A_e := G_{e^+} \cup G_{e^-}$ as the subgraph in G that connects to either node e^- or node e^+. Define $B_e := \cup_{e* \in A_e} A_{e*}$ as the subgraph in G that connects to any node in A_e. Let $k = O(n^\alpha)$, where $0 \le \alpha < 1$. The sufficient conditions in Theorem 4.1 from [9] are as follows:

1. $|G| = O(n^\beta), 1 \le \beta < 1.25$;
2. $\sum_{e\in G} |A_e||B_e| = o(|G|^{1.5})$;
3. $\sum_{e\in G} |A_e|^2 = o(|G|\sqrt{n})$;
4. $\sum_{i=1}^n |G_i|^2 - 4|G|^2/n = O(\sum_{i=1}^n |G_i|^2)$.

The conditions in Theorem 3.1 require k to be at least $O(1)$ and $\sum_{i=1^n} |G_i|^2 = o(kn^2)$. The requirement that k is at least $O(1)$ is automatically satisfied by the first condition in [9], as the number of edges in the graph must be at least $O(n)$. Moreover, unlike the first condition in [9], our condition does not impose the additional restriction that $k \prec n^{0.25}$. The condition $\sum_{i=1^n} |G_i|^2 = o(kn^2)$ follows naturally from the third condition in [9]. Specifically, since $\sum_{i=1}^n |G_i|^2 \le 2\sum_{e\in G} |A_e|^2$, and the third condition in [9] is $\sum_{e\in G} |A_e|^2 = o(|G|\sqrt{n})$, it follows that $\sum_{i=1}^n |G_i|^2 = o(kn^{1.5}) = o(kn^2)$. Thus, the condition $\sum_{i=1^n} |G_i|^2 = o(kn^2)$ is much less restrictive than the third condition in Theorem 4.1 of [9].

The conditions in Theorem 3.2 require k to be at least $O(1)$ and $\sum_{i=1}^n |G_i|^2 - 4|G|^2/n$ to be at least $O(k^2)$. According to the fourth condition in [9], $\sum_{i=1}^n |G_i|^2 - 4|G|^2/n = O(\sum_{i=1}^n |G_i|^2)$. Since $\sum_{i=1}^n |G_i|^2 \ge 4|G|^2/n = 4k^2n$ and $4k^2n$ is way larger than $O(k^2)$, it follows that $\sum_{i=1}^n |G_i|^2 - 4|G|^2/n$ is at least $O(k^2)$. Thus, the condition in Theorem 3.2 is much less restrictive than the fourth condition in Theorem 4.1 of [9].

It follows that the conditions for tightness are significantly less restrictive than those provided in [9] for establishing convergence of finite-dimensional distributions. Consequently, if the conditions in [9] are satisfied, the tightness conditions will automatically hold, allowing us to confidently use the asymptotic properties of the maximum scan statistic for change-point detection.

5.2. Implications of conditions on tightness

To illustrate the sufficiency of our conditions, we construct similarity graphs that intentionally violate the conditions in Theorems 3.1 and 3.2 and examine

how Billingsley's tightness criterion (defined in [3]) behaves when the graph conditions are not met. The modulus of continuity in D is defined as

$$\omega'(x,\delta) = \inf_{t_i} \max_i \sup_{s,t\in[t_{i-1},t_i)} |x(s) - x(t)|.$$

If the stochastic process $x(\cdot)$ is tight, then asymptotically, when t and s are close, the values of $x(t)$ and $x(s)$ should also be close. We assess this for a finite sample size by evaluating the maximum jump in the process for sequential time points t_{i-1} and t_i. The simulation setup is as follows: we construct a similarity graph, calculate the graph-based stochastic process, and then evaluate $\sup|Z_w(t_{i-1}) - Z_w(t_i)|$ and $\sup|Z_{\text{diff}}(t_{i-1}) - Z_{\text{diff}}(t_i)|$ for $i \in [n_0, n_1]$. We let $n = 500$, $n_0 = 0.05 * n$, and $n_1 = n - n_0$. We repeat this process 100 times. We consider two types of graphs that violate the conditions in Theorem 3.1 and Theorem 3.2, respectively.

In Theorem 3.1, we require that $\sum_{i=1}^n |G_i|^2 = o(kn^2)$ in order for the stochastic process $\{Z_w(\lfloor nu \rfloor) : 0 < u < 1\}$ to be tight. Consider that $\sum_{i=1}^n |G_i|^2 < \sum_{i=1}^n \max(|G_i|)|G_i|$. Then even if $\max(|G_i|) = n$, we have $\sum_{i=1}^n |G_i|^2 \precsim kn^2$. Therefore, to violate this condition, we must consider a complete graph where $|G| = n(n-1)/2$. It follows that, if $\max(|G_i|) = n$, we have $\sum_{i=1}^n |G_i|^2 < \sum_{i=1}^n \max(|G_i|)|G_i| = n\sum_{i=1}^n |G_i| = 2n|G| \asymp n^3$, which violates our condition in Theorem 3.1. In Figure 3, we evaluate the tightness criterion for $Z_w(t)$ for graphs generated from a complete graph with varying number of random edges deleted from the graph. Since $n = 500$, the complete graph has $n_E = 124{,}749$ edges. We remove an increasing proportion of edges (from a single edge to 90%). We see that when the condition is violated (for example, the complete graph with only 1 edge removed), the jumps between $Z_w(t_{i-1})$ and $Z_w(t_i)$ tend to be larger. As we progressively delete more edges, the severity of the condition violation decreases and we observe that difference in jumps also begins to decrease. In Figure 4, it is clear that as the similarity graph moves away from the complete graph (i.e. the number of edges deleted increases), the value of $\sum_{i=1}^n |G_i|^2/kn^2$ begins to decrease as well.

In Theorem 3.2, we require $\sum_i |G_i|^2 - \frac{4|G|^2}{n}$ to be at least $O(k^2)$ in order for the stochastic process $\{Z_{\text{diff}}(\lfloor nu \rfloor) : 0 < u < 1\}$ to be tight. This condition can be violated when the similarity graph is very flat. For example, if every node degree is k, then $|G_i| = k$, for all $i = 1, \ldots, n$ and $\sum_i |G_i|^2 - \frac{4|G|^2}{n} = nk^2 - nk^2 = 0$. This violates our condition in Theorem 3.2. In Figure 5, we evaluate ω' for the graph-based stochastic process $Z_{\text{diff}}(t)$ constructed from a k-MDP (minimum distance pairing) similarity graph [6]. The MDP graph is constructed such that n nodes are divided into $n/2$ non-overlapping pairs in such a way as to minimize the total of $n/2$ distances between the pairs. The MDP can be extended to k-MDPs as well, where a k-MDP is defined similarly to a k-MST. By construction, the k-MDP constrains the node-degrees to be exactly k. In this setting, we set $k = \sqrt{n}$. We then randomly add an increasing number of edges to the k-MDP, gradually making the graph less flat. As shown in Figure 5, as the graph becomes less flat and the condition is less violated,

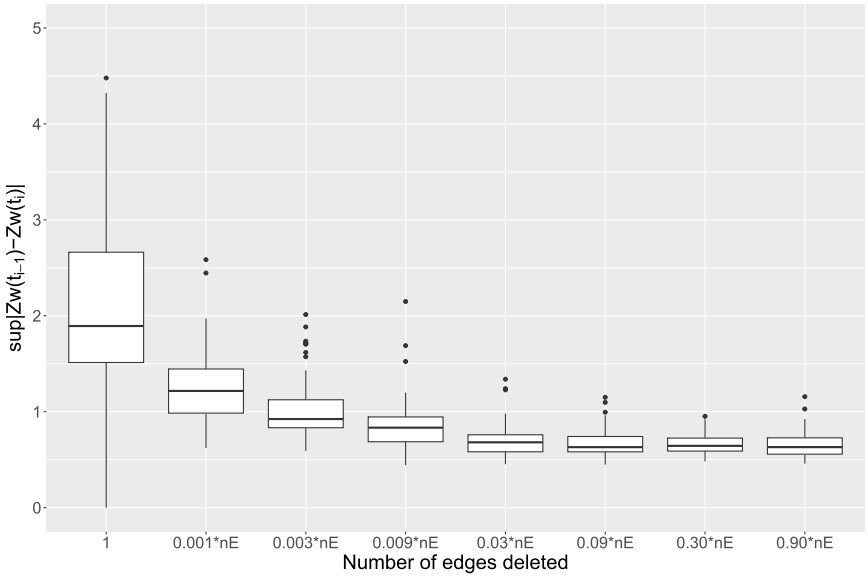

FIG 3. *Boxplots of* $\sup|Z_w(t_{i-1}) - Z_w(t_i)|$ *for graphs under different settings. 100 simulations are conducted for each setting. For each simulation, the graph is constructed from a complete graph, with a varying proportion of edges randomly deleted.*

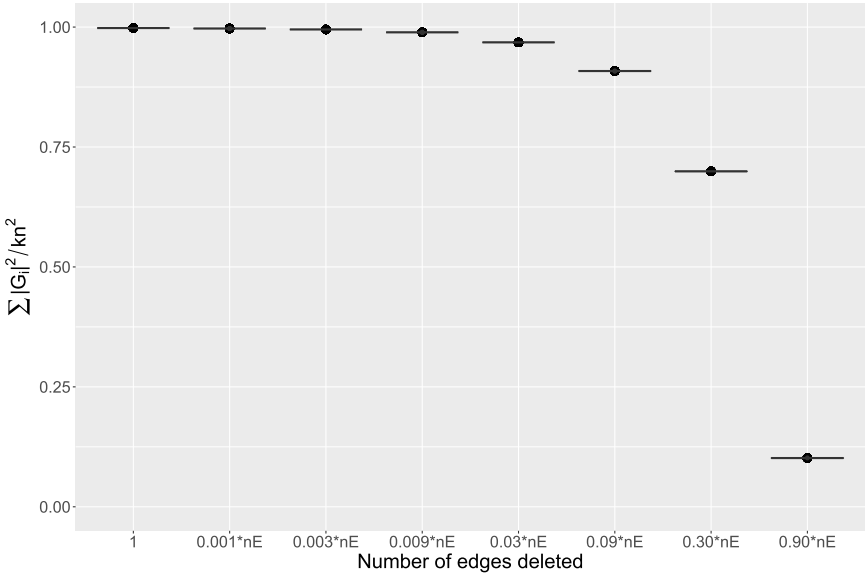

FIG 4. *Boxplots of* $\sum|G_i|^2/kn^2$ *for graphs under the same settings as in Figure 3.*

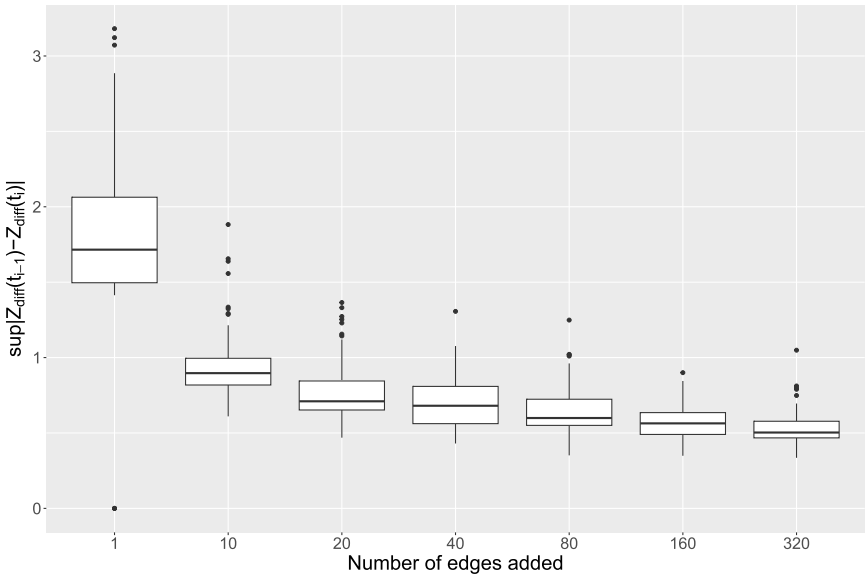

FIG 5. *Boxplots of* $\sup|Z_{diff}(t_{i-1}) - Z_{diff}(t_i))|$ *for graphs under different settings. 100 simulations are conducted for each setting. For each simulation, the graph is constructed from a k-MDP, with a varying number of edges randomly added.*

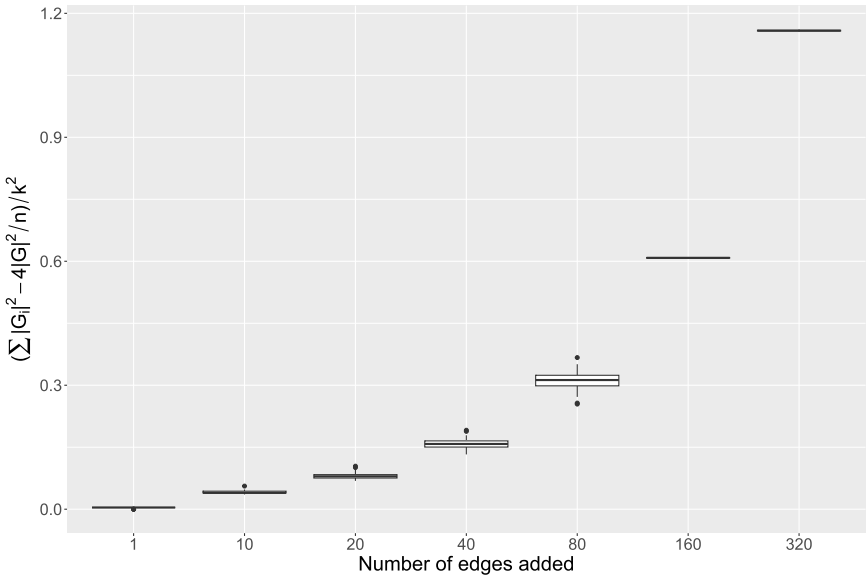

FIG 6. *Boxplots of* $(\sum_i |G_i|^2 - \frac{4|G|^2}{n})/k^2$ *for graphs under the same settings as in Figure 5.*

the modulus of continuity ω' becomes smaller and the magnitude of the jumps decreases. Moreover, as we increase the number of edges, Figure 6 shows that the condition $\sum_i |G_i|^2 - \frac{4|G|^2}{n}/k^2$ also improves.

6. Conclusion

We demonstrate that the graph-based stochastic processes are tight under mild conditions on the graph. To establish this, we utilize the Kolmogorov-Chentsov tightness criterion, which requires bounding higher moments. These moments are derived using combinatorial analysis, with their leading terms bounded by studying the graph configurations. This completes the framework for establishing the limiting distribution of these graph-based processes, facilitating p-value approximations for real applications. Our framework for establishing tightness can be applied to other scenarios where higher moments can be derived and bounded. For example, the rank-based change-point setting studied in [22] extends the existing graph-based framework by applying weights, in the form of ranks, to each edge in the similarity graph. For illustration, consider the 5-NN graph, where each observation is connected by edges to its five nearest neighbors. Each edge is then weighted by importance, with the first nearest neighbor (which contains the most relevant similarity information) assigned the highest rank. Based on these graph-induced ranks, the (weighted) graph-based test statistics $Z_w(t)$ and $Z_{\text{diff}}(t)$ are constructed. The explicit first and second moments of these rank-based test statistics are derived in [22]. Using combinatorial analysis, higher product moments can also be derived, following a similar approach to what we outline in this manuscript, albeit with more careful treatment of the ranks. To establish tightness for the rank-based stochastic processes, one can demonstrate that the higher moments are bounded, though this may require a different set of conditions on the graph.

We also conjecture that the corresponding stochastic random field under the changed-interval alternative is tight. Bickel and Wichura [2] provide a multidimensional analogue of Billingsley's definition of tightness for multi-parameter stochastic process. Specifically, let $T = T_1 \times T_2$, and let block B in T be of the form $(s_1, t_1] \times (s_2, t_2]$. Similarly, let block C in T be of the form $(t_1, u_1] \times (s_2, t_2]$, where $s_1 < t_1 < u_1$. For block B, define

$$X(B) = X(s_1, s_2) - X(t_1, s_2) - X(s_1, t_2) + X(t_1, t_2),$$

and for block C

$$X(C) = X(t_1, s_2) - X(u_1, s_2) - X(t_1, t_2) + X(u_1, t_2).$$

A sufficient condition for tightness is

$$E(|X(B)|^{\gamma_1}|X(C)|^{\gamma_2}) \leq \mu(B)^{\beta_1}\mu(C)^{\beta_2},$$

where $\gamma_1, \gamma_2, \beta_1$, and β_2 satisfy $\gamma_1 + \gamma_2 > 0$ and $\beta_1 + \beta_2 > 1$, and μ is a finite nonnegative measure of T.

We conjecture that the tightness of the graph-based scan statistics under the changed-interval alternative setting also hold. However, proving this is not straightforward due to the complexity of bounding the higher moments in this version, as the two-dimensional process introduces additional challenges. This line of research, which involves studying the increments of X around B, is reserved for future work.

Appendix A: Proof of Theorem 3.1

Under the condition for Theorem 3.1, we have $k^2n^4 \asymp (kn^2 - \sum_{i=1}^n |G_i|^2)^2$, and then $\frac{k^2n^4}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \asymp O(1)$.

For x_{14}, we have

$$\begin{aligned}
x_{14} =& \sum_i \sum_{j\neq i} (|G_i \setminus \{j \in G_i\}|)(|G_i \setminus \{j \in G_i\}| - 1)\times \\
& (|G_j \setminus \{i \in G_j\}|)(|G_j \setminus \{i \in G_j\}| - 1) \\
& < \sum_i \sum_{j\neq i} |G_i|^2 |G_j|^2 \\
=& |G_1|^2 \sum_{j\neq 1} |G_j|^2 + |G_2|^2 \sum_{j\neq 2} |G_j|^2 + \ldots + |G_n|^2 \sum_{j\neq n} |G_n|^2 \\
=& |G_1|^2 (\sum_{i=1}^n |G_i|^2 - |G_1|^2) \\
& + |G_2|^2 (\sum_{i=1}^n |G_i|^2 - |G_2|^2) + \ldots + |G_n|^2 (\sum_{i=1}^n |G_i|^2 - |G_n|^2) \\
=& (\sum_{i=1}^n |G_i|^2)^2 - \sum_{i=1}^n |G_i|^4 > 0.
\end{aligned}$$

Then $x_{14} < (\sum_{i=1}^n |G_i|^2)^2$ and $\frac{x_{14}}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$. Following similar arguments, since $\sum_{i=1}^n |G_i|^2 = o(kn^2)$, we have $\frac{\sum_{i=1}^n |G_i|^4}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$ and $\frac{n\sum_{i=1}^n |G_i|^3}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$.

For $\sum_i \sum_{j\in G_i; j\neq i} (|G_i| - 1)^2(|G_j| - 1)$, we have

$$\sum_i \sum_{j\in G_i; j\neq i} (|G_i| - 1)^2(|G_j| - 1) < 2|G| \sum_{i=1}^n |G_i|^3.$$

Since the the largest $|G_i|$ can be is $n-1$ (every other observation connects to node $\mathbf{y}_i$), it follows that $2|G| \sum_{i=1}^n |G_i|^3 \precsim 2|G|kn^3 \asymp k^2n^4$ and $\frac{k^2n^4}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$.

Similarly, since $kn^2 \sum_{i=1}^n |G_i|^2 \precsim k^2n^4$, we have $\frac{kn^2 \sum_{i=1}^n |G_i|^2}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$.

We have $\sum_{(i,j)\in G}(|G_i|-1)(|G_j|-1) < \sum_{i=1}^n |G_i|(|G|-|G_i|) = |G|\sum_{i=1}^n |G_i| - \sum_{i=1}^n |G_i|^2 < 2|G|^2 \asymp 2k^2n^2$, and so $\frac{\sum_{i,j\in G}(|G_i|-1)(|G_j|-1)}{(kn^2-\sum_{i=1}^n |G_i|^2)^2} \precsim O(1)$.

Finally, since

$$
\begin{aligned}
x_7 &< \sum_{i=1} |G_i|(|G| - |G_i|) < 2|G|^2 \asymp k^2n^2,\\
x_8 &= \sum_{(i,j),(j,l),i\neq l} |\{m : (i,m),(l,m) \in G\} \precsim kn^3,\\
x_9 &= \sum_{(i,j)} \sum_{l:(i,l),(j,l)\in G} (|G_l| - 2) \asymp k^2n^2.
\end{aligned}
$$

it follows that the ratio of the these configurations with $(kn^2 - \sum_{i=1}^n |G_i|^2)^2$ are bounded asymptotically by $O(1)$.

Appendix B: Proof of Theorem 3.2

1. $K_1(u,v,w)$: The leading coefficient for k^4n^2 can be expanded as

$$
\begin{aligned}
K_1(u,v,w) &= C_{d,1}(w-v)^2 + C_{d,2}(v-u)(w-v)\\
&\quad + C_{d,3}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})(w-v)\\
&\quad + C_{d,4}\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_w})\Big(\sqrt{u(1-v)}(\sqrt{v(1-u)} - \sqrt{u(1-v)})\\
&\quad - 2\sqrt{v(1-u)}(\sqrt{v(1-u)} - \sqrt{u(1-v)})\Big)
\end{aligned}
$$

with

$$
\begin{aligned}
C_{d,1} &= 64v(1-v)\sqrt{e_u}(u(1-v) + 2v(1-u) - 3\sqrt{e_ue_v}),\\
C_{d,2} &= 32\sqrt{e_u}(12u-8)(v-u)^3\\
&\quad + 192\sqrt{e_u}(8u^2 - 9u + 2 + \sqrt{e_u}(2\sqrt{e_v} - \sqrt{e_w}))(v-u)^2\\
&\quad + 32(2\sqrt{e_u}(36u^3 - 57u^2 + 24u - 2) - 18u\sqrt{e_v}(1-u)(1-2u)\\
&\quad + 2u\sqrt{e_w}(1-u)(5-9u) + 2\sqrt{e_u}\sqrt{e_v}\sqrt{e_w}(3u-2))(v-u)\\
&\quad - 64u\sqrt{e_u}(24u^2 - 25u + 5)(1-u) + 192u\sqrt{e_v}(1-u)(6u^2 - 6u + 1)\\
&\quad - 64u\sqrt{e_w}(1-u)(9u^2 - 10u + 2) + 32\sqrt{e_u}\sqrt{e_v}\sqrt{e_w}(12u^2 - 14u + 3),\\
C_{d,3} &= 64e_u(\sqrt{e_w}(3u-2) - 3\sqrt{e_u}(1-2u)),\\
C_{d,4} &= 64e_v\sqrt{e_u}.
\end{aligned}
$$

It is clear that $C_{d,1}(w-v)^2 + C_{d,2}(v-u)(w-v) \leq C(w-u)^2$ since $C_{d,1}(w-v)^2 + C_{d,2}(v-u)(w-v) \leq (C_{d,1} + C_{d,2})(w-u)^2$ and C can be chosen to be large enough such that $C_{d,1} + C_{d,2} \leq C$. In the following we focus on the next two terms. For the third term, we need to show that $\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v}) \leq (v-u)$. Let $\delta = v - u$ and define

$$
g(\delta) = \sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})
$$

$$= \sqrt{u(1-u)}\left(\sqrt{u(1-u)} - \sqrt{(u+\delta)(1-u-\delta)}\right)$$

which is continuous everywhere on $0 \leq \delta \leq 1-u$.
If $g(\delta)$ is convex for $0 \leq \delta \leq 1-u$, it follows that $g(\delta) \leq \delta$. Since $g(0) = 0$ and $g(1-u) = u(1-u) \leq 1(-u)$, what remains is to check its second derivative is non-negative:

$$g'(\delta) = \frac{-(1-2u-2\delta)\sqrt{u(1-u)}}{2\sqrt{(u+\delta)(1-u-\delta)}},$$
$$g''(\delta) = \frac{\sqrt{u(1-u)}}{2}\left(\frac{2}{\sqrt{(u+\delta)(1-u-\delta)}} + \frac{(1-2u-2\delta)^2}{2\sqrt{(u+\delta)(1-u-\delta)^3}}\right),$$

and it follows that $g''(\delta) > 0$. Since we have established that $g(\delta) = \sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})$ is convex, it follows that $\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v}) \leq (v-u)$ and $\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_w}) \leq (w-v)$. Moreover, the minimum of $g(\delta)$ is achieved when $\delta = 0.5 - u$ and $-g(0.5-u) = \sqrt{e_u}(\frac{1}{2} - \sqrt{e_u}) \leq \frac{1}{2} - u$, for $u < \frac{1}{2}$. Therefore $|\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})| \leq (v-u)$.
Following a similar argument, we can establish that $\sqrt{u(1-v)}(\sqrt{v(1-u)} - \sqrt{u(1-v)}) \leq (v-u)$.
Let $h(\delta) = \sqrt{(u+\delta)(1-u)}\left(\sqrt{(u+\delta)(1-u)} - \sqrt{u(1-u-\delta)}\right)$. We have $h(0) = 0$ and $h(1-u) = 1-u$. Its first and second derivatives are

$$h'(\delta) = (1-u) - \frac{(1-2u-2\delta)\sqrt{u(1-u)}}{2\sqrt{(1-u-\delta)(u+\delta)}},$$
$$h''(\delta) = \frac{\sqrt{u(1-u)}}{2}\left(\frac{2}{\sqrt{(1-u-\delta)(u+\delta)}} + \frac{(1-2u-2\delta)^2}{2\sqrt{(1-u-\delta)(u+\delta)}^3}\right),$$

and since $h''(\delta) > 0$, therefore $\sqrt{v(1-u)}(\sqrt{v(1-u)} - \sqrt{u(1-v)}) \leq (v-u)$. Since $\sqrt{u(1-v)} < \sqrt{v(1-u)}$, it follows that $\sqrt{u(1-v)}(\sqrt{v(1-u)} - \sqrt{u(1-v)}) \leq (v-u)$. Note that $\sqrt{v(1-u)} - \sqrt{u(1-v)} > 0$.
Therefore, $K_1(u,v,w) \leq C(w-u)^2$ for some constant C.

2. $K_2(u,v,w)$: The leading coefficient for $k^2 n(\sum_{i=1}^{n} |G_i|^2)$ is

$$\begin{aligned} K_2(u,v,w) = &\sqrt{e_u}\sqrt{e_v}\sqrt{e_w}\Big\{C_{d,5}(v-u)^2 + C_{d,6}(w-u)(v-u) \\ &+ C_{d,7}(\sqrt{e_u} - \sqrt{e_v})(v-u) \\ &+ C_{d,8}(\sqrt{e_u} - \sqrt{e_w})(v-u) + C_{d,9}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})^2 \\ &+ C_{d,10}(\sqrt{e_u} - \sqrt{e_v})(\sqrt{u(1-w)}(\sqrt{w(1-u)} - \sqrt{u(1-w)}) \\ &- 2\sqrt{w(1-u)}(\sqrt{w(1-u)} - \sqrt{u(1-w)}))\Big\} \end{aligned}$$

with

$$C_{d,5} = 16(-6uw + 4w + 2u - 1)$$

$$
\begin{aligned}
C_{d,6} =& 16(-12u^2 + 14u - 6\sqrt{e_u}\sqrt{e_v} - 3)\\
C_{d,7} =& 16(-2\sqrt{e_w}(2-3u) - 2\sqrt{e_u}(1-3u))\\
C_{d,8} =& 32(3u-2)\sqrt{e_u}\\
C_{d,9} =& -96\sqrt{e_w}\\
C_{d,10} =& 32\sqrt{e_u}
\end{aligned}
$$

Since $u < v < w$, we have $C_{d,5}(v_u)^2 + C_{d,6}(w-u)(v-u) \leq C(w-u)^2$ for some constant C. In order to show that remaining terms can also be bounded by $C(w-u)^2$, we follow that same argument detailed above for $K_1(u,v,w)$. Observe that $|\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})| \leq (v-u)$ and $|\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_w})| \leq (w-u)$. It follows that terms with $C_{d,7}, C_{d,8}$, and $C_{d,9}$ of $K_2(u,v,w)$ can be by bounded by $C(w-u)^2$ as well.
Finally, for the last term in $K_2(u,v,w)$, we see that $\sqrt{u(1-w)}(\sqrt{w(1-u)} - \sqrt{u(1-w)}) \leq (w-u)$ and $\sqrt{w(1-u)}(\sqrt{w(1-u)} - \sqrt{u(1-w)}) \leq (w-u)$.
It follows that $K_2(u,v,w) \leq C(w-u)^2$ for some constant C.

3. $K_3(u,v,w)$: The leading coefficient for $\sum_{i=1}^{n} |G_i|^4$ is

$$
\begin{aligned}
K_3(u,v,w) =& C_{d,11}(w-v)^2 + C_{d,12}(v-u)(w-v)\\
&+ C_{d,13}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})(w-v)\\
&+ C_{d,14}\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_u})(w-v)\\
&+ C_{d,15}\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_w})\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})\\
&+ C_{d,16}\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_w})(v-u)
\end{aligned}
$$

with

$$
\begin{aligned}
C_{d,11} =& -2v\sqrt{e_u}(3u + 5v - 8uv + 6uv^2 - 4v^2 - 2)\\
&- 4e_u\sqrt{e_v}(3v^2 - 3v + 1),\\
C_{d,12} =& 8\sqrt{e_u}(2-3u)(v-u)^3(w-v),\\
&- (4\sqrt{e_u}(24u^2 - 27u + 7) + 24e_u\sqrt{e_v} + 12e_u\sqrt{e_w})(v-u)^2(w-v),\\
&+ (-2\sqrt{e_u}(72u^3 - 114u^2 + 56u - 9) + 36u\sqrt{e_v}(2u^2 - 3u + 1),\\
&- 4u\sqrt{e_w}(9u^2 - 14u + 5) - 4\sqrt{e_u}\sqrt{e_v}\sqrt{e_w}(3u-2))(v-u)(w-v),\\
&- 2\sqrt{e_u}(48u^4 - 98u^3 + 70u^2 - 21u + 2)\\
&+ 4u\sqrt{e_v}(18u^3 - 36u^2 + 23u - 5),\\
&- 2u\sqrt{e_w}(18u^3 - 38u^2 + 25u - 5)\\
&- \sqrt{e_u}\sqrt{e_v}\sqrt{e_w}(24u^2 - 28u + 7),\\
&- 2\sqrt{e_w}(1-u-v)(6u^3 - 10u^2 + 5u - 1),\\
C_{d,13} =& -4\sqrt{e_u}(14u^2 - (6u-1)(u^2 + u + 1)),\\
C_{d,14} =& 2\sqrt{e_w}(6u^3 - 10u^2 + 5u - 1),\\
C_{d,15} =& 4\sqrt{e_v}(3v^2 - 3v + 1),
\end{aligned}
$$

$$C_{d,16} = 2v\sqrt{e_u}(6v^2 - 8v + 3).$$

Again, the first two terms involving $C_{d,11}$ and $C_{d,12}$ can be bounded by $C(w-u)^2$. Repeating the convexity argument, $|\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_u})| \leq (v - u)$, which allow us to bound the remaining terms by $C(w-u)^2$ as well. Therefore, the entire expression $K_3(u, v, w)$ can also be bounded by $C(w-u)^2$.

4. $K_4(u, v, w)$: The leading coefficient for $k\sum_{i=1}^{n} |G_i|^3$ is

$$\begin{aligned}
&C_{d,17}(w-v)^2 + C_{d,18}(w-v)(v-u) + C_{d,19}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})(w-v) \\
&+ C_{d,20}(v-u)^2 + C_{d,21}\sqrt{e_u}(\sqrt{e_v} - \sqrt{e_w})(v-u) \\
&+ C_{d,22}e_u(\sqrt{e_u} - \sqrt{e_v})(\sqrt{e_v} - \sqrt{e_w})
\end{aligned}$$

with

$$\begin{aligned}
C_{d,17} =& \sqrt{e_u}(16uv(12v^2 - 16v + 5) - 16v(8v^2 - 9v + 2)) \\
&+ 32e_u\sqrt{e_v}(6v^2 - 6v + 1), \\
C_{d,18} =& (128\sqrt{e_u}(3u - 2))(v-u)^3 \\
&+ (32\sqrt{e_u}(48u^2 - 54u + 13) + 384e_u\sqrt{e_v} - 192e_u\sqrt{e_w})(v-u)^2 \\
&+ (16\sqrt{e_u}(144u^3 - 228u^2 + 104u - 13) - 576\sqrt{e_v}u(2u^2 - 3u + 1) \\
&+ 64\sqrt{e_w}u(9u^2 - 14u + 5) + 64\sqrt{e_u}\sqrt{e_v}\sqrt{e_w}(3u - 2))(v-u) \\
&+ 16\sqrt{e_u}(96u^4 - 196u^3 + 130u^2 - 31u + 2), \\
C_{d,19} =& 8(\sqrt{e_u}(1-2u)(4(1-6u(1-u))) + 2\sqrt{e_w}(1-u)(1-8u+3u^2)), \\
C_{d,20} =& (\sqrt{e_v} - \sqrt{e_w})(-192e_u(v-u)^2 \\
&+ 64(6u - 2\sqrt{e_u}\sqrt{e_v} - 18u^2 + 12u^3 + 3\sqrt{e_u}\sqrt{e_v}u)(v-u) \\
&- 32e_u(7 - 36u(1-u)) + 16\sqrt{e_u}\sqrt{e_v}(9 - 40u + 36u^2)), \\
C_{d,21} =& 32\sqrt{e_u}(1-2u)(1-12u+12u^2) \\
&+ 16\sqrt{e_v}(u(36u^2 - 56u + 23) - 2 + 5u - 14u^2 + 9u^3) \\
C_{d,22} =& 32\sqrt{eu}(1 - 6u + 6u^2)
\end{aligned}$$

The first two terms involving $C_{d,17}$, $C_{d,18}$, and $C_{d,20}$ can be bounded by $C(w-u)^2$. Repeating a combination of the convexity arguments from above, the remaining terms can also be bounded by $C(w-u)^2$. It follows that $K_4(u, v, w) \leq C(w-u)^2$.

5. $K_5(u, v, w)$: The leading coefficient for x_{14} is

$$\begin{aligned}
K_5(u, v, w) =& C_{d,23}(w-v)^2 + C_{d,24}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})(w-v) \\
&+ C_{d,25}(v-u)(w-v) \\
&+ C_{d,26}\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_w})(v-u) \\
&+ C_{d,27}\sqrt{e_v}(\sqrt{e_v} - \sqrt{e_w})\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v}).
\end{aligned}$$

with

$$\begin{aligned}
C_{d,23} =& 4v(1-v)((u+2v-3uv)\sqrt{e_u} - 3\sqrt{e_v}u(1-u)) \\
C_{d,24} =& -2\sqrt{e_v}((6u\sqrt{e_w} - 6\sqrt{e_u} + 12\sqrt{e_u}v)(\sqrt{e_u} - \sqrt{e_v}) \\
& -2(1-2v)(u+2v-3uv) + 12u(1-u)(1-2v) \\
& +2\sqrt{e_u}\sqrt{e_w}(-6u+3v-2)) \\
C_{d,25} =& -2\sqrt{e_u}\sqrt{e_v}((4\sqrt{e_w} - 8\sqrt{e_u} + 12\sqrt{e_u}u)(v-u) \\
& +2\sqrt{e_u}(2u-1)(3u-2) + \sqrt{e_w}(8u-3) \\
& -6\sqrt{e_u}^2\sqrt{e_w}) \\
C_{d,26} =& -4\sqrt{e_v}((-3u^2+3u)(v-u) + 2\sqrt{e_u}\sqrt{e_v} \\
& -3u + 9u^2 - 6u^3 - 3u\sqrt{e_u}\sqrt{e_v}) \\
C_{d,27} =& 12\sqrt{e_v}u(1-u).
\end{aligned}$$

and utilizing the arguments above, this term is also bounded by $C(w-u)^2$.

6. $K_6(u,v,w)$: The leading coefficient for $\sum_i \sum_{j\in G_i; j\neq i}(|G_i|-1)^2(|G_j|-1)$ is

$$K_6(u,v,w) = C_{d,28}(v-u) + C_{d,29}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v})$$

with

$$\begin{aligned}
C_{d,28} =& 16\sqrt{e_u}\sqrt{e_w}(\sqrt{e_u} + 2\sqrt{e_w} - 3\sqrt{e_w}u - 3\sqrt{e_u}w)(v-u)^2 \\
& -8\sqrt{e_u}\sqrt{e_w}(-2\sqrt{e_u}(3u+5w-9uw-1) \\
& +\sqrt{e_v}(2u+4w-6uw-1) \\
& +2\sqrt{e_w}(9u^2-10u+2) + 6\sqrt{e_u}\sqrt{e_v}\sqrt{e_w})v-u) \\
& -8\sqrt{e_w}(-2u(1-u)w(-9u^2+10u-2) + 2u^2(1-u)(2-3u) \\
& -\sqrt{e_u}\sqrt{e_v}(3u+3w-14uw+12u^2w-4u^2) \\
& +2\sqrt{e_u}\sqrt{e_w}u(9u^2-14u+5) \\
& -6\sqrt{e_v}\sqrt{e_w}u(2u^2-3u+1)), \\
C_{d,29} =& -\sqrt{e_u}(48uw(1-u)(1-w) + 16\sqrt{e_u}\sqrt{e_w}(u(1-w) + 2w(1-u))).
\end{aligned}$$

We have that $C_{d,28}(v-u)$ can be bounded by a constant $C(w-u)$ and by convexity, $C_{d,29}\sqrt{e_u}(\sqrt{e_u} - \sqrt{e_v}) \leq C(w-u)$.
Therefore, the leading coefficient $K_6(u,v,w)$ is bounded by $C(w-u)$.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

Lynna Chu was supported in part by NSF Grant DMS-1513653. Hao Chen was supported in part by NSF Grants DMS-1513653, DMS-1848579, and DMS-2311399.

Supplementary Material

Supplementary Materials for "On the Tightness of Graph-based Statistics"
(doi: 10.1214/25-EJS2367SUPP; .pdf). Supplement A: Comparison of asymptotic and permutation critical values for different values of n, n_0, and n_1. This supplement includes numerical comparisons of the asymptotic and permutation critical values of the graph-based scan statistics for different values of n, n_0, and n_1. Supplement B: Examples of derivation of product moments. This supplement illustrates the derivation of the probability for each of the 19 configurations. Supplement C: Exact analytical expressions for a list of product moments of R_1 and R_2. This supplement provides analytical expressions for all 86 product moments of R_1 and R_2.

References

[1] Aue, A., Hörmann, S., Horváth, L., Reimherr, M. et al. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics* **37** 4046–4087. MR2572452
[2] Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics* **42** 1656–1670. MR0383482
[3] Billingsley, P. (1999). *Convergence of probability measures*. 2nd edition, John Wiley & Sons. MR1700749
[4] Chen, H. Chen, X. and Su, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association* **113** 1146–1155. MR3862346
[5] Chen, H. and Chu, L. (2023). Graph-based change-point analysis. *Annual Review of Statistics and Its Application* **10** 475–499. MR4567802
[6] Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association* **112** 397–409. MR3646580
[7] Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *The Annals of Statistics* **43** 139–176. MR3285603
[8] Chentsov, N. N. (1956). Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov-Smirnov tests. *Theory of Probability & Its Applications* **1** 140–144.

[9] Chu, L. and Chen, H. (2019). Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data. *The Annals of Statistics* **47** 382–414. MR3910545

[10] Chu, L. and Chen, H. (2025). Supplement to "On the Tightness of Graph-based Statistics". DOI: https://doi.org/10.1214/25-EJS2367SUPP.

[11] Dubey, P. and Müller, H.-G. (2020). Fréchet change-point detection. *The Annals of Statistics* **48** 3312–3335. MR4185810

[12] Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76** 495–580. MR3210728

[13] Garreau, D. and Arlot, S. (2018). Consistent change-point detection with kernels. *Electronic Journal of Statistics* **12** 4440–4486. MR3892345

[14] Harchaoui, Z., Moulines, E. and Bach, F. R. (2009). Kernel change-point analysis. In *Advances in Neural Information Processing Systems* 609–616.

[15] Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* **109** 334–345. MR3180567

[16] Padilla, C. M. M., Xu, H., Wang, D., Padilla, O. H. M. and Yu, Y. (2023). Change point detection and inference in multivariable nonparametric models under mixing conditions. *arXiv preprint arXiv:2301.11491*.

[17] Padilla, O. H. M., Yu, Y., Wang, D. and Rinaldo, A. (2021). Optimal nonparametric multivariate change point detection and localization. *IEEE Transactions on Information Theory* **68** 1922–1944. MR4395506

[18] Shi, X., Wu, Y. and Rao, C. R. (2018). Consistent and powerful non-Euclidean graph-based change-point test with applications to segmenting random interfered video data. *Proceedings of the National Academy of Sciences* 201804649. MR3827579

[19] Wang, D., Yu, Y. and Rinaldo, A. (2021). Optimal change point detection and localization in sparse dynamic networks. *The Annals of Statistics* **49** 203–232. MR4206675

[20] Wang, T. and Samworth, R. J. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 57–83. MR3744712

[21] Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645. MR2672488

[22] Zhou, D. and Chen, H. (2022). Asymptotic Distribution-free Change-point Detection for Modern Data Based on a New Ranking Scheme. *arXiv preprint arXiv:2206.03038*.

[23] Zou, C., Wang, G. and Li, R. (2020). Consistent selection of the number of change-points via sample-splitting. *Annals of statistics* **48** 413. MR4065168