The Annals of Statistics
2019, Vol. 47, No. 3, 1381–1407
https://doi.org/10.1214/18-AOS1718

SEQUENTIAL CHANGE-POINT DETECTION BASED ON NEAREST NEIGHBORS

BY HAO CHEN[1]

University of California, Davis

We propose a new framework for the detection of change-points in online, sequential data analysis. The approach utilizes nearest neighbor information and can be applied to sequences of multivariate observations or non-Euclidean data objects, such as network data. Different stopping rules are explored, and one specific rule is recommended due to its desirable properties. An accurate analytic approximation of the average run length is derived for the recommended rule, making it an easy off-the-shelf approach for real multivariate/object sequential data monitoring applications. Simulations reveal that the new approach has better performance than likelihood-based approaches for high dimensional data. The new approach is illustrated through a real dataset in detecting global structural changes in social networks.

1. Introduction. Sequential change-point models are widely used in many fields to detect events of interest as data are generated. One of its early applications is in quality control where a summary statistic reflecting a manufacture process is monitored over time. When the statistic begins to exhibit values that are unlikely to be achieved by random fluctuations, there is a high probability that something went wrong and an investigation is needed. Therefore, it is important to detect the change-point, the time when the event of interest happens, as soon as possible if it occurs, while keeping the false discovery rate low; refer to monographs Wald (1973), Siegmund (1985) and Tartakovsky, Nikiforov and Basseville (2015) for more background information.

Sequential change-point detection has been extensively studied for univariate data, that is, for data where the observations are scalar at each time point. However, many recent applications involve the detection of change-points over a sequence of multivariate, or even non-Euclidean, observations. Following are some motivating examples.

Multiple sensor framework: In a sensor network, hundreds or thousands of sensors are deployed to detect events of interest. For example, hundreds of monitors are placed worldwide to detect solar flares, which are large energy releases by the

Received February 2017; revised April 2018.
[1]Supported in part by NSF Grant DMS-1513653.
MSC2010 subject classifications. Primary 62G32; secondary 60K35.
Key words and phrases. Change-point, sequential detection, graph-based tests, nonparametrics, scan statistic, tail probability, high-dimensional data, network data, non-Euclidean data.

sun that can affect Earth's ionosphere and disrupt long-range radio communication [Kappenman (2012), Qu et al. (2005)]. Often, the structure of the sensor network can be used to boost the power of the detection. Then each observation can be viewed as a vector with some structures among its elements that reflect the spacial information of the sensors.

Social network evolution: Technological advances provide us with rich resources of social network data, such as networks constructed by Facebook friendship relations, email communications, phone calls or online chat records. The detection of abrupt events, such as shifts in network connectivity, dissociation of communities, or formation of new communities, can be formulated as a change-point problem. Here, the observation at each time point is a graphical encoding of the network.

Epidemic disease outbreak: It is important to detect the emergence of new infectious diseases as early as possible to prevent their spreadings. In the United States, the current practice is that the Centers for Disease Control gathers data from hospitals and then integrates information together to tell if there is an outbreak. This process usually takes weeks to draw conclusions. Researchers have tried to incorporate other information, such as online searches on disease related topics and climate information, which had success in shortening the prediction lag time for flu outbreaks [Yang, Lipsitch and Shaman (2015), Yang, Santillana and Kou (2015)]. It can be foreseen that, in the future, information from multiple sources will be used to predict disease outbreaks. Then each observation can be quite complicated and may include hospital admission rates, online search frequencies on related topics, personal posts on related symptoms and whether information.

Image analysis: Image data are collected over time in many areas. It is of tremendous interest to automatically detect abrupt events, such as security breaches from surveillance videos or extreme weather conditions, for example, storms, from climatology. In these applications, the data at each time point is the digital encoding of an image.

In all of these examples, the problem can be formulated in the following way: We denote the data sequence by $\{\mathbf{Y}_i\}$, $i = 1, 2, \ldots, n, \ldots$, indexed by time or some other meaningful orderings. Here, $\mathbf{Y}_i$'s can be vectors, networks or images. The sequence is identically distributed as F_0 until a time τ the distribution changes abruptly to F_1:

$$\mathbf{Y}_i \sim F_0, \qquad i = 1, \ldots, \tau - 1,$$
$$\mathbf{Y}_i \sim F_1, \qquad i = \tau, \tau + 1, \ldots,$$

where F_0 and F_1 are two different probability measures.

There is a burst of works recently on the change-point detection in multiple sequences where the sequences are assumed to be independent, such as in multiple sensor framework where the sensors are assumed to be indenpendent. These works also in general assume the observations over time are independent. Some

nice algorithms and theorems have been developed under these assumptions. For example, Tartakovsky and Veeravalli (2008) and Mei (2010) studied statistics that sum signals over all streams with further assumptions that the density functions before and after the change are known and the change happens to all streams at the same time. Siegmund (2013) and Chan and Walther (2015) allow the change only happen to a subset of the data streams under the assumption that F_0 and F_1 are multivariate normal distributions with identity covariance. The latter paper also studied the optimality of several statistics. These statistics are useful if the assumptions under which the statistic was developed hold for the data. However, in many applications, it would be too stringent to assume that all data streams are independent.

To another end, the change-point detection problem for dynamic network data is gaining more and more attention. A number of works have been done if the networks are generated in some specific ways. For example, Heard et al. (2010) developed Bayesian methods by modeling each pair of nodes independently and they modeled the communications between nodes over time as a counting process with the increments of the process following a Bayesian probability model. A multinomial extension that relaxed the independence assumption among pairs of nodes was also studied by the authors. Wang et al. (2014) considered the setting that the series of networks are generated by a stochastic block model with the block membership of the vertices fixed across time. They use locality-based scan statistic to find change-point where the connectivity probability matrix varies. Again, these methods are useful if the data do satisfy the assumptions, while these assumptions could be too specific for many applications.

In this paper, we describe a nonparametric framework to approach the problem. This framework can be applied to data in arbitrary dimension and to non-Euclidean data, with a general, analytic formula for false discovery control. The proposed method adopts the idea of making use of similarity graphs, such as nearest neighbors, among the observations in Chen and Zhang (2015).

In the following, we do not impose specific assumptions on F_0 or F_1. However, we assume the observations over time are independent. When there is weak dependence over time, the graph-based approach could still provide meaningful results for change-point analysis as shown in Chen and Zhang (2015). Also, the independence assumption is a natural starting point for more sophisticated models that consider dependency over time.

Chen and Zhang (2015) studied the problem of *offline* change-point detection, where all observations are completely observed at the time when data analysis is conducted. However, in many applications, it is desirable to detect change-points on the fly. There are both theoretical and computational challenges to extend the method in Chen and Zhang (2015) to the *online* framework. In particular, adding new observations usually changes the similarity structure among existing observations, when the most similar observation for an existing observation may be changed to the newest observation. This makes the theoretical analysis on false

discovery control much harder as it requires an analysis of the dynamics of similarity structural change when new observations are added.

In this paper, we consider the similarity structure represented by nearest neighbors (NN). We studied the dynamics in NN updates as new observations are added. It turns out that the characterization of a small number of events, in particular, the updates of mutual NNs and shared NNs, and all three-way interactions among the NN relations, could capture the majority of the dynamics (see Section 5 for details). This makes the task tractable. We can also easily implement the method for real data applications.

The rest of the paper is organized as follows: In Section 2, we briefly review a two-sample test based on NNs, which is a building block for the change-point analysis. In Section 3, we discuss details of the proposed detection method and three stopping rules. We recommend the use of the stopping rule that relies on recent observations for its desirable properties. In Section 4, we study the updating dynamics of NNs and derive an analytic formula for false discovery control that is accurate for finite samples. In Section 5, we compare the proposed method to parametric methods for multivariate data. We illustrate the proposed method on a real dataset in Section 6. In Section 7, we briefly discuss the choice of the number of nearest neighbors, the performance of the proposed method on gradual changes, and possible extensions of the method to other similarity graphs.

2. A brief review of the two-sample test on k-NN. In this section, we review the two-sample test on k-NN proposed by Schilling (1986) and Henze (1988). Here, k is a fixed integer. Let k-NN be the directed graph with the pooled observations as the nodes and each node points to its first k NNs. It is assumed that the observations are distinct with uniquely defined neighbors. (This happens with probability 1 if $\mathbf{Y}_i$'s follow continuous multivariate distributions and the Euclidean distance is used.)

Let $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_1}\}$ and $\{\mathbf{Y}_{n_1+1}, \ldots, \mathbf{Y}_{n_1+n_2}\}$ be random samples from two populations, and let $n = n_1 + n_2$ be the total sample size. For any event x, let $\mathbf{I}(x)$ be the indicator function that takes value 1 if x is true or 0 if otherwise. Let

$$b_{ij} = \mathbf{I}\big((i \le n_1, j > n_1) \text{ or } (i > n_1, j \le n_1)\big),$$

then b_{ij} is the indicator function that $\mathbf{Y}_i$ and $\mathbf{Y}_j$ belong to different samples. We want to test whether these two population distributions are the same or not. Let

$$A_{ij}^{(r)} = \mathbf{I}(\mathbf{Y}_j \text{ is the } r\text{th nearest neighbor of } \mathbf{Y}_i), \qquad A_{ij}^{+} = \sum_{r=1}^{k} A_{ij}^{(r)}.$$

Then A_{ij}^{+} is the indicator function that $\mathbf{Y}_j$ is among the first k NNs of $\mathbf{Y}_i$. We have $A_{ij}^{+} \in \{0, 1\}$ and $\sum_{j=1}^{n} A_{ij}^{+} = k$, $1 \le i \le n$. Then

$$\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^{+} b_{ij} \equiv \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ji}^{+} b_{ij}$$

is the number of edges in the k-NN that connect between the two samples.

Expressing in a more symmetric way, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(A_{ij}^{+}+A_{ji}^{+})b_{ij} \tag{2.1}$$

being twice the number of edges in the k-NN that connect between the two samples. Given the observations $\mathbf{Y}_i = \mathbf{y}_i$, $1 \le i \le n$, the test statistic is

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}^{+}+a_{ji}^{+})b_{ij},$$

where $a_{ij}^{+} = \sum_{r=1}^{k} a_{ij}^{(r)}$ with $a_{ij}^{(r)} = \mathbf{I}(\mathbf{y}_j \text{ is the } r\text{th nearest neighbor of } \mathbf{y}_i)$. In Schilling (1986) and Henze (1988), the authors proposed to reject the null hypothesis of no difference if the test statistic is significantly *smaller* than its expectation under the permutation null distribution. The rationale is that, if the two samples are from the same distribution, they are well mixed and are likely to find their nearest neighbors from the other sample. So if the observations tend to not find nearest neighbors from the other sample, they are from different distributions.

We denote the random variable under the permutation distribution as follows: Let $B_{ij} = b_{\mathbf{P}(i)\mathbf{P}(j)}$ be the indicator function that $\mathbf{Y}_i$ and $\mathbf{Y}_j$ belong to different samples under random permutation. Here, $\mathbf{P}(i)$ is the index of $\mathbf{Y}_i$ under permutation. Let

$$X = \sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}^{+}+a_{ji}^{+})B_{ij}. \tag{2.2}$$

Then its expectation and variance are

$$\begin{aligned}
\mathsf{E}(X) &= \frac{4kn_1n_2}{n-1},\\
\mathsf{Var}(X) &= \frac{4n_1n_2}{n-1}\left(h(n_1,n_2)\left(\frac{1}{n}\sum_{i,j=1}^{n}a_{ij}^{+}a_{ji}^{+}+k-\frac{2k^2}{n-1}\right)\right.\\
&\qquad \left.+\bigl(1-h(n_1,n_2)\bigr)\left(\frac{1}{n}\sum_{i,j,l=1}^{n}a_{ji}^{+}a_{li}^{+}-k^2\right)\right),
\end{aligned}$$

where $h(n_1,n_2) = \frac{4(n_1-1)(n_2-1)}{(n-2)(n-3)}$. It has been shown that

$$\frac{X-\mathsf{E}(X)}{\sqrt{\mathsf{Var}(X)}}$$

converges to the standard normal distribution under the null hypothesis as long as $n_1/n_2 \to \lambda \in (0,\infty)$ for multivariate data [Schilling (1986), Henze (1988)].

3. Sequential change-point detection based on k-NN. We use

$$\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n, \ldots$$

to denote the data sequence, where $\mathbf{Y}_n$ is the observation at time n. In the following, we assume that we have a well defined norm $\|\cdot\|$ on the sample space such that the distance between two observations $\mathbf{y}_i$ and $\mathbf{y}_j$ can be calculated as $d(\mathbf{y}_i, \mathbf{y}_j) = \|\mathbf{y}_i - \mathbf{y}_j\|$. We also assume that the observations are distinct points in the sample space and have uniquely defined nearest neighbors. In the following, k is fixed. The choice of k is briefly discussed in Section 8.

We assume that there are N_0 historical observations with no change-point. That is, $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{N_0}$ follow the same distribution. This can be determined from prior information or we can use offline change-point detection methods to test whether there is any change-point among the first N_0 observations, such as the method in Chen and Zhang (2015). We begin our test from observation $N_0 + 1$.

For any n, $1 \le i, j \le n$, let

$$A_{n,ij}^{(r)} = \mathbf{I}(\mathbf{Y}_j \text{ is the } r\text{th NN of } \mathbf{Y}_i \text{ among the first } n \text{ observations}).$$

Then $A_{n,ij}^{+} = \sum_{r=1}^{k} A_{n,ij}^{(r)}$ is the indicator function that $\mathbf{Y}_j$ is one of the first k NNs of $\mathbf{Y}_i$ among the first n observations.

We can perform a two-sample test for each $t \in \{1, \ldots, n-1\}$ with one sample being the observations before t and the other sample being the observations between t and n. Define

$$b_{ij}(t, n) = \mathbf{I}\big((i \le t, t < j \le n) \text{ or } (t < i \le n, j \le t)\big),$$

and $B_{ij}(t, n) = b_{\mathbf{P}_n(i)\mathbf{P}_n(j)}(t, n)$, where $\mathbf{P}_n(\cdot)$ is a random permutation among the first n indices. Let

$$R(t, n) = \sum_{i=1}^{n} \sum_{j=1}^{n} \big(A_{n,ij}^{+} + A_{n,ji}^{+}\big) B_{ij}(t, n).$$

We use $\mathbf{y}_i$'s to denote the realizations of $\mathbf{Y}_i$'s, and let

$$Z_{|\mathbf{y}}(t, n) = -\frac{R(t, n) - \mathsf{E}(R(t, n))}{\sqrt{\mathsf{Var}(R(t, n)|\mathbf{y})}}.$$

Note that $\mathsf{E}(R(t, n)|\mathbf{y}) = \mathsf{E}(R(t, n))$.

If a change-point $\tau > N_0$ occurs in the sequence, we would expect $Z_{|\mathbf{y}}(t, n)$ to be *large* (notice the negative sign in the standardization) when $n > \tau$ and t close to τ. In the following, we consider three stopping rules:

$$T_1(b_1) = \inf\Big\{n - N_0 : \Big(\max_{n_0 \le t \le n - n_0} Z_{|\mathbf{y}}(t, n)\Big) > b_1, n \ge N_0\Big\}, \tag{3.1}$$

$$T_2(b_2) = \inf\Big\{n - N_0 : \Big(\max_{n - n_1 \le t \le n - n_0} Z_{|\mathbf{y}}(t, n)\Big) > b_2, n \ge N_0\Big\}, \tag{3.2}$$

$$T_3(b_3) = \inf\Big\{n - N_0 : \Big(\max_{n-n_1 \le t \le n-n_0} Z_{L|\mathbf{y}}(t,n)\Big) > b_3, n \ge N_0\Big\}. \tag{3.3}$$

Here, b_1, b_2 and b_3 are chosen so that the false discovery rate for each of the stopping rule is controlled at a prespecified level.

In the above stopping rules, n_0, n_1 and L are prespecified values. Usually, n_0 is set to be small so as to detect the change as soon as possible, while not too small, such as 1, to avoid the high fluctuations at the very ends. So T_1 is a straightforward stopping rule. Sometimes, when τ is large, we may not want to put too much emphasizes on the early observations. This leads to T_2 and T_3. It is easy to see that T_2 is a more relaxed version of T_1. In T_2, if we set n_1 to be $n - n_0$, then it is the same as T_1, while we could set n_1 tactically to achieve performance similar to (or even better than) T_1 and at the same reduce computation time.

For both T_1 and T_2, at time n, we find k NNs among the first n observations. One modification we can make is that we use the most recent observations to compute the test statistic. In T_3, $Z_{L|\mathbf{y}}(t,n)$ is defined the same as $Z_{|\mathbf{y}}(t,n)$ but only based on the L most recent observations: $\mathbf{Y}_{n-L+1}, \ldots, \mathbf{Y}_n$. That is, for $i, j \in n_L \triangleq \{n - L + 1, \ldots, n\}$, we let

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

$A^{+}_{n_L,ij} = \sum_{r=1}^{k} A^{(r)}_{n_L,ij}$, and $R_L(t,n) = \sum_{i,j \in n_L} (A^{+}_{n_L,ij} + A^{+}_{n_L,ji}) B_{ij}(t, n_L)$ with $B_{ij}(t, n_L) = b_{\mathbf{P}_{n_L}(i)\mathbf{P}_{n_L}(j)}(t)$, where $\mathbf{P}_{n_L}(\cdot)$ is a random permutation among indices $\{n - L + 1, \ldots, n\}$. Then

$$Z_{L|\mathbf{y}}(t,n) = -\frac{R_L(t,n) - \mathsf{E}(R_L(t,n))}{\sqrt{\mathsf{Var}(R_L(t,n)|\mathbf{y})}}.$$

3.1. *Comparisons of the three stopping rules.* Two key objectives of sequential detection are (i) to detect the change-point as soon as possible when it occurs; and (ii) to keep the false discovery rate low. These can be characterized by two quantities: The expected detection delay, $\mathsf{E}_{\tau^*}(T - \tau^* | T > \tau^*)$, where $\tau^* = \tau - N_0$ is the time index of the change-point if we set the time we begin the test to be 1; and the average run length, $\mathsf{E}_\infty(T)$, the expectation of T when there is no change-point or the change-point is at infinity.

In the following, we use Monte Carlo simulations to better understand the three stopping rules. To make a fair comparison, the critical values $b_i, i = 1, 2, 3$ are chosen (through simulation runs) so that $\mathsf{E}_\infty(T_i) = 2000$ for each stopping rule. We then compare their detection delays. The detailed simulation setup is as follows: There are $N_0 = 200$ historical observations from the same distribution. We begin our test from $t = 201$. In the simulation, the change-point is at τ. Before the change-point τ, the distribution is a d-dimensional Gaussian distribution with mean μ_1 and covariance matrix I_d, $\mathcal{N}_d(\mu_1, I_d)$; after the change, the distribution is $\mathcal{N}_d(\mu_2, I_d)$. Let $\|\mu_2 - \mu_1\|_2 = 2$ where $\|\cdot\|_2$ is the L_2 norm. We consider

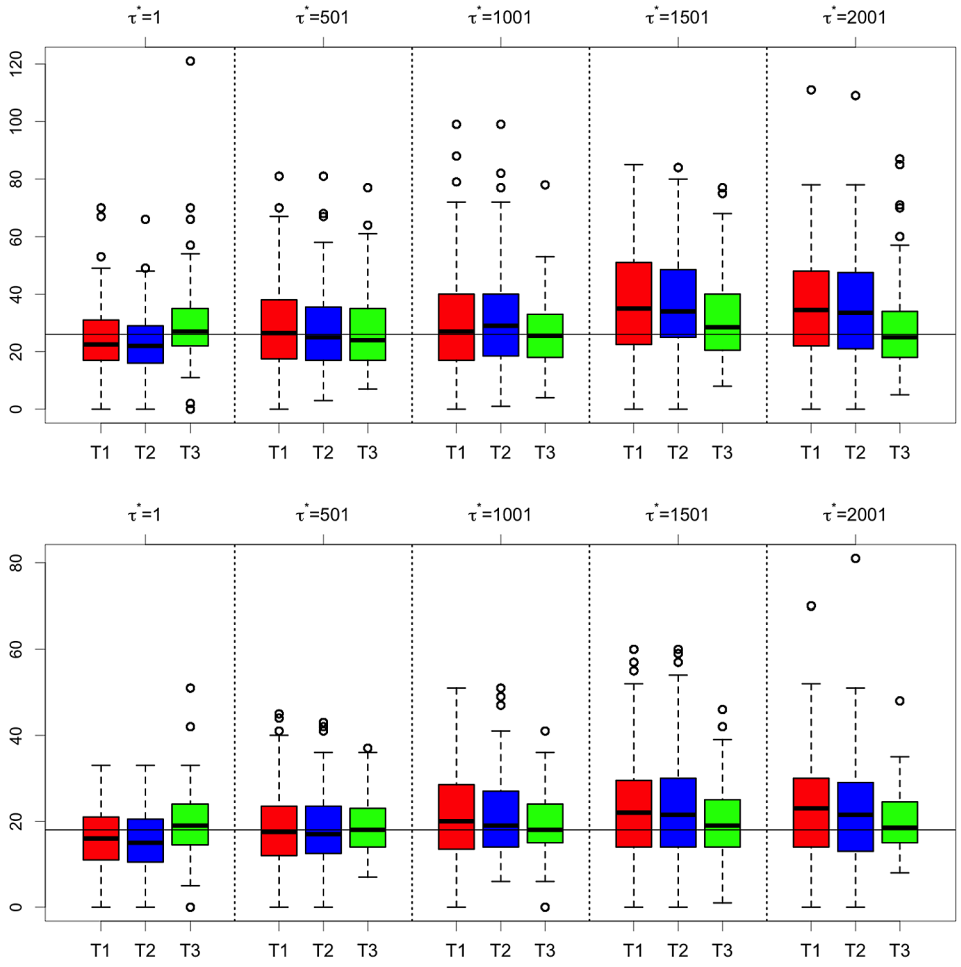

FIG. 1. *Boxplots of detection delays of the three stopping rules based on* 1000 *simulation runs for each* τ^*. *Top panel*: $k = 1$; *bottom panel*: $k = 3$. *Other parameters are set as*: $n_0 = 3, n_1 = 197$ *and* $L = 200$. *The horizontal line is the median of the detection delays for* T_3 *across all* 5000 *simulation runs*.

$\tau = 201$ (the change occurs right at the time when we begin to perform the test, that is, $\tau^* = 1$) till $\tau = 2201$ (the change occurs 2000 observations after we begin to perform the test, i.e., $\tau^* = 2001$) for an increment of 500. We consider 1-NN and 3-NN graphs.

Figure 1 shows boxplots of the detection delays ($T - \tau^*$) of the three stopping rules under different τ^*'s. Here, we aim for shorter detection delays. We can see that T_2 is in general slightly better than T_1 as the boxes are shifted downward a little bit overall. When τ^* is small, T_3 has a longer detection delay than T_1 or T_2. As τ^* increases, the detection delay for T_3 is almost the same, while that for T_1

or T_2 increases substantially. When $\tau^* = 1501$, the detection delay for T_1 or T_2 is clearly larger than T_3. One reason for the increasing detection delay for T_1 or T_2 is that $Z_{|\mathbf{y}}(t, n)$ is left skewed when the ratio t/n is small and this problem becomes severer as n increases.

On the other hand, since T_3 is based on the same number of observations for all n, its detection delay is not affected by where τ^* locates. Its detection delay is longer than T_1 and T_2 when the change occurs at a very early stage, but it is on par with T_1 or T_2 when the change occurs later, and shorter than T_1 and T_2 when the change occurs in a late stage. As the first work on sequential detection based on k-NN graphs, we recommend to use T_3. For T_1 and T_2, one way to overcome the problem of increasing detection delay is to make the thresholds in T_1 and T_2 to be functions of n; for example, we could consider $T_1(b_1(n))$ and $T_2(b_2(n))$ with $b_1(n)$ and $b_2(n)$ monotone increasing functions in n. This is, however, a large topic, and we reserve it for future studies.

In the following, if not further noted, T and b refer to T_3 and b_3, respectively.

4. Average run length $\mathsf{E}_\infty(T(b))$. Given the stopping rule $T(b)$, the remaining question is how to determine the detection threshold b, in particular, how to choose b so that the average run length $\mathsf{E}_\infty(T(b))$ is a prespecified value, such as 10,000.

First of all, we usually do not know the underlying distribution of the observations, so we could not directly simulate observations to obtain b as done in Section 3.1. Second, resampling based methods, such as permutation and bootstrap, are not appropriate here as new observations keep arriving and the limited existing observations are usually not representative enough, especially for complicated data. Even if one could come up with some approaches through resampling methods, they would be very time consuming and not practical for online applications. Therefore, we seek to obtain an analytic formula for $\mathsf{E}_\infty(T(b))$.

Given the nonparametric nature of the proposed method, we would not be able to get an exact analytic formula for $\mathsf{E}_\infty(T(b))$ for finite L, the number of observations used at each time, so we approach the problem asymptotically, that is, $L \to \infty$. We then make further modifications so that the analytic formula is a good approximation for finite L.

4.1. *Asymptotic results.* We first consider the asymptotic scenario, $L \to \infty$. In this context, $\{Z_{L|\mathbf{y}}(t, n)\}_{t,n}$, with t and n rescaled by L, can be shown to converge to a two-dimensional Gaussian random field under very mild conditions. The properties of the supremum of a two-dimensional Gaussian random field was well studied [Siegmund and Venkatraman (1995)], and the remaining task is to quantify the covariance function of the Gaussian random field, as well as its partial derivatives. They can be obtained by studying the dynamics of the NN relations. The main results are given in Lemma 4.1 and Theorems 4.2 and 4.4.

We assume the following condition.

CONDITION 1. There is a positive constant $\mathbb{C}$, $1 \le \mathbb{C} < \infty$, depending only on k, such that

$$\sup_{1\le j\le n}\left(\sum_{i=1}^{n} A^{+}_{n,ij}\right) \le \mathbb{C}, \qquad n \in \mathbb{N}.$$

In k-NN, each observation points to its first k NNs, so the out-degree of each observation (the number of arrows pointing from the observation) is k, while the in-degree of each observation (the number of arrows pointing to the observation) can vary. This condition says that the in-degree of each observation is bounded. It is satisfied almost surely for multivariate data [Bickel and Breiman (1983), Henze (1988)]. For non-Euclidean data, if the distance is chosen properly, this condition is also easy to hold as many non-Euclidean data can be embedded into a Euclidean space.

Before stating the main results, we define some useful quantities. According to Propositions 3.1 and 3.2 in Henze (1988), under Condition 1, the quantities

$$\frac{1}{L}\sum_{i,j\in n_L} A^{(r)}_{n_L,ij}A^{(s)}_{n_L,ji}, \qquad \frac{1}{L}\sum_{i,j,l\in n_L, j\ne l} A^{(r)}_{n_L,ji}A^{(s)}_{n_L,li},$$

converge in probability to constants as $L \to \infty$ and the limits can be calculated through complicated integrals [Henze (1988)]. We denote the limits as

$$p_\infty(r,s) = \lim_{L\to\infty}\frac{1}{L}\sum_{i,j\in n_L} A^{(r)}_{n_L,ij}A^{(s)}_{n_L,ji}, \tag{4.1}$$

$$q_\infty(r,s) = \lim_{L\to\infty}\frac{1}{L}\sum_{i,j,l\in n_L, j\ne l} A^{(r)}_{n_L,ji}A^{(s)}_{n_L,li}. \tag{4.2}$$

Let

$$p_{k,\infty} = \sum_{r=1}^{k}\sum_{s=1}^{k} p_\infty(r,s), \tag{4.3}$$

$$q_{k,\infty} = \sum_{r=1}^{k}\sum_{s=1}^{k} q_\infty(r,s). \tag{4.4}$$

Then $p_{k,\infty}$ is the limiting expected number of mutual NNs a node has in k-NN and $q_{k,\infty}$ the limiting expected number of nodes that share a NN with a node in k-NN. We also define their finite sample versions by taking expectations

$$p_L(r,s) = \frac{1}{L}\mathsf{E}\Big(\sum_{i,j\in n_L} A^{(r)}_{n_L,ij}A^{(s)}_{n_L,ji}\Big), \tag{4.5}$$

$$q_L(r,s) = \frac{1}{L}\mathsf{E}\Big(\sum_{i,j,l\in n_L, j\ne l} A^{(r)}_{n_L,ji}A^{(s)}_{n_L,li}\Big), \tag{4.6}$$

$$p_{k,L} = \frac{1}{L}\mathsf{E}\Big(\sum_{i,j\in n_L} A^+_{n_L,ij}A^+_{n_L,ji}\Big), \tag{4.7}$$

$$q_{k,L} = \frac{1}{L}\mathsf{E}\Big(\sum_{i,j,l\in n_L, j\neq l} A^+_{n_L,ji}A^+_{n_L,li}\Big). \tag{4.8}$$

Then

$$\lim_{L\to\infty} p_L(r,s) = p_\infty(r,s), \qquad \lim_{L\to\infty} q_L(r,s) = q_\infty(r,s),$$
$$\lim_{L\to\infty} p_{k,L} = p_{k,\infty}, \qquad \lim_{L\to\infty} q_{k,L} = q_{k,\infty}.$$

We next state the main results.

LEMMA 4.1. *Under Condition 1, when* $t-(n-L), (n-t) = O(L)$, *as* $L\to\infty$, $Z_{L|\mathbf{y}}(t,n) \to Z_L(t,n)$ *almost surely, where*

$$Z_L(t,n) = -\frac{R_L(t,n) - \mathsf{E}(R_L(t,n))}{\sqrt{\mathsf{Var}(R_L(t,n))}}.$$

This lemma follows immediately from Propositions 3.1 and 3.2 in Henze (1988).

THEOREM 4.2. *Under Condition 1, as* $L\to\infty$, *the finite dimensional distributions of* $\{Z_L([vL],[wL]) : 0 < w-1 < v < w < \infty\}$ *weakly converges to the finite dimensional distributions of a two-dimensional Gaussian random field, which we denote as* $\{Z^\star(v,w) : 0 < w-1 < v < w < \infty\}$. (*Here,* $[x]$ *denotes the largest integer smaller than or equal to* x *for any real number* x.)

A main challenge to prove this theorem is how to deal with the holistic dependencies among $A^+_{n_L,ij}$'s. Even for different i,j,l,r, $A^+_{n_L,ij}$ and $A^+_{n_L,lr}$ are dependent. This is because of the constraints $\sum_j A^+_{n_L,ij} = k$ for all $i\in n_L$ [see details in the Supplementary Material, Chen (2019), Section A.1].

We consider a similar set of Bernoulli random variables $\{\widetilde{A}^+_{n_L,ij}\}_{i,j\in n_L}$ but with relaxed dependencies. We keep the following probabilities unchanged:

$$\mathsf{P}\big(\widetilde{A}^+_{n_L,ij}=1\big) = \mathsf{P}\big(A^+_{n_L,ij}=1\big),$$
$$\mathsf{P}\big(\widetilde{A}^+_{n_L,ij}=1, \widetilde{A}^+_{n_L,ji}=1\big) = \mathsf{P}\big(A^+_{n_L,ij}=1, A^+_{n_L,ji}=1\big),$$
$$\mathsf{P}\big(\widetilde{A}^+_{n_L,ji}=1, \widetilde{A}^+_{n_L,li}=1\big) = \mathsf{P}\big(A^+_{n_L,ji}=1, A^+_{n_L,li}=1\big).$$

That is, two-way NN relations are retained. However, we relax the other dependencies. We let $\widetilde{A}^+_{n_L,ij}$ be independent of $\{\widetilde{A}^+_{n_L,il}, \widetilde{A}^+_{n_L,li}\}_{l\neq j}$. We also let $\widetilde{A}^+_{n_L,ij}$ and $\widetilde{A}^+_{n_L,lr}$ be independent when i,j,l,r are all different.

Then $\widetilde{A}^{+}_{n_L,ij}$'s are only locally dependent. But $\sum_j \widetilde{A}^{+}_{n_L,ij}$'s are no longer necessarily k. However, $\{\widetilde{A}^{+}_{n_L,ij}\}_{i,j\in n_L}$ becomes $\{A^{+}_{n_L,ij}\}_{i,j\in n_L}$ if we condition on the events $\{\sum_j \widetilde{A}^{+}_{n_L,ij} = k\}_{i\in n_L}$. Thus, $Z_L(t,n)$ can be studied through the joint distribution of summations of locally dependent terms. We then use Stein's method to deal with local dependencies. The complete proof is in the Supplementary Material [Chen (2019), Section A.1].

REMARK 4.3. The tightness of the two-dimensional field can be shown for $\{Z_L([vL],[wL]) : 0 < w-1+\delta \le v \le w-\delta < \infty\}$ for any $\delta \in (0,1)$. For v too close to $w-1$ or w, the fluctuation in the random field could be too wild to have the field being uniformly tight.

Based on Theorem 4.2, we approximate $\mathsf{E}_\infty(T(b))$ by that of the corresponding quantity defined for the limiting random field:

$$T^\star(b) = \inf\Big\{n - N_0 : \Big(\max_{n-n_1\le t\le n-n_0} Z^\star(t/L, n/L)\Big) > b, n \ge N_0\Big\}. \tag{4.9}$$

According to Siegmund and Venkatraman (1995), when $b, L, n_0, n_1 \to \infty$ in such a way that $b = c\sqrt{L}$ for some fixed $0 < c < \infty$, $n_0 = u_0 L$ and $n_1 = u_1 L$ for some fixed $0 < u_0 < u_1 < 1$, and when there is no change-point, $T^\star(b)$ is asymptotically exponentially distributed with mean

$$\mathsf{E}_\infty\big(T^\star(b)\big) \sim \frac{\sqrt{2\pi}\exp(b^2/2)}{c^2 b\int_{u_0}^{u_1} g_1(u)g_2(u)\nu(c\sqrt{2g_1(u)})\nu(c\sqrt{2g_2(u)})\,du}, \tag{4.10}$$

where

$$g_1(u) = \frac{\partial_-\rho^\star_{(u,w)}(\delta_1,0)}{\partial\delta_1}\bigg|_{\delta_1=0} \equiv -\frac{\partial_+\rho^\star_{(u,w)}(\delta_1,0)}{\partial\delta_1}\bigg|_{\delta_1=0},$$

$$g_2(u) = \frac{\partial_-\rho^\star_{(u,w)}(0,\delta_2)}{\partial\delta_2}\bigg|_{\delta_2=0} \equiv -\frac{\partial_+\rho^\star_{(u,w)}(0,\delta_2)}{\partial\delta_2}\bigg|_{\delta_2=0},$$

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

Here, $\rho^\star_{(u,w)}(\delta_1,\delta_2) = \mathsf{Cov}(Z^\star(w-u,w), Z^\star(w-u+\delta_1, w+\delta_2))$ and $\nu(\cdot)$ is closely related to the Laplace transform of the overshoot over the boundary of a random walk. A simple approximation given in Siegmund and Yakir (2007) is sufficient for numerical purpose:

$$\nu(x) \approx \frac{(2/x)(\Phi(x/2) - 0.5)}{(x/2)\Phi(x/2) + \phi(x/2)}, \tag{4.11}$$

where $\Phi(\cdot)$ is the cumulate distribution function of the standard normal distribution and $\phi(\cdot)$ the density function of the standard normal distribution.

Thus, the remaining task is to derive the directional partial derivatives of the covariance function of the Gaussian random field. Their analytic expressions are given in the following theorem.

THEOREM 4.4. *For the two-dimensional field* $\{Z^\star(v,w) : 0 < w-1 < v < w < \infty\}$, *the directional partial derivatives are*

$$
\begin{aligned}
g_1(u) &= \frac{\partial_- \rho^\star_{(u,w)}(\delta_1,0)}{\partial \delta_1}\bigg|_{\delta_1=0} \equiv -\frac{\partial_+ \rho^\star_{(u,w)}(\delta_1,0)}{\partial \delta_1}\bigg|_{\delta_1=0} \\
&= \frac{16u(1-u)(k+p_{k,\infty}) + 2(1-2u)^2(q_{k,\infty}-k^2+k)}{\sigma^2(u)}, \qquad (4.12)\\
g_2(u) &= \frac{\partial_- \rho^\star_{(u,w)}(0,\delta_2)}{\partial \delta_2}\bigg|_{\delta_2=0} \equiv -\frac{\partial_+ \rho^\star_{(u,w)}(0,\delta_2)}{\partial \delta_2}\bigg|_{\delta_2=0} \\
&= \frac{16u^2(1-u)^2(p_{k,\infty}+q_{k,\infty}+k^2+2p^{(k)}_{k,\infty}-2q^{(k)}_{k,\infty})}{\sigma^2(u)} \qquad (4.13)\\
&\quad + \frac{4u(1-u)(2q^{(k)}_{k,\infty}-3q_{k,\infty}+k^2+k)+2(q_{k,\infty}-k^2+k)}{\sigma^2(u)},
\end{aligned}
$$

where

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

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

The complete proof of this theorem is in the Supplementary Material [Chen (2019), Section A.2]. We studied the dynamics of the k-NN series as new observations are added through combinatorial analysis and it turned out that a few key quantities are enough to characterize the dynamics in the asymptotic domain.

4.2. *Finite* L. We now consider the practical scenario where L is finite. Based on Theorems 4.2 and 4.4, $\mathsf{E}_\infty(T(b))$ can be approximated by

$$
\mathsf{E}_\infty\big(T(b)\big) \approx \frac{L\sqrt{2\pi}\exp(b^2/2)}{b^3 \int_{\frac{n_0}{L}}^{\frac{n_1}{L}} g_1(u)g_2(u)\nu\big(\sqrt{2b^2 g_1(u)/L}\big)\nu\big(\sqrt{2b^2 g_2(u)/L}\big)\,du}
$$

with the analytic expressions for $g_1(u)$ and $g_2(u)$ given in (4.12) and (4.13), respectively, and $\nu(\cdot)$ given in (4.11).

When deriving the limiting expressions for $g_1(u)$ and $g_2(u)$, we evaluate $\sum_j \mathsf{E}(A^{(r)}_{n_L,ij}A^{(s)}_{n_L,ji})$ and $\sum_{j\neq l}\mathsf{E}(A^{(r)}_{n_L,ji}A^{(s)}_{n_L,li})$ under $L\to\infty$ and the two quantities become $p_\infty(r,s)$ and $q_\infty(r,s)$, respectively. In practice, when L is finite,

$p_\infty(r,s)$ and $q_\infty(r,s)$ are not the best estimates for these two expectations, yet the expectations could be better estimated through historical data. Therefore, we use the following formula to approximate $\mathsf{E}_\infty(T(b))$ in practice:

$$\mathsf{E}_\infty\big(T(b)\big) \approx \frac{L\sqrt{2\pi}\exp(b^2/2)}{b^3\int_{\frac{n_0}{L}}^{\frac{n_1}{L}} g_{L,1}(u)g_{L,2}(u)v\big(\sqrt{2b^2g_{L,1}(u)/L}\big)v\big(\sqrt{2b^2g_{L,2}(u)/L}\big)\,du}, \tag{4.14}$$

where $g_{L,1}(u)$ and $g_{L,2}(u)$ are the same as $g_1(u)$ and $g_2(u)$, respectively, except that $p_{k,\infty}$, $q_{k,\infty}$, $p_{k,\infty}^{(k)}$ and $q_{k,\infty}^{(k)}$ are replaced by $p_{k,L}$, $q_{k,L}$, $p_{k,L}^{(k)}$ and $q_{k,L}^{(k)}$, respectively, with $p_{k,L}$ given in (4.7), $q_{k,L}$ given in (4.8), and

$$p_{k,L}^{(k)} = \sum_{r=1}^{k} p_L(k,r), \qquad q_{k,L}^{(k)} = \sum_{r=1}^{k} q_L(k,r). \tag{4.15}$$

For $p_{k,L}$, $q_{k,L}$, $p_{k,L}^{(k)}$ and $q_{k,L}^{(k)}$, they usually do not have analytical expressions. However, they can be easily estimated from historical data. These estimates can further be updated by new observations as long as no change-point is detected.

We next check how this analytic approximation works. We compare the threshold b such that $\mathsf{E}_\infty(T(b)) = 10{,}000$ based on this analytic approximation and that based on 10,000 Monte Carlo simulations. The threshold obtained through 10,000 Monte Carlo simulations can be regarded as the true threshold. Results under different choices of n_0, k and d for multivariate Gaussian data are shown in Table 1. We checked two values of L, namely $L = 200$ and $L = 50$, and let $n_1 = L - n_0$.

Unfortunately, the thresholds obtained through the analytic approximation (4.14) are not that close to the Monte Carlo results except for a few occasions. The analytic approximation (4.14) gives similar thresholds for different dimensions when all other parameters are fixed. However, the thresholds from Monte Carlo simulations are quite different for different dimensions with those under a higher dimension much smaller. Thus, (4.14) is still missing some major components for finite L due to the fact that $Z_L(t,n)$ can be quite left skewed for finite L and small $(n-t)$. In the following, we incorporate skewness of $Z_L(t,n)$ to improve the analytic approximation.

REMARK 4.5. The reason of the discrepancy between the asymptotic results and finite samples was discussed in details in the offline counterpart of the work [Chen and Zhang (2015)]. Briefly, the convergence rate of $Z_L(t,n)$ to the Gaussian distribution is slow if $(n-t)/L$ is close to 0 or 1. In this online detection setting, the problem is even severer as we would like to set n_0 very small (such as 3) so as to detect the change as soon as it happens. For finite L, $Z_L(t,n)$ is quiet left skewed when $(n-t)$ is close to 0 or L, and the tail probability is overestimated, making the threshold b obtained based on the asymptotic results too conservative.

TABLE 1

The threshold b, such that $\mathsf{E}_\infty(T(b)) = 10{,}000$, through 10,000 Monte Carlo simulations, through analytic formula (4.14) based on asymptotic results, and through analytic formula (4.17) with additional skewness correction. Each observation in the data sequence follows a d-dimensional normal distribution

		$n_0 = 3$			$n_0 = 10$		
		Monte Carlo	**Asymp. (4.14)**	**Skewness corrected (4.17)**	**Monte Carlo**	**Asymp. (4.14)**	**Skewness corrected (4.17)**
$L = 200$							
$d = 10$	$k = 1$	4.04	4.40	4.07	4.04	4.31	4.07
	$k = 3$	4.14	4.34	4.14	4.14	4.23	4.14
	$k = 5$	4.16	4.31	4.18	4.16	4.17	4.18
$d = 100$	$k = 1$	3.76	4.37	3.79	3.76	4.26	3.79
	$k = 3$	3.78	4.33	3.79	3.78	4.20	3.79
	$k = 5$	3.79	4.31	3.81	3.79	4.18	3.81
$d = 1000$	$k = 1$	3.73	4.38	3.73	3.73	4.28	3.73
	$k = 3$	3.71	4.33	3.71	3.71	4.21	3.71
	$k = 5$	3.75	4.32	3.72	3.75	4.18	3.72
$d = 10{,}000$	$k = 1$	3.71	4.38	3.70	3.71	4.27	3.70
	$k = 3$	3.65	4.33	3.69	3.65	4.21	3.69
	$k = 5$	3.68	4.32	3.69	3.68	4.18	3.69
$L = 50$							
$d = 10$	$k = 1$	4.00	4.38	4.10	3.99	4.24	4.10
	$k = 3$	4.36	4.32	4.37	4.36	4.19	4.37
	$k = 5$	4.57	4.28	4.50	4.57	4.15	4.50
$d = 100$	$k = 1$	3.86	4.36	3.94	3.83	4.23	3.94
	$k = 3$	3.92	4.31	4.02	3.92	4.18	4.02
	$k = 5$	3.95	4.29	4.09	3.95	4.15	4.09
$d = 1000$	$k = 1$	3.83	4.36	3.91	3.83	4.23	3.91
	$k = 3$	3.92	4.32	3.93	3.92	4.18	3.93
	$k = 5$	3.95	4.29	3.97	3.95	4.15	3.97
$d = 10{,}000$	$k = 1$	3.79	4.36	3.90	3.79	4.23	3.90
	$k = 3$	3.86	4.32	3.90	3.86	4.18	3.90
	$k = 5$	3.91	4.29	3.92	3.91	4.15	3.92

4.2.1. *Skewness correction.* We adapt the skewness correction approach in Chen and Zhang (2015). In particular, when we derive the average run length for the limiting two-dimensional Gaussian random field (4.10), the term in the integral related to the marginal distribution of $Z^\star(w-u, w)$ is $\mathsf{P}(Z^\star(w-u, w) \in b + du)$. (Here, du is the differential of the variable u, and similar definition for dt in the following.) To make the analytic approximation more accurate for fi-

nite L and small n_0, we replace $\mathsf{P}(Z^\star(w-u,w)=b+du)$ by an estimate of $\mathsf{P}(Z_L([n(w-u)],[nw])\in b+du)$. Following the method based on cumulant-generating functions and change of measure [details refer to Chen and Zhang (2015)], we have

$$\begin{aligned}&\frac{\mathsf{P}(Z_L(t,n)\in b+dt/b)}{\mathsf{P}(Z^\star(n/L-t/L,n/L)\in b+dt/b)}\\&\quad\approx\frac{\exp((b-\theta_b)^2/2+\theta_b^2\gamma_L(t,n)\theta_b/6)}{\sqrt{1+\gamma_L(t,n)\theta_b}}:=S_L\big((n-t)/L\big).\end{aligned}\tag{4.16}$$

Here, $\theta_b=(-1+\sqrt{1+2\gamma_L(t,n)b})/\gamma_L(t,n)$ and $\gamma_L(t,n)=\mathsf{E}(Z_L^3(t,n))$. The denotation for $S_L((n-t)/L)$ holds because $\gamma_L(t,n)$ relates to t and n only as a function of $n-t$ (see Lemma 4.6 below). Then, the analytic approximation for $\mathsf{E}_\infty(T)$ incorporating skewness becomes

$$\frac{L\sqrt{2\pi}\exp(b^2/2)}{b^3\int_{n_0/L}^{n_1/L}\mathbf{S_L(u)}g_{L,1}(u)g_{L,2}(u)\nu(\sqrt{2b^2g_{L,1}(u)/L})\nu(\sqrt{2b^2g_{L,2}(u)/L})\,du}.\tag{4.17}$$

When $L\to\infty$ and $(n-t),(L-(n-t))=O(L)$, $\gamma_L(t,n)$ goes to 0 and $S_L(u)$ goes to 1 for $0<u<1$, so this formula converges to (4.10) in the limit.

The exact analytic expression for $\gamma_L(t,n)$ is given in the following lemma.

LEMMA 4.6. *We have*

$$\gamma_L(t,n)=\frac{(\mathsf{E}(R_L(t,n)))^3+3\mathsf{E}(R_L(t,n))\mathsf{Var}(R_L(t,n))-\mathsf{E}(R_L^3(t,n))}{(\mathsf{Var}(R_L(t,n)))^{3/2}},$$

where $\mathsf{E}(R_L(t,n))$ *and* $\mathsf{Var}(R_L(t,n))$ *are given in* (A.3) *and* (A.4), *respectively, and*

$$\begin{aligned}\mathsf{E}\big(R_L^3(t,n)\big)&=8k^3L^3r_4+12k^2L^2\big(r_2+3k(r_2-2r_4)\big)\\&\quad+4kL\big(3r_2-r_1+2r_3-4r_4+3k(3r_1-2r_2-4r_3-4r_4)\\&\quad+8k^2(r_3-3r_2+5r_4)\big)\\&\quad+24p_{k,L}\big(kL^2r_4+kL(r_1+r_2-2r_3-4r_4)\\&\quad+2L(2r_3-r_1+2r_4)\big)\\&\quad+12q_{k,L}\big(kL^2(r_2-2r_4)+kL(2r_3-5r_2+8r_4)\\&\quad+L(r_1+r_2-2r_3-4r_4)\big)\\&\quad+4(2r_3-3r_2+4r_4)\mathsf{E}\Big(\sum_{i,j,l,v}A^+_{n_L,ji}A^+_{n_L,li}A^+_{n_L,vi}\Big)\\&\quad+24(r_1+r_2-2r_3-4r_4)\mathsf{E}\Big(\sum_{i,j,l}A^+_{n_L,ij}A^+_{n_L,ji}A^+_{n_L,li}\Big)\end{aligned}$$

$$+ 24(2r_4 - r_2)\mathsf{E}\Big(\sum_{i,j,l,v} A^+_{n_L,ij}A^+_{n_L,li}A^+_{n_L,vj}\Big)$$
$$- 16r_4\Big(\mathsf{E}\Big(\sum_{i,j,l} A^+_{n_L,ij}A^+_{n_L,jl}A^+_{n_L,li}\Big)$$
$$+ 3\mathsf{E}\Big(\sum_{i,j,l} A^+_{n_L,ij}A^+_{n_L,il}A^+_{n_L,jl}\Big)\Big)$$

with

$$r_1 = \frac{2x(L-x)}{L(L-1)}, \qquad x = L-(n-t),$$
$$r_2 = \frac{4x(x-1)(L-x)(L-x-1)}{L(L-1)(L-2)(L-3)},$$
$$r_3 = \frac{x(L-x)((x-1)(x-2)+(L-x-1)(L-x-2))}{L(L-1)(L-2)(L-3)},$$
$$r_4 = \frac{8x(x-1)(x-2)(L-x)(L-x-1)(L-x-2)}{L(L-1)(L-2)(L-3)(L-4)(L-5)}.$$

To prove this lemma, we have

$$\mathsf{E}\big(R^3_L(t,n)\big) = \mathsf{E}\big(\mathsf{E}\big(R^3_L(t,n)|\mathbf{Y}\big)\big)$$
$$= \sum_{i,j,l,r,u,v} \mathsf{E}\big((A^+_{n_L,ij}+A^+_{n_L,ji})(A^+_{n_L,lr}+A^+_{n_L,rl})(A^+_{n_L,uv}+A^+_{n_L,vu})\big)$$
$$\times \mathsf{E}\big(B_{ij}(t,n_L)B_{lr}(t,n_L)B_{lr}(t,n_L)\big).$$

Adapting similar arguments in calculating the covariance in the proof of Theorem 4.4 but with more careful treatment of the summation indices, we could get the result in the lemma.

From Lemma 4.6, we see that $\mathsf{E}(R^3_L(t,n))$ depends on the probability of having certain structures in the nearest neighbor graph. The relevant structures in k-NN are shown in Figure 2. The first two structures represent mutual NNs and shared NNs. The other five structures are three-way interactions among the NN relations. The probability of having each of them can be estimated through historical data, and can also be updated by new observations when no change-point is detected.

FIG. 2. *The configurations in k-NN that relate to the third moment of* $R_L(t,n)$.

We now check how skewness correction performs. Table 1 also lists the thresholds obtained through the analytic approximation with skewness correction. We see that, after skewness correction, the analytic formula gives much better estimates to the thresholds. When $L = 200$, all thresholds estimated by (4.17) are very accurate. Even for small L ($L = 50$), the analytic approximated with skewness correction is doing a reasonable job. When the dimension becomes larger, the threshold estimated by the analytic formula with skewness correction is smaller, exhibiting the same trend as the Monte Carlo results.

These results show that the formula with skewness correction could capture the major factors and gives quite reliable estimates. It would be reasonable to use the analytic formula with skewness correction to get the threshold b in real applications.

5. Power analysis. Given the procedure and the fast analytic way of determining the detection threshold, the proposed method can be easily applied to real problems. Now, the question is how powerful this method is. To get some idea, we compare it to the test based on Hotelling's T^2 test for multivariate Gaussian data as Hotelling T^2 test is asymptotically the most powerful for testing two multivariate Gaussian distributions with the same covariance matrix.

The simulation setup is as follows: There are $N_0 = 200$ historical observations and a change occurs at $t = 400$ (200 new observations after the start of the test). The observations are independent and follow d-dimensional Gaussian distribution with a mean shift (Δ) at the change-point. (The L_2 distance between the two means is Δ.) The amount of change, Δ, is chosen so that the tests have moderate power. Results are given in Table 2. "Successful detection" is defined the same as in Section 3.1 that the test detects the change-point within 100 observations after the change occurred. We compare all tests on the same ground by controlling the early stop probability to be 0.01.

Table 2 shows the results under different scenarios with 1000 simulation runs for each scenario. The fraction of the runs that the change-point is successfully

TABLE 2
Fraction of trials (out of 1000) that the change-point is successfully detected for the proposed test and for the test based on the Hotelling's T^2 test

	Normal data				**Log-normal data**	
	$d = 10$, $\Delta = 0.7$	$d = 100$, $\Delta = 1.8$	$d = 1000$, $\Delta = 2.7$	$d = 10{,}000$, $\Delta = 5$	$d = 10$, $\Delta = 1.5$	$d = 100$, $\Delta = 2$
Proposed test: 1-NN	0.02	0.21	0.12	0.16	0.48	0.08
Proposed test: 3-NN	0.07	0.55	0.41	0.52	0.87	0.48
Proposed test: 5-NN	0.15	**0.81**	**0.57**	**0.70**	**0.95**	**0.77**
Hotelling's T^2	**0.69**	0.63	–	–	0.34	0.02

detected is reported. When the data is multivariate Gaussian, we see that the test based on the Hotelling T^2 test is doing very well in low dimension. When the dimension becomes higher, the power of the proposed test catches up. When $d = 100$, the proposed test based on 5-NN is outperforming the test based on the Hotelling T^2 test. When d is even higher, the dimension is larger than the number of observations that the method based on the Hotelling's T^2 cannot be applied. For the proposed tests, we see that the we do need to increase the strength of the signal to achieve a similar detection power. However, the number of fold we need for the increase of the signal is much smaller than that for the dimension. When the dimension increase from 100 to 10,000 (by a fold of 100), we only need to increase the signal by a fold about 3 to achieve the same detection power. Hence, the proposed method is relatively mildly affected by the dimensionality.

We also did the comparison for log-normal data and the change is in the mean parameter. Now, the assumptions for the Hotelling T^2 test do not hold and we see that the proposed test is outperforming the test based on the Hotelling T^2 test even when the dimension is low.

The results show that the proposed test has satisfying power and works for various distributions.

6. An illustration example from real data. Here, we apply the proposed method to a real dataset on network analysis. The dataset has been completely collected at the time of analysis. We treat it as if the data were being observed to illustrate how the proposed method works. It is conceivable to apply the proposed method in a sequential manner if the data keep arriving.

The MIT Media Laboratory conducted a study following 106 subjects, students and stuff in an institute, who used mobile phones with preinstalled software that can record all activities on their phones from July 2004 to June 2005 [Eagle, Pentland and Lazer (2009)]. A natural question of interest is whether there is any change in the phone-call pattern among these people over time. This is one way to assess their friendship along time.

We bin the phone calls by day, and for each day, construct a phone-call network with the subjects as nodes and a directed edge pointing from subject i to subject j if subject i called subject j on that day. We encode the directed network of each day by an adjacency matrix, with 1 for element $[i, j]$ if there is a directed edge pointing from subject i to subject j, and 0 otherwise. Let M_i be the 106×106 adjacency matrix on day i. We consider two distance measures defined as:

(1) the number of different entries: $\|M_i - M_j\|_F^2$, where $\|\cdot\|_F$ means the Frobenius norm of a matrix,

(2) the number of different entries, normalized by the geometric mean of the total edges in each day: $\frac{\|M_i - M_j\|_F^2}{\|M_i\|_F \|M_j\|_F}$.

For this dataset, since there is no further information to tell whether there is any change-point for the first few observations, we applied the offline change-point

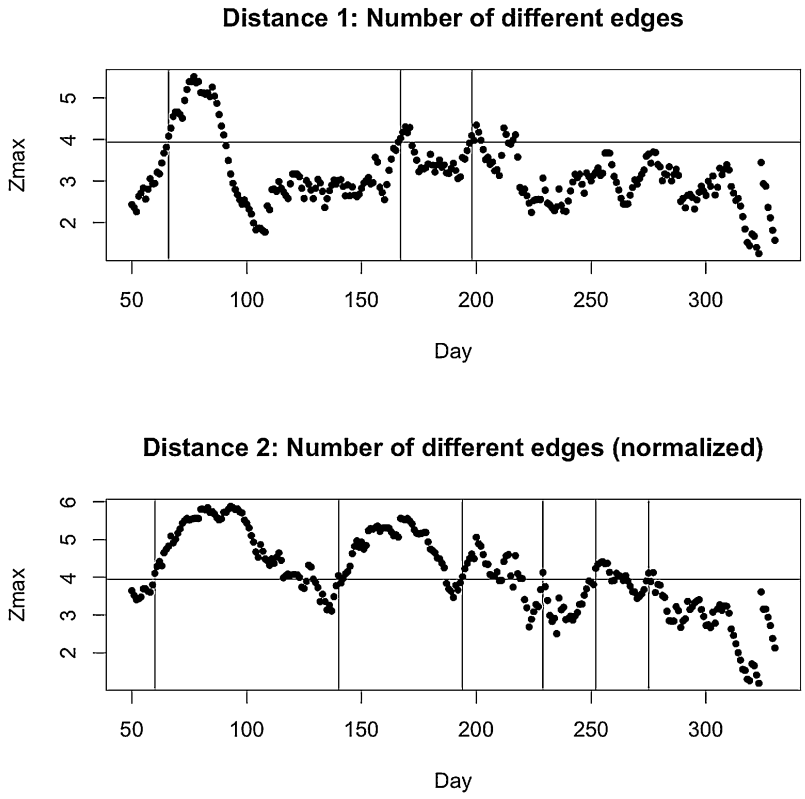

FIG. 3. *Zmax for the network data based on two distances. The horizontal line in each plot is the threshold b such that* $\mathsf{E}_\infty(T(b)) = 10{,}000$. *The vertical lines are the valid stopping times.*

detection method in Chen and Zhang (2015) on the first 50 days/observations. No change-point was found for either distance measure. So we treat the first 50 observations as historical observations. We let $L = 50$, $n_0 = 3$ and determine the threshold based on (4.17).

Figure 3 plots $\mathrm{Zmax}(n) = \max_{n-L+n_0 \le t \le n-n_0} Z_{L|\mathbf{y}}(t, n)$ against n, the index of days, based on the two distances. The detection thresholds for the two distances are $b = 3.92$ and $b = 3.98$, respectively. Since multiple stopping times might be called for one change-point, we disregard time n if $\max(\mathrm{Zmax}(n-5), \mathrm{Zmax}(n-4), \ldots, \mathrm{Zmax}(n-1)) > b$, that is, we consider them to be caused by the same event. We call the remaining stopping times the "candidate stopping times." Then three candidate stopping times for distance 1 and six candidate stopping times for distance 2 are found. They are summarized in Table 3, together with their nearby academic events.

From Table 3, we see that the proposed method based on either distance finds change-points at around the beginning of the Fall term, the end of the Fall term, and the beginning of the Spring term. The proposed method using distance 2 finds

TABLE 3
Valid stopping times and their nearby academic events

Distance 1	**Distance 2**	**Nearby academic event***
$n = 66$: 2004/9/23	$n = 60$: 2004/9/17	9/9: First day of class for Fall term
$n = 167$: 2005/1/2	$n = 140$: 2004/12/6	12/18: Last day of class for Fall term
$n = 198$: 2005/2/2	$n = 194$: 2005/1/29	2/2: First day of class for Spring term
–	$n = 229$: 2005/3/5	3/5: Registration deadline for Spring term
–	$n = 252$: 2005/3/28	3/21: Spring vacation
–	$n = 275$: 2005/4/20	4/21: Drop deadline for Spring term

*The dates of the academic events are from the 2015–2016 academic calendar of MIT as the 2004–2005 academic calendar of MIT cannot be found online.

additional change-points in the middle of the Spring term. These are all reasonable times to have some significant call pattern changes.

One may wonder if these change-points could be found by a 1-dimensional summary statistic. We plot in Figure 4 the number of edges in each network over time. We could see clearly the change-points at around the beginning of the Fall term and the end of the Fall term, reflected by the change of the call volume. Starting from the winter break ($n = 160$), the call volume stabilizes. There is a slight call volume decrease starting from the spring vacation (at around $n = 250$). However, the call volumes from $n = 160$ toward $n = 250$ are quite similar, and there is no significant change within this period. For example, we apply the function *cpt.meanvar()* in `R` package `changepoint`, a 1-dimensional change-point detection approach for detecting either mean or variance change, to this segment of data and no change-point is found. Hence, there is no significant change in the call volume transiting from the winter break to the Spring term.

On the other hand, the proposed method on either distance finds the change-point at the beginning of Spring term (around $n = 198$), indicating that there are some structural changes in the phone-call network which would not be captured by only examining the call volume. Also, since distance 2 is normalized by the

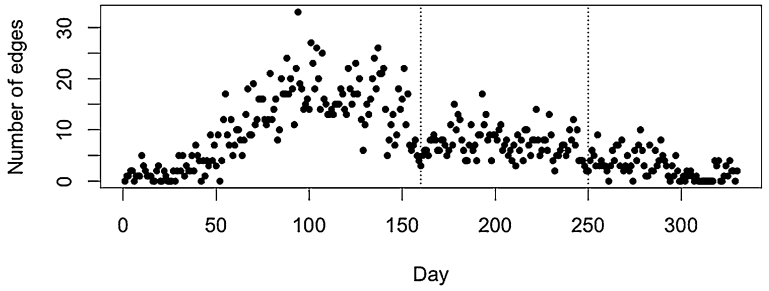

FIG. 4. *Number of edges in the phone-call network for each day.*

total number of edges in each network, there are probably structural changes in the phone-call network besides call volume change in the other five change-points detected based on distance 2.

For further details on what the changes are, one could pick some networks before and after the change-point and conduct more detailed comparisons. Moreover, if one is interested in some specific characteristics of the network, a distance reflecting such characteristics can be used for the proposed method.

REMARK 6.1. This phone-call network example is for illustration. The proposed method will be more useful in real data applications where data are indeed being collected (not completely collected at the time of the analysis) and the observations cannot be characterized through a simple model, such as a long vector with unknown structures among the elements, a combination of quantitative and qualitative components, a networks, or an image, with the type of change not specified.

7. Discussion. In the section, we briefly discuss the choice of k, the number of NNs to be included in the test, how the test works for gradual changes, and possible extensions of the tests to other graphs.

7.1. *Choice of k*. Heuristically, if we choose a very small value of k, some useful similarity information among the observations is not used by the test. We see from Table 2 that the power of the test increases from 1-NN to 5-NN. On the other hand, if we set k to be too large, it may include some irrelevant information, which would also harm the power.

Figure 5 plots the power of the test as k varies. The different symbols corresponds to different dimensions of the observations. For each dimension, the amount of the change is fixed and only k varies. The amount of the change for each dimension is chosen so that the highest power is around 0.8. We can see clearly from the plot the relation between the power and k: The power first increases as k increases and becomes steady for a wide range of k's and then decreases as k increases. Therefore, the optimal k should be chosen before the test reaches the plateau to achieve a high power and low computation time at the same time. Another nice thing exhibited by the plot is that the dimension of the observations does not play a significant role in the choice of k. The profiles for different dimensions, from $d = 10$ to $d = 10{,}000$, are almost the same. If we increase the strength of the signal (Figure 6), the whole curve shifts upward, while the profiles for different dimensions still remain the same.

When we set L to be larger (Figure 7, $L = 200$, versus Figure 5, $L = 50$), a similar shape is observed. It is worthwhile to note that the power of the test increases dramatically as L increase: For $d = 1000$, the power achieved by $\Delta = 4.5$ for $L = 50$ is achieved at $\Delta = 2.2$ for $L = 200$. If we set $\Delta = 4$ for $L = 200$, the power is almost 100% for 3-NN and 5-NN (shown as circles in Figure 7).

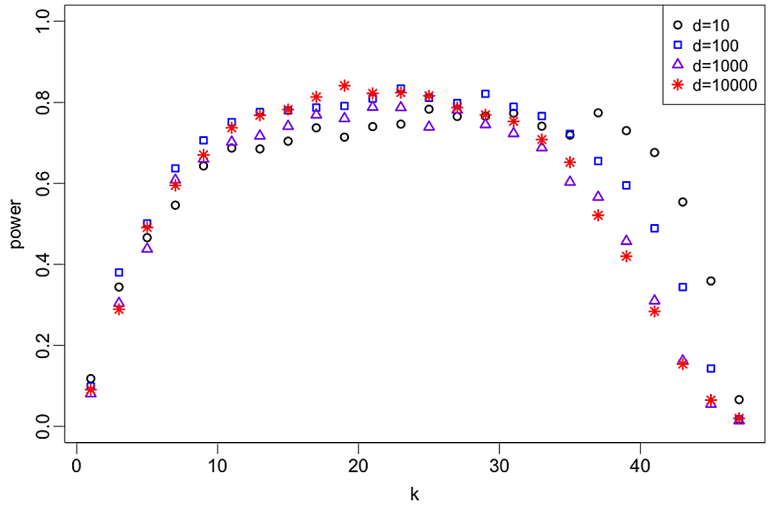

FIG. 5. *Power of the test based on k-NN for detecting change-points in sequences of multivariate normal data over a range of dimensions*: $d = 10$ (*black circle*), $d = 100$ (*blue square*), $d = 1000$ (*purple triangle*) *and* $d = 10{,}000$ (*red star*). *The change is a shift in mean with the* L_2 *distance between the means before and after the change* Δ: $\Delta = 1.7$ ($d = 10$), $\Delta = 2.7$ ($d = 100$), $\Delta = 4.5$ ($d = 1000$) *and* $\Delta = 8$ ($d = 10{,}000$). *The parameter* L *is set to be* 50, *and the power is estimated through* 1000 *simulation runs*.

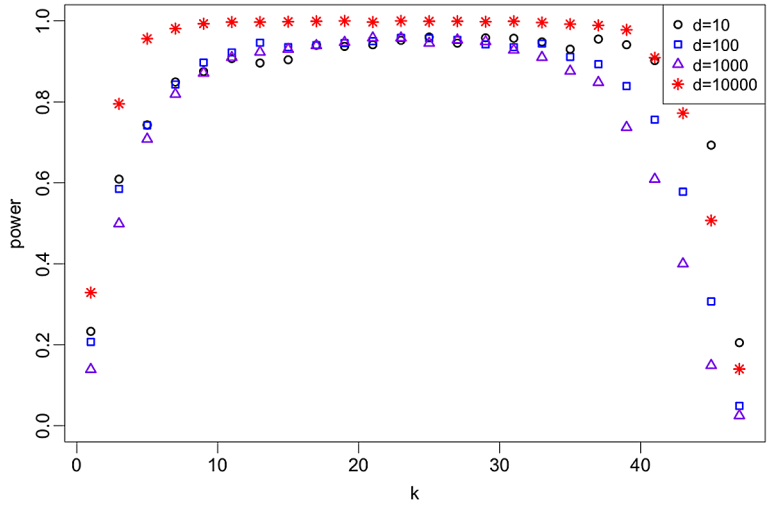

FIG. 6. *The same set up as in Figure 5 while the strength of the signal is increased for each dimension*: $\Delta = 2$ ($d = 10$), $\Delta = 3$ ($d = 100$), $\Delta = 5$ ($d = 1000$) *and* $\Delta = 10$ ($d = 10{,}000$).

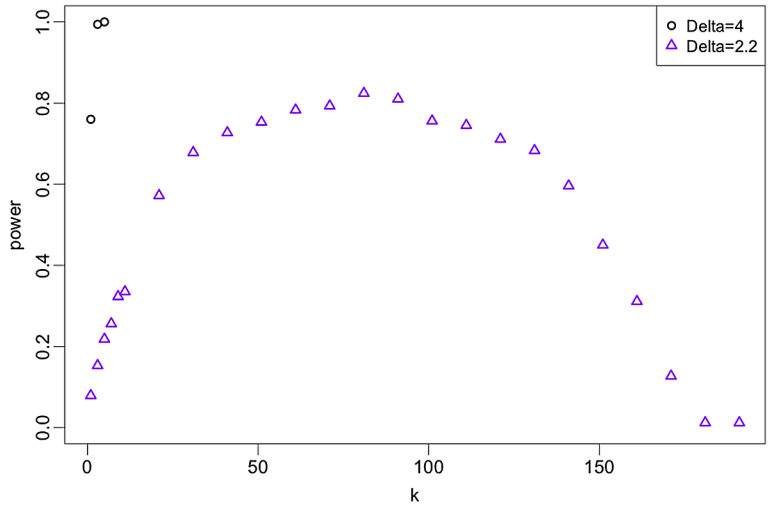

FIG. 7. *The same set up as in Figure* 5 *while* L *is set to be* 200. *The dimension of the observations in the sequence is* 1000.

In practice, for high-dimensional data or non-Euclidean data, sometimes, only large changes may be of interest, then a relative small k would be preferred as large k may detect small changes. On the other hand, if all small changes are of interest, then a relatively large k would be recommended. Also, since the statistics are easy and fast to compute, it might be helpful to run the detection for a number of k's simultaneously.

7.2. *Gradual change*. In some applications, the change may happen gradually rather than abruptly. Even though the proposed method is designed for detecting abrupt changes, it also works for gradual change as long as the change per unit time is relatively strong.

Figure 8 plots the power of the test based on 5-NN for a change of mean. In all scenarios, the L_2 distance between the mean before the change and the mean after the change is 3. However, the change could take more than one unit of time to finish. For example, if the "gradual change length" is 10, then $\|\mathsf{E}(\mathbf{Y}_{\tau+9}) - \mathsf{E}(\mathbf{Y}_{\tau-1})\|_2 = 3$ where τ is the time the change starts to happen. For simplicity, we let the change speed to be the same over the gradual change period. We see that the power decreases as the change takes longer for the same amount of change. However, the decrease in power is not too bad if the length of the change does not take too long to finalize. For example, when the "gradual change length" is 20, the power is 0.64, about 80% of the power if the same amount of change happens abruptly.

7.3. *Possible extensions to other graphs*. In this work, the focus is on the tests based on k-NNs. However, similar tests could be defined for other types of similar-

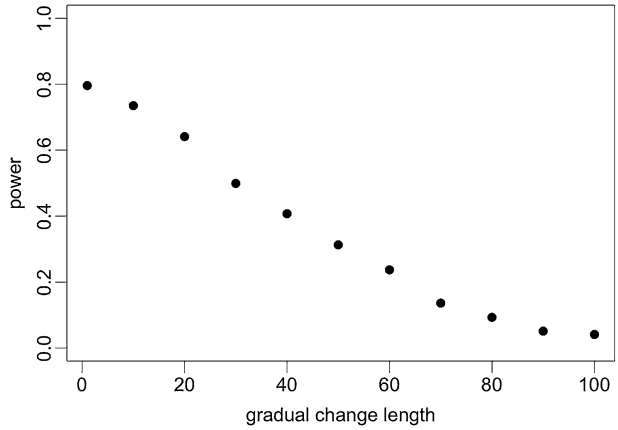

FIG. 8. *The power of the test based on* 5*-NN for the same amount of change in the mean with the change speed differs. The "gradual change length" is the amount of time the change takes to finalize. The longer the gradual change length, the slower the change happens. The dimension of the observations in the sequence is* 1000 *and* L *is set to be* 200. *The power is estimated from* 1000 *simulation runs and we call the detection successful if it detects the change within* 100 *observations from the change starts to happen.*

ity graphs. For example, we could constructed the minimum spanning tree (MST) constructed on the most recent L observations for each n, which is a graph that connects to the most recent L observations with the sum of the distances on the edges minimized, and denote the graph to be $\mathcal{M}_{n_L}$. Then $R_L(t, n)$ could be defined as the number of edges in $\mathcal{M}_{n_L}$ connecting observations before t and after t, and the standardization could be done correspondingly. Most of the theoretical treatments in this work could be adopted while we need to figure out the dynamics of the MSTs along time. In particular, for $\mathcal{M}_{m_L}$ and $\mathcal{M}_{n_L}$, we would need to figure out the expected number of edges that are shared by the two graphs, and the expected number of pairs of edges with one from $\mathcal{M}_{m_L}$ and the other from $\mathcal{M}_{n_L}$ that share a node. These expectations are not as straightforwardly obtainable as the counterparts in k-NN, but they are tractable. Also, if other ways of the constructing the similarity graph are used rather than the MST, similar arguments follows. Hence, this current work sets up the basics for graph-based methods for online change-point detection and the special treatments for different similarity graphs are more or less graph-specific. These specific treatments for other classic similarity graphs will be carried out in future works.

8. Conclusion. We propose a new framework for detecting change-points sequentially as data are generated. Motivated by the complexity of observations in many real applications, we propose to use nearest neighbor information among the observations for sequential detection. These information can usually be provided by domain experts, and thus the proposed method has a wide range of applications.

We explored several stopping rules and the one based on the most recent observations is recommended as it has the desirable property that the detection power is the same across the time. The asymptotic properties of this stopping rule is studied and the analytic approximation for calculating the average run length works well for finite samples after skewness correction. The proposed test exhibits higher power than the parametric method based on normal theory when the dimension of the data is high and/or distributional assumptions for the parametric method are violated. The proposed method is illustrated on the analysis of friendship network data over time and some interesting insights are obtained.

Acknowledgments. The author thank David Siegmund, Nancy Zhang and Jie Peng for helpful discussions.

SUPPLEMENTARY MATERIAL

Proofs for theorems (DOI: 10.1214/18-AOS1718SUPP; .pdf). This supplement contains proofs for Theorem 4.2 and Theorem 4.4.

REFERENCES

BICKEL, P. J. and BREIMAN, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probab.* **11** 185–214. MR0682809

CHAN, H. P. and WALTHER, G. (2015). Optimal detection of multi-sample aligned sparse signals. *Ann. Statist.* **43** 1865–1895. MR3375870

CHEN, H. (2019). Supplement to "Sequential change-point detection based on nearest neighbors." DOI:10.1214/18-AOS1718SUPP.

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

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

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

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783. MR0947577

KAPPENMAN, J. (2012). A perfect storm of planetary proportions. *IEEE Spectrum* **49** 26–31.

MEI, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **97** 419–433. MR2650748

QU, M., SHIH, F. Y., JING, J. and WANG, H. (2005). Automatic solar filament detection using image processing techniques. *Sol. Phys.* **228** 119–135.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

SIEGMUND, D. (1985). *Sequential Analysis*: *Tests and Confidence Intervals*. Springer, New York. MR0799155

SIEGMUND, D. and VENKATRAMAN, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.* **23** 255–271. MR1331667

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

TARTAKOVSKY, A., NIKIFOROV, I. and BASSEVILLE, M. (2015). *Sequential Analysis*: *Hypothesis Testing and Changepoint Detection*. *Monographs on Statistics and Applied Probability* **136**. CRC Press, Boca Raton, FL. MR3241619

TARTAKOVSKY, A. G. and VEERAVALLI, V. V. (2008). Asymptotically optimal quickest change detection in distributed sensor systems. *Sequential Anal*. **27** 441–475. MR2460208

WALD, A. (1973). *Sequential Analysis*. Dover, Mineola, NY.

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

XIE, Y. and SIEGMUND, D. (2013). Sequential multi-sensor change-point detection. In 2013 *Information Theory and Applications Workshop* (*ITA*) 1–20. IEEE, Los Alamitos, CA.

YANG, W., LIPSITCH, M. and SHAMAN, J. (2015). Inference of seasonal and pandemic influenza transmission dynamics. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **112** 2723–2728.

YANG, S., SANTILLANA, M. and KOU, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **112** 14473–14478.

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