Spectral CUSUM for Online Network Structure Change Detection

Minghe Zhang, Liyan Xie, *Student Member, IEEE*, and Yao Xie, *Member, IEEE*

***Abstract*— Detecting abrupt changes in the community structure of a network from noisy observations is a fundamental problem in statistics and machine learning. This paper presents an online change detection algorithm called Spectral-CUSUM to detect unknown network structure changes through a generalized likelihood ratio statistic. We characterize the average run length (ARL) and the expected detection delay (EDD) of the Spectral-CUSUM procedure and prove its asymptotic optimality. Finally, we demonstrate the good performance of the Spectral-CUSUM procedure and compare it with several baseline methods using simulations and real data examples on seismic event detection using sensor network data.**

***Index Terms*— Change-point detection, false-alarm control, graph community change detection, spectral method.**

I. Introduction

DETECTING network structure change from sequential data is a fundamental problem in high-dimensional data analysis, emerging from multiple applications, including seismic sensor networks [1], traffic networks [2], swarm behavior monitoring [3], and social network change detection [4]. A community corresponds to a subset of nodes with much higher connectivity within the group than across groups. Real-world communication structure changes can be complicated. In various settings, the change may correspond to the emergence of a community, switching community memberships, changes in the number of communities, etc.

The need for community structure change detection is high for high-dimensional sequential data, which tend to have complex inter-dependent relationships between different dimensions. Such interdependence structure can be explicit, where the network topology be inferred from data. In the network settings, *the characteristic of the changes* will be a shift in *structures* of the underlying parameters, which is fundamentally different from a simple mean-shift considered in the change-point detection literature. As has been recognized, it is possible to exploit the underlying dependence structures to design asymptotically optimal detection algorithms [5]. In addition, we typically need a detection procedure to be computationally and memory efficient, as the data streams in this setting are often very high-dimensional and generated at high speed.

We first give a few examples of detecting changes exploring community structures:

- Seismic sensor networks: The seismic research has been based on the massive amount of continuous data recorded by ultra-dense seismic sensor arrays, and many such data are publicly available on IRIS (https://www.iris.edu). In the old days, network seismology treated seismic signals individually - one sensor at a time - and detected an earthquake when multiple impulsive arrivals were consistent with a source within the Earth [6]. Recently, with advances in sensor technology, which bring densely sampled data and high-performance computing and communication, we may be able to use a *network-based detection* by exploiting correlations between sensors to extract coherence signals. This will enhance the systematic detection of weak and unusual events that currently go undetected using individual sensors. Detecting such weak events is very crucial for earthquake prediction [7], [8], oil field exploration, volcano monitoring, and deeper earth studies [9].
- Social networks: The widespread use of social networks leads to a large amount of user-generated continuous data, which is quite valuable in studying many social phenomena. One important application is to detect change points using social network data. These change points may represent the collective anticipation of or response to external events or system "shocks" [10]. Detecting such changes can provide a better understanding of patterns of social life. In other cases, early detection of change-point can predict or even prevent social stress due to disease or international threats. In social network data [11], each node represents one individual, and the edge represents the relationship between two individuals.
- Manifold is a common low-dimensional structure that lies in high-dimensional data, which can be captured using a similar network such as Isomap and Laplacian Eigenmaps [12]. Thus, change in the manifold structure can be detected from similar graphs.

In this paper, we present a new online change-point detection procedure, the Spectral CUSUM, to detect network structure changes by observing node features. We model the

Manuscript received 16 June 2022; revised 27 February 2023; accepted 3 March 2023. Date of publication 20 March 2023; date of current version 16 June 2023. This work was supported in part by NSF CAREER under Grant CCF-1650913; and in part by NSF under Grant DMS-2134037, Grant CMMI-2015787, Grant CMMI-2112533, Grant DMS-1938106, and Grant DMS-1830210. *(Corresponding author: Yao Xie.)*

Minghe Zhang and Yao Xie are with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: mzhang388@gatech.edu; yao.xie@isye.gatech.edu).

Liyan Xie is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: xieliyan@cuhk.edu.cn).

Communicated by L. Lai, Associate Editor for Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2023.3259185.

Digital Object Identifier 10.1109/TIT.2023.3259185

network structure through the inverse covariance of the noisy Gaussian features. Based on such a model, Spectral CUSUM is derived based on generalized likelihood ratios, where the unknown post-change parameters are estimated sequentially. This approach enables us to detect general types of structural changes, including the emergence of community and switching membership. The main theoretical contribution is to show the first-order asymptotic optimality of Spectral CUSUM and characterize the optimal choice of the parameters. We also present an online scheme for computing the detection statistic based on subspace tracking that is computationally and memory efficient. We demonstrate the meritorious performance of Spectral CUSUM through simulated and real-data examples of detecting changes in Yellowstone seismic sensors data.

The rest of the paper is organized as follows. Section II provides a detailed formulation of the emerging communities problem as well as the switching membership problem. Section III introduces the exact-CUSUM and the proposed Spectral-CUSUM procedure. Section IV presents the asymptotic analysis of the proposed detection scheme together with parameter optimization and proof of the first-order asymptotic optimality. Section V presents an efficient gradient-based algorithm to keep track of the underlying community structure. Section VI gives simulation and real data examples to verify the theoretical findings and show the good performance of the proposed method. We delegate all proofs to the appendix.

A. Literature

Community detection in the *offline* setting (i.e., when the samples are collected beforehand and inference is made in one-shot) is a well-studied problem (see [13] for a survey). For example, spectral methods based on eigenvectors of the graph Laplacian are used in [14] and [15]. Besides, there are many practical algorithms for community detection (see, e.g., [16]), including the so-called Kernighan andLin algorithm [17], which uses a greedy algorithm to improve an initial division of the network and a genetic algorithm named Ganet [18] to detect the community structure by calculating the community score. However, offline community detection cannot be used to detect community changes if the graph is dynamic.

Online community detection has also been considered in the literature. Peel and Clauset [10] first formalized the change-point detection problem as identifying the times at which the large-scale patterns of interaction change fundamentally. They choose from a parametric family of probability distribution to describe the data and then use the Bayesian method to detect the change. A Markov-process-based approach is presented by [19], which is based on MCMC. Each graph snapshot depends on the current generative model and the previously observed snapshot. Moreover, [20] proposes a method called Spotlight to detect anomalies in streaming graphs by composing a K-dimensional sketch containing K subgraphs to detect changes inside the dynamic graph. Recently, a Laplacian anomaly detection method for dynamic graphs is presented in [21], which uses the spectrum of the Laplacian matrix of the graph structure at each snapshot to obtain low dimensional embeddings. However, none of these works gives a statistical perspective and asymptotic analysis for the quickest change detection in communities. While estimating the community structure through a dynamic network, our paper presents novel methods of handling estimation and change detection, which are supported by theoretical asymptotic optimality.

Spectral graph change detection is also related to our work. The spectral property of the graph is one of the essential theories that can capture the community structure of the graph. The spectral method proposed by [22] is used to detect changes in Noisy Dynamic Networks efficiently by transforming a graph to a lower-dimensional latent Euclidean space. In our work, we also use the spectral method to get a reduced dimensional representation for each node, and our procedure can detect different types of changes and has theoretically proved optimally.

Change Detection of Gaussian Graphical Model is similar to our work as well. We consider the Gaussian graphical model to capture the network structure through the inverse covariance matrix of the nodal features. The Gaussian model is useful for modeling the correlation between observations at different nodes. There are many previous works that are focused on the estimation of the Gaussian graphical model. For instance, penalized likelihood method is proposed by [23] for estimating the concentration matrix in the Gaussian graphical model while [24] presents an alternative tuning-insensitive approach to efficiently choose the tuning parameter in finite sample settings. But none of these estimation methods of the Gaussian graphical model can be directly used to detect the structure change. Recently, a piecewise stationary graphical model has been presented in [25], and it is able to detect the change of the graph by monitoring the conditional log-likelihood of all nodes in the network. However, this method does not consider the community structure of the network and thus is unable to distinguish different types of changes inside the network.

Manifold change detection is another type of related work. In the work of [26], multi-scale online manifold learning is used to extract change-point detection test statistics from high-dimensional data. But they do not consider the network property for the high-dimensional data and thus cannot be applied directly to graph scenarios.

II. Problem Setup

Consider a dynamic network with n nodes. Assume at each time t, we observe a feature value v_{ti} for each node i, $i = 1, \ldots, n$. We collect all features at each time into one vector and denote as $v_t \in \mathbb{R}^n, t = 1, 2, \ldots$. We will focus on the online setting where we observe feature values sequentially. Such a setting is widely applicable in real datasets, see below for some examples.

- In sensor networks, the feature $\{v_{ti}, t = 1, 2, \ldots\}$ represent a sequence of signals recorded by the i-th sensor. The sensors may form communities, and thus the features can be correlated. It can be the observed seismic/solar activity measurement from each sensor at each time in a seismic/solar system.
- In social networks, a vector of features represents user activities at each time. For instance, it can be social

activities such as twittering at time t from each user in the Twitter network.

Assume the features for nodes within the same community have a higher correlation than those that are not in the same community. The correlation can be estimated using observed nodal features [27]. The underlying community structure may change at some time, which leads to a change in the correlations between the affected nodes and, thus the distribution of feature vectors. We aim to detect such a change as quickly as possible from sequential observations.

In the following, we first give the community definition in Section II-A and statistical modeling for the community structure within graphs, then discuss two kinds of changes respectively in Section II-B.

A. Adjacency Matrix for Community

Suppose there are m communities within a network with n nodes. Denote these communities as m sets $\{C_1, \cdots, C_m\}$, where the k-th community is represented using the index set C_k of the nodes belonging to this community. Assume the communities do not overlap with each other, i.e., the index sets are mutually disjoint. For each node i, we introduce an indicator vector $a_i \in \{0,1\}^m$ representing its true membership: the k-th entry equal to 1 and all other entries equal to 0 if node i belongs to the k-th community C_k, i.e.,

$$a_i = [0 \ \cdots \ \underbrace{1}_{k\text{-th entry}} \ \cdots \ 0]^\top, \forall i \in C_k.$$

Define the global indicator matrix as

$$A = \left[a_1, \cdots, a_n\right]^\top \in \{0,1\}^{n\times m}. \tag{1}$$

Notice that we have

$$a_i^\top a_j = \begin{cases} 1 & \text{if } \exists k \text{ s.t. nodes } i,j \in C_k, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the matrix $AA^\top \in \{0,1\}^{n\times n}$ defines an *adjacency matrix*, whose (i,j)-th entry equals to 1 if and only if nodes i and j belong to the same community. This setup for A can also be generalized beyond 0-1 matrices. For example, we may let $a_i \in \mathbb{R}^m$ represent a feature embedding vector for node i, and each entry of a_i represents the weight/probability for node i belonging to the corresponding community.

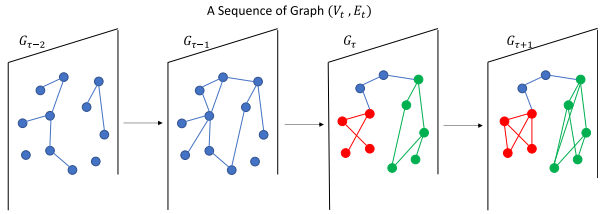
B. Community Change-Point Detection

We aim to monitor two types of structural changes in dynamic networks:

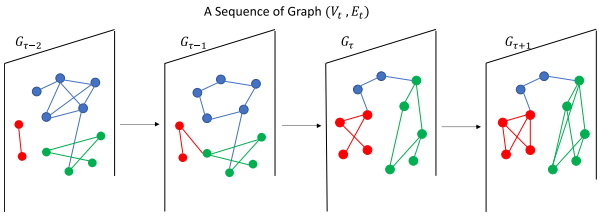
i) The emergence of new communities: Before the change happens, there is no clear community formed in the graph, and the community structure emerges after the change, as indicated in Figure 1(a).

ii) Switching membership: Some community members are switched after the change, such as the increase or decrease of a single community or the membership flow from one community to another, as indicated in Figure 1(b).

We now formulate such change detection problems based on the adjacency matrix representation in Section II-A.



(a) Emergence problem illustration



(b) Switching membership problem illustration

Fig. 1. Case of the emergence of communities after change time τ as shown in (a) and switching membership case shown in (b). Three communities are marked as red, blue, and green dots. For the emergence problem, it can be seen that there is no community or one single community before changes happen, then after that, three separate communities emerge. For the switching membership case, it can be seen that before the change happens, the blue community is the biggest one, while after the change, the membership of it begins switching to the other two communities.

1) Emerging Community: In many applications, the change can be modeled as the emergence of several disjoint communities; the nodes inside the same community are more correlated with each other. Thus, we start by considering the emerging community detection problem, which assumes that the network has no community structure at the beginning but forms m communities $(C_1, \cdots, C_m)$ after the change where C_k is the node index set for the k-th community.

With the indicator matrix A defined in (1), we assume the feature observations v_t are multivariate Gaussian and its covariance matrix is modeled based on the adjacency matrix $AA^\top$. More specifically, we cast the emerging community problem as follows.

$$\begin{aligned} H_0: \quad & v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{\sigma^2} I), \quad t = 1, 2, \ldots \\ H_1: \quad & v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{\sigma^2} I), \quad t = 1, 2, \ldots, \tau, \\ & v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (AA^\top + \sigma^2 I)^{-1}), \quad t = \tau{+}1, \tau{+}2, \ldots \end{aligned} \tag{2}$$

Here we introduce $\sigma^2 I$ as a noise term since the network is usually not perfectly separable in practice. Note that in the formulation here and below for the switching subspace, we consider the structure of the inverse covariance matrix, which is a common approach for Gaussian graphical models [28]. Following this, we can regard $AA^\top$ as the underlying community structure of the graph and σ^2 as the noise level.

The above representation can be related to the Gaussian graphical model, which is a common approach to exploring the relationships between nodes in an undirected graph through the inverse covariance matrix. Given a Gaussian graphical model with covariance matrix Σ, there is an edge between node i and node j, if and only if $\Sigma^{-1}_{ij} \neq 0$. Note that this can be related

to our model: the zero-valued entry in the inverse covariance matrix means that the corresponding edge does not exist in the graphical model. As the inverse covariance matrix changes from $\sigma^2 I$ to $AA^\top + \sigma^2 I$ in a model (2), the corresponding off-diagonal entries change from zero to non-zero after the community emerges.

2) Switching Membership: Another type of community change is called the switching membership problem. As shown in Figure 1(b), some of the nodes belong to one community at first and switch to a different one after the change happens at time τ. Similar to (2), the switching membership problem can be formulated as follows:

$$\begin{aligned} H_0: &\quad v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (A_1 A_1^\top + \sigma^2 I)^{-1}), \quad t = 1, 2, \dots \\ H_1: &\quad v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (A_1 A_1^\top + \sigma^2 I)^{-1}), \quad t = 1, 2, \dots, \tau, \\ &\quad v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (A_2 A_2^\top + \sigma^2 I)^{-1}), \quad t = \tau+1, \tau+2, \dots \end{aligned} \tag{3}$$

Here A_1 represents the pre-change community structure while A_2 represents the post-change community structure. This general model can denote cases when the sizes of certain communities change. It can also model the case when the total number of communities increases (one community splits into smaller ones) or decreases (several small communities merge into a bigger one). As a result, the emergence problem can be seen as a special case of the switching membership problem, which is capable of detecting various types of graph structure changes. We will show later the detecting procedure of switching membership can also be equivalently treated as a community emergence problem.

III. Detection Procedures

In this section, we first review the well-known cumulative sum (CUSUM) detection rule and then propose the Spectral-CUSUM procedure under both emergence and switching membership scenarios.

A. Exact-CUSUM Procedure

Let $f_\infty(\cdot)$ and $f_0(\cdot)$ denote the pre- and post-change probability density function (pdf) of the observations, and $\mathbb{E}_\infty$ and $\mathbb{E}_0$ denotes the expectation under f_∞ and f_0, respectively. The CUSUM statistic [29] is defined by maximizing the log-likelihood ratio statistic over all possible change-point locations:

$$S_t = \max_{1 \le k \le t} \sum_{i=k}^{t} \log \frac{f_0(v_i)}{f_\infty(v_i)}.$$

S_t has a recursive formulation with $S_0 = 0$ as follows:

$$S_t = (S_{t-1})^+ + \log \frac{f_0(v_t)}{f_\infty(v_t)}, \ t \ge 1,$$

where $(x)^+ := \max\{x, 0\}$. The corresponding CUSUM stopping time T_C is defined as:

$$T_C = \inf\{t > 0 : S_t \ge b\}, \tag{4}$$

where b is a pre-set constant threshold. Under the model (2), we have that

$$\log \frac{f_0(v_t)}{f_\infty(v_t)} = -\frac{1}{2} v_t^\top A A^\top v_t + \frac{1}{2} \log \frac{\det(AA^\top + \sigma^2 I)}{\sigma^{2n}}.$$

Since the multiplicative factor $1/2$ is positive, we can omit it from the log-likelihood ratio when forming the CUSUM statistic, thus yielding an equivalent formulation:

$$S_t = (S_{t-1})^+ - v_t^\top A A^\top v_t + \underbrace{\log \frac{\det(AA^\top + \sigma^2 I)}{\sigma^{2n}}}_{d}, \tag{5}$$

where d is a drift parameter that is fixed in the CUSUM procedure.

Similarly, for the switching membership problem under the model (3), the log-likelihood ratio is:

$$\begin{aligned} \log \frac{f_0(v_t)}{f_\infty(v_t)} = &-\frac{1}{2} v_t^\top (A_2 A_2^\top - A_1 A_1^\top) v_t \\ &+ \frac{1}{2} \log \frac{\det(A_2 A_2^\top + \sigma^2 I)}{\det(A_1 A_1^\top + \sigma^2 I)}. \end{aligned}$$

Therefore, the CUSUM statistic in this case becomes:

$$S_t = (S_{t-1})^+ - v_t^\top (A_2 A_2^\top - A_1 A_1^\top) v_t + \underbrace{\log \frac{\det(A_2 A_2^\top + \sigma^2 I)}{\det(A_1 A_1^\top + \sigma^2 I)}}_{d'}. \tag{6}$$

We note that in the exact CUSUM procedure, all the parameters are assumed known so that the drift terms d and d' can be computed explicitly beforehand.

B. Spectral-CUSUM Procedure

The implementation of the exact CUSUM procedure requires that all parameters are known, and it has been proved to be optimum [30], [31]. However, if the post-change distribution is unknown, the exact-CUSUM is not applicable. Usually, we can estimate the pre-change distribution using historical data (training data), but the post-change community structures are unknown since it represents anomaly information and cannot be predicted. Therefore, the post-change distribution, i.e., the post-change structure A, has to be estimated sequentially from data. A natural estimate for the post-change covariance matrix $(AA^\top + \sigma^2 I)^{-1}$ is the sample covariance matrix. To eliminate the matrix inversion computation in estimating A, we use eigendecomposition on the sample covariance matrix to estimate A directly. Therefore, we propose the Spectral-CUSUM procedure below, where we approximate the covariance matrix by rank-m eigendecomposition.

Define $U \in \mathbb{R}^{n \times m}$ and $\Lambda \in \mathbb{R}^{m \times m}$ as the eigenvectors and eigenvalues of the term $AA^\top$ within the post-change covariance matrix, i.e., $AA^\top = U \Lambda U^\top$. Here Λ is a diagonal matrix corresponding to the eigenvalues of the matrix $AA^\top$ defined using (1):

$$\Lambda = \begin{bmatrix} |C_1| & & \\ & \ddots & \\ & & |C_m| \end{bmatrix},$$

where $|C_k|$ denotes the number of nodes inside the k-th community. This shows that the eigenvalue decomposition of the adjacency matrix would reflect the community structure of a graph, which can also provide empirical guidance on how to determine the potential number m of communities.

We can then jointly estimate matrices U and Λ. Using observations $\{v_{t+1}, \ldots, v_{t+w}\}$ in a *future* sliding window with length w, we define

$$\widehat{G}_t = (v_{t+1}v_{t+1}^\top + \cdots + v_{t+w}v_{t+w}^\top)/w. \tag{7}$$

Note that $\widehat{G}_t$ serves as an approximation for the covariance matrix $(AA^\top + \sigma^2 I)^{-1}$ if samples $\{v_{t+1}, \ldots, v_{t+w}\}$ are drawn from post-change distribution; for ease of presentation, we assume eigenvalues are distinct. Let $\{\hat{u}_{t1}, \ldots, \hat{u}_{tm}\}$ be the m unit-norm eigenvectors corresponding to the m *smallest* eigenvalues $\{\hat{\lambda}_{t1} < \cdots < \hat{\lambda}_{tm}\}$ of $\widehat{G}_t$. Let

$$\widehat{U}_t = [\hat{u}_{t1}, \ldots, \hat{u}_{tm}] \in \mathbb{R}^{n\times m},$$

and

$$\widehat{\Lambda}_t = \text{diag}(\hat{\lambda}_{t1}, \ldots, \hat{\lambda}_{tm}) \in \mathbb{R}^{m\times m}.$$

Then A can be approximately estimated using $\widehat{A}_t = \widehat{U}_t\widehat{\Lambda}_t^{-1/2}$, as the noise term σ^2 is usually relatively small. Then we can substitute the estimate $\widehat{A}_t$ into (5) to obtain alternative detection statistics, which we call Spectral-CUSUM, when the post-change distribution is unknown:

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ - v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t + d. \tag{8}$$

Here d is a tunable drift parameter that plays a similar role as the last term in (5). And the stopping time is defined as follows

$$\mathcal{T}_C = \inf\{t > 0 : \mathcal{S}_t \geq b\}. \tag{9}$$

Similarly, for the switching membership problem, the alternative to (6) is:

$$\mathcal{S}_t = (\mathcal{S}_{t-1})^+ - v_t^\top(\widehat{A}_t\widehat{A}_t^\top - A_1A_1^\top)v_t + d'. \tag{10}$$

In the proposed Spectral-CUSUM procedure, the drift parameter d should be chosen to ensure the detection statistics are capable of detecting the change. More specifically, the CUSUM type procedure requires the increment term in $\mathcal{S}_t$ to have a *negative* mean under the pre-change distribution, and a *positive* mean under the post-change distribution. Therefore, for the emerging community problem, we need:

$$\mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t] < d < \mathbb{E}_\infty[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]. \tag{11}$$

For the switching membership problem, the same rule means that we need:

$$\mathbb{E}_0[v_t^\top(\widehat{A}_t\widehat{A}_t^\top - A_1A_1^\top)v_t] < d' < \mathbb{E}_\infty[v_t^\top(\widehat{A}_t\widehat{A}_t^\top - A_1A_1^\top)v_t]. \tag{12}$$

Since A_1 is a constant matrix known beforehand from historical data, we can calculate the expectations $\mathbb{E}[v_t^\top A_1A_1^\top v_t]$ explicitly. Thus, the switching membership problem can be treated as an emerging community problem. Therefore, the theoretical analysis for the switching membership problem is similar to the emerging community problem, and we will only discuss the emergence problem in Section IV.

Due to the above property (11), which is also mentioned in [32] and [33], the detection statistic in (8) will deviate from 0 and increase gradually after the change happens.

Algorithm 1 Spectral-CUSUM Procedure

Input: Sequence of observations $\{v_t, t = 1, 2, \ldots\}$, number of communities m, sliding window size w, carefully selected drift parameter d (or d'), detection threshold b.

Output: Stopping time $\mathcal{T}_C$.

1: Initialize $\mathcal{T}_C = \infty$, $t = 0$, $S_0 = 0$.
2: **while** $S_t < b$ **do**
3: Calculate sample covariance matrix $\widehat{G}_t$ from future observations using (7);
4: Compute $\widehat{U}_t$ and $\widehat{\Lambda}_t$ through the eigenvalue decomposition of $\widehat{G}_t$;
5: Estimate community structure $\widehat{A}_t = \widehat{U}_t\widehat{\Lambda}_t^{-1/2}$;
6: Let $t = t + 1$ and update Spectral-CUSUM statistics $\mathcal{S}_t$ following (8) or (10).
7: Set $\mathcal{T}_C = t$.
8: **return** $\mathcal{T}_C$

IV. Theoretical Analysis

This section provides a theoretical analysis of the proposed Spectral-CUSUM procedure under the emerging community setting. The main result is presented in Theorem 2, which shows the asymptotic optimality of Spectral-CUSUM under the optimal choices of parameters. We also derive the form of the optimal parameters. The analysis techniques are related and extended to those used in prior works in [34], [35] and [36] for subspace change detection.

A. Preliminary

We first state the following assumptions that are being made in order to derive the main results.

Assumption IV.1: We make the following assumptions for the true community structure.

(1) The total number of post-change communities remains a constant m.
(2) (Community sizes are comparable but not not identical) The size of different communities are not identical and is ordered as $|C_1| > |C_2| > \cdots > |C_m|$, and $|C_1|/|C_m| \leq (1+\theta)^2$ for some $\theta > 0$.
(3) (Noise variance bounded)

$$\frac{\sigma^2}{|C_i|} < \left(\frac{1}{4} + \eta \cdot \sqrt{\frac{m-1}{n-m}} \cdot \frac{1}{\left(1+\theta-\frac{1}{1+\theta}\right)^2}\right)^{1/2} - \frac{1}{2}$$

, $\forall i$, for a small constant $0 < \eta < 1$.

We remark that Assumption IV.1(1) can be treated as a pre-defined number of communities to detect. Assumption IV.1(2) assumes that the community sizes are comparable. Assumption IV.1(3) requires that the noise variance relative to the community sizes is not too large, but we do not

need the noise variance to be 0. In fact, when $\theta \to 0$, i.e., when the community size is very similar to each other, the ratio $\frac{1}{(1+\theta-\frac{1}{1+\theta})^2} \to \infty$ and thus the condition can be easily satisfied; we can even allow high noise σ^2 when the community sizes are comparable.

Define the following quantities to simplify the presentation of the main results:

$$B_i = \sum_{k=1,k\neq i}^{m} \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2} > 0,$$
$$D = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda_i+\sigma^2}\left(1-\frac{B_i}{w}+\frac{3B_i^2}{w^2}\right),$$
$$\widetilde{D} = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda_i+\sigma^2}\left(1-\frac{B_i}{w}\right); \tag{13}$$

recall that $\lambda_i = |C_i|$ is the i-th largest eigenvalue of the matrix $AA^\top$ (thus we always have $\lambda_i > 0$), w is the sliding window size, and σ^2 is the noise level. For theoretical analysis purposes, we assume the eigenvalues are distinct, i.e., no community sizes are exactly the same (see Assumption IV.1(2)); not that when this does not hold, we can still apply the algorithm although the theory needs to be extended. Because of Assumption IV.1(2), we have

$$\frac{m-1}{\left(1+\theta-\frac{1}{1+\theta}\right)^2} < B_i < \frac{m-1}{\left(\sqrt{\frac{|C_1|+1}{|C_1|}}-\sqrt{\frac{|C_1|}{|C_1|+1}}\right)^2}, \quad \forall i = 1,\dots,m. \tag{14}$$

Moreover, a basic bound can be derived: $D \geq \frac{11}{12}\sum_{i=1}^{m}\frac{\lambda_i}{\lambda_i+\sigma^2} > 0$ based the property of the quadratic equation, and $\frac{11}{12}\sum_{i=1}^{m}\frac{\lambda_i}{\lambda_i+\sigma^2} \leq \widetilde{D} \leq D \leq \sum_{i=1}^{m}\frac{\lambda_i}{\lambda_i+\sigma^2} < m$ when w is sufficiently large.

We start with a useful result from [37].

Theorem 1 (Asymptotic Property of Sample Covariance, [37]): The asymptotic distribution of eigenvectors and eigenvalues of sample covariance matrix (7) from w samples under post-change distribution (2) are: (i) independent and (ii) has the following distribution under the assumption that the eigenvalues are distinct:

$$\sqrt{w}(\hat{u}_i - u_i) \xrightarrow{d} \mathcal{N}\left(0, \sum_{k=1,k\neq i}^{m} \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2}u_k u_k^\top\right),$$
$$\sqrt{w}(\hat{\lambda}_i - \lambda_i) \xrightarrow{d} \mathcal{N}(0, 2\lambda_i^2),$$

where $\xrightarrow{d}$ denotes convergence in distribution when $w \to \infty$. Here $\hat{u}_i$ and $\hat{\lambda}_i$ denotes the sample eigenvectors and eigenvalues, while u_i and λ_i denotes the eigenvectors and eigenvalues of the true covariance matrix, and n is the data dimension.

For the emerging community problem in (2), the post-change covariance matrix is $(AA^\top+\sigma^2 I)^{-1}$. Recall that we use the window size w to construct the post-change sample covariance matrix in (7) and then estimate the eigenvalues and eigenvectors. Denote $\rho_1 < \cdots < \rho_m$ as the smallest m eigenvalues of $(AA^\top+\sigma^2 I)^{-1}$; based on the eigendecomposition $AA^\top = U\Lambda U^\top$, we can show $\rho_i = 1/(\sigma^2+|C_i|)$. Let u_i, $i = 1,\dots,m$ denote the corresponding eigenvectors. Let $\hat{\rho}_i$ and $\hat{u}_i$, $i = 1,\dots,m$, be the smallest m eigenvalues and corresponding eigenvectors obtained from the sample covariance matrix $\widehat{G}_t$. By Theorem 1, we have

$$\sqrt{w}(\hat{\rho}_i - \rho_i) \xrightarrow{d} \mathcal{N}(0, 2\rho_i^2),$$

and

$$\sqrt{w}(\hat{u}_i - u_i) \xrightarrow{d} \mathcal{N}\Bigg(0, \sum_{k=1,k\neq i}^{m} \frac{\rho_i\rho_k}{(\rho_i-\rho_k)^2}u_k u_k^\top + \frac{\rho_i/\sigma^2}{(\rho_i-1/\sigma^2)^2}(I-UU^\top)\Bigg).$$

We can show in the above estimation error covariance matrix (IV-A), the second term in the summation is negligible compared with the first term under AssumptionIV.1, when η is sufficiently small (e.g., 0.1, 0.2) — η is desired upper bound for the ratio of the second term relative to the first term. The derivation can be found in the appendix.

With the help of Theorem 1 we can show the following properties for Spectral-CUSUM.

Lemma 1 (Properties of Detection Statistic for Spectral-CUSUM): When $w \to \infty$, the expected drifts under the pre- and post-change distributions for Spectral-CUSUM are given by

$$\mathbb{E}_\infty[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t] = m, \quad \widetilde{D} \leq \mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t] \leq D.$$

The proof of this Lemma is shown in the Appendix. By Lemma 1, the necessary condition for the drift parameter d in (11) translates into $D < d < m$. Later in Section IV-C, we can show that by construction $\mathbb{E}_\infty[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t] > \mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]$ (i.e., $m > D$) can be satisfied, and hence we can choose a suitable drift parameter d in between such that the Spectral-CUSUM procedure work. Indeed, we emphasize that we focus on the regime that w is sufficiently large such that $D < m$ is guaranteed.

B. ARL/EDD Performance Analysis

The standard performance of change detection procedures is measured by average run length (ARL) and expected detection delay (EDD). ARL represents the average time interval between two consecutive false alarms, while EDD measures the worst-case detection delay. When ARL is fixed, it is known that the exact CUSUM procedure minimizes EDD, which can be calculated directly. In the following, we will analyze ARL and EDD for the proposed Spectral-CUSUM procedure given in (8) for the emerging community problem, under the assumption that the window size w^* scales in order $\sqrt{\log\gamma}$ in the asymptotic analysis as $\gamma \to \infty$. In this case, $D/\widetilde{D}$ approaches to 1, and then we can approximate $\mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]$ by D in our derivation.

Given a constant $\gamma > 1$ as the desired lower bound of ARL, we need to set the threshold b in (4) (and similarly in (9) for Spectral-CUSUM) accordingly such that ARL $\geq \gamma$. Recall that T_C denotes the stopping time for the exact CUSUM procedure. Thus $\mathbb{E}_0[T_C]$ and $\mathbb{E}_\infty[T_C]$ are EDD and ARL of CUSUM. Similarly, $\mathbb{E}_0[\mathcal{T}_C]$ and $\mathbb{E}_\infty[\mathcal{T}_C]$ are the EDD and

ARL of Spectral-CUSUM. According to classic results from [38], we have the following:

$$\mathbb{E}_\infty[T_C] = \frac{e^b}{\mathcal{I}_\infty}(1+o(1)),\ \mathbb{E}_0[T_C] = \frac{b}{\mathcal{I}_0}(1+o(1)), \quad (15)$$

where $\mathcal{I}_0$ and $\mathcal{I}_\infty$ are the Kullback-Leibler (KL) divergence:

$$\mathcal{I}_\infty = \mathbb{E}_\infty\{\log[f_\infty(x)/f_0(x)]\}, \quad \mathcal{I}_0 = \mathbb{E}_0\{\log[f_0(x)/f_\infty(x)]\},$$

The constraint ARL $\geq \gamma$ will be satisfied with threshold $b = (\log\gamma)(1+o(1))$ according to (15).

Lemma 2 (KL Divergence for Emerging Subspace Case): For emerging subspace problem (2), we have the Kullback–Leibler divergence (K-L) divergence

$$\mathcal{I}_0 = \mathbb{E}_0\left[\log\frac{f_0(v)}{f_\infty(v)}\right] = -\frac{1}{2}\sum_{i=1}^{m} h\left(\frac{\lambda_i}{\sigma^2+\lambda_i}\right),$$

where $h(x) = x + \log(1-x)$.

Note that $h(x) < 0$, for $x \in (0,1)$, so this also verifies that $\mathcal{I}_0 > 0$.

Using the K-L divergence in Lemma 2, we can obtain the EDD expression:

$$\mathbb{E}_0[T_C] = \frac{\log\gamma}{\mathcal{I}_0}(1+o(1)). \quad (16)$$

For performance analysis of Spectral-CUSUM, we follow a similar strategy as [35], where the analysis is done for a different subspace detection problem under the special rank-one case; here, to generalize the analysis, we extend it for rank more than one. Since the increment term $-v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t + d$ in (8) is not a log-likelihood ratio. Thus we cannot use the ARL and EDD expressions in (15) directly, which are derived for log likelihood-ratio based CUSUM procedures. To compute the ARL and EDD of $\mathcal{T}_C$, we introduce an *equalizer* $\delta_\infty \in \mathbb{R}$ such that:

$$\mathbb{E}_\infty\left[\exp\{\delta_\infty(-v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t + d)\}\right] = 1. \quad (17)$$

And then when (17) holds, $\delta_\infty[-v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t + d]$ is the log-likelihood ratio between $\tilde{f}_0$ and f_∞ where

$$\tilde{f}_0 = \exp\{\delta_\infty[-v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t + d]\}f_\infty.$$

This allows us to compute the threshold b asymptotically as $b = (\log\gamma)(1+o(1))/\delta_\infty$. Similarly, we can find a $\delta_0 > 0$ and define $\tilde{f}_\infty = \exp\{\delta_0[v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t - d]\}f_0$ so that $\delta_0[v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t - d]$ is the log-likelihood ratio between f_0 and $\tilde{f}_\infty$, leading to $\mathbb{E}_0[\mathcal{T}_C] = b(1+o(1))/(\mathbb{E}_0[-v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t] + d)$ where the dependence on δ_0 being an $o(1)$ term. Now after substituting b, using the ARL and EDD expression for the likelihood-ratio based CUSUM in (15), we obtain:

$$\begin{aligned}\mathbb{E}_0(\mathcal{T}_C) &= \frac{\log\gamma\big(1+o(1)\big)}{\delta_\infty(\mathbb{E}_0[-v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t] + d)} + w \\ &= \frac{\log\gamma\big(1+o(1)\big)}{\delta_\infty(-D+d)} + w, \end{aligned} \quad (18)$$

where the second equality is due to Lemma 1; and the window length w is added for the reason that we are using additional data to perform estimation to detect the potential change at time t; thus, the actual detection time is $t+w$.

Now would like to find the equalizer δ_∞ that satisfies (17) and thus can lead to the desired likelihood ratio construction above. Using standard computations involving Moment Generating Function for Gaussian random variables, we can write:

$$\begin{aligned}&\mathbb{E}_\infty[e^{\delta_\infty[-v_t^\top \hat{A}_t\hat{A}_t^\top v_t + d]}] \\ &= e^{\delta_\infty d}\mathbb{E}\left[\mathbb{E}_\infty[e^{-\delta_\infty[v_t^\top \hat{A}_t\hat{A}_t^\top v_t]}|\hat{A}_t]\right] \\ &= e^{\delta_\infty d}\mathbb{E}\left[\int e^{-\delta_\infty(v_t^\top \hat{A}_t\hat{A}_t^\top v_t)}\frac{e^{-v_t^\top v_t\sigma^2/2}}{\sqrt{(2\pi)^n(1/\sigma^2)^n}}dv_t\right] \\ &= e^{\delta_\infty d}\mathbb{E}_\infty\left[|I + 2\sigma^{-2}\delta_\infty\hat{A}_t\hat{A}_t^\top|^{-1/2}\right] \\ &= \frac{e^{\delta_\infty d}}{\sqrt{\prod_{i=1}^m(1+2\sigma^{-2}\delta_\infty/\rho_i)}} = 1. \end{aligned} \quad (19)$$

In evaluating the integral above, we use the standard technique of "completing the square" in the exponent for Gaussian distribution. With proper normalization, we generate an alternative Gaussian probability density function which integrates to 1. Solving (19) we obtain that the drift term d is related to the equalizer δ_∞: $d = (2\delta_\infty)^{-1}\sum_{i=1}^m \log(1+2\sigma^{-2}\delta_\infty/\rho_i)$. We notice that under the pre-change measure, all eigenvalues ρ_i of the covariance matrix are equal to $1/\sigma^2$, thus the drift parameter can be further written as:

$$d = \frac{m\log(1+2\delta_\infty)}{2\delta_\infty}. \quad (20)$$

Using (20), we can eliminate d from the EDD expression and only leave δ_∞ dependence. We will later validate that the drift d in (20) under the optimal choice δ_∞^* will indeed satisfy the condition (11) to ensure a valid CUSUM procedure. Combining Equation (18) and Equation (20) we have the expression for EDD is:

$$\mathbb{E}_0(\mathcal{T}_C) = \frac{2\log\gamma\big(1+o(1)\big)}{-2\delta_\infty D + m\log(1+2\delta_\infty)} + w. \quad (21)$$

In the following, we will further derive the optimal value of δ_∞ (thus the optimal value of the drift parameter since they are equivalent through (20)) to minimize the EDD, as a function of w.

C. Optimal Parameters for Spectral-CUSUM to Minimize EDD

Note that the formulation (21) contains two parameters: the drift d (or the equalizer δ_∞) and the window size w. We will first optimize over these parameters to minimize EDD, then show that the Spectral-CUSUM procedure based on optimal window size and draft is first-order asymptotically optimum.

We first find the optimal value of δ_∞ and the corresponding drift parameter d (according to (20)). We observe that the denominator in (21) is a concave function of δ_∞ therefore, it exhibits a single maximum. Setting the derivative of the denominator as a function of δ_∞ to be 0, we obtain the optimum value of δ_∞ (we omit the high order terms of

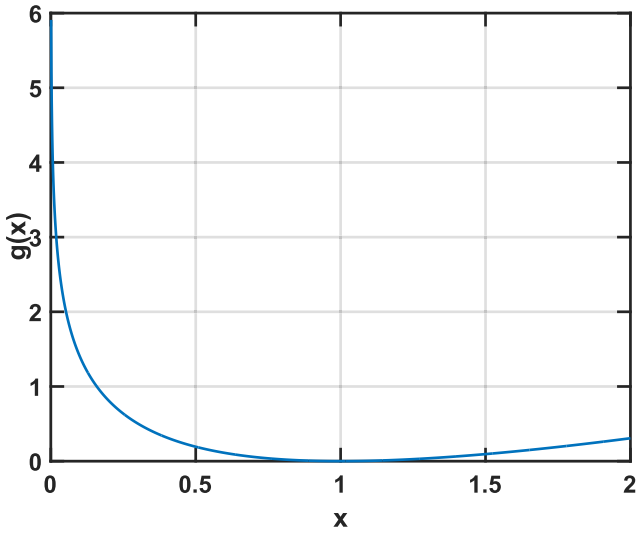


Fig. 2. Plot of $g(x) = x - 1 - \log(x)$ in the expression of EDD (24) with δ_∞^*; note that $g(x) \geq 0$ which ensures (24) is meaningful; $g(x) \to \infty$ when $x \to 0$.

eigenvalue in the product):

$$\delta_\infty^* = \frac{m}{2D} - \frac{1}{2}. \tag{22}$$

Note that the optimal drift $d^* = \frac{m\log(1+2\delta_\infty{}^*)}{2\delta_\infty^*}$ corresponding to the optimal δ_∞^* satisfies the condition (11). Substitute the optimal δ_∞^*, we can show that

$$D < \frac{m\log(1+2\delta_\infty^*)}{2\delta_\infty^*} = \frac{m\log(\frac{m}{D})}{\frac{m}{D}-1} < m, \tag{23}$$

where we have used that $x\log x \geq x-1$ for $x > 1$. By letting $x = m/D$, we have $x > 1$ and prove the left hand side of the inequality; similarly by $\log(x) < x - 1$ for $x > 1$, we have the right hand side of the inequality.

Substituting δ_∞^* to (21), we obtain the optimized EDD (with respect to δ_∞) as a function of w:

$$\mathbb{E}_0(\mathcal{T}_C) = \frac{2\log\gamma\big(1+o(1)\big)}{mg(D/m)} + w. \tag{24}$$

where $g(x) = x - 1 - \log(x)$. From (23) we know $D/m < 1$. It should be noted that in the expression above, the dependence on w also comes from D defined in (13).

Now, using (24) we can further find the optimized window size w^* (for a finite γ) such that the EDD in (24) is minimized. After taking the derivative with respect to w, note that $1/w^2$ is in the order of $o(1/w)$ and thus can be ignored when w is large. Thus we omit all the higher-order terms of w in D and obtain the following result:

Lemma 3 (Optimal Window Size and Drift Parameter): For each ARL level γ, the optimal window size w^* which minimizes the Expected Detection Delay (EDD) in (21) is given by:

$$w^* = \frac{\sqrt{2\,(1/\Delta - 1)\left(\sum_{i=1}^m \frac{\lambda_i}{\sigma^2+\lambda_i}B_i\right)}}{mg\,(\Delta)} \cdot \sqrt{\log\gamma}. \tag{25}$$

where $\Delta := \frac{1}{m}\sum_{i=1}^m \frac{\lambda_i}{\sigma^2+\lambda_i} \in (0,1)$. Substituting the optimal w^* and δ_∞^* back into (20) gives the optimal value for drift parameter:

$$d^* = \frac{m\log(m/D^*)}{m/D^* - 1}, \tag{26}$$

where $D^* = \sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\big(1 - B_i/w^* + 3B_i^2/w^{*2}\big)$.

The implication of the lemma above is that the optimal window length w^* grows in the order of $\sqrt{\log\gamma}$. Clearly, w^* in (25) is positive; in particular, the denominator in (25) is non-negative, since $g(x) > 0$ for $x \in (0,1)$ and $\Delta \in (0,1)$. Moreover, it can be shown that when w^* goes to infinity (since the optimal $w^* = \sqrt{\log\gamma}$, $\gamma \to \infty$ in our asymptotic analysis), then $D^* \to m\Delta$, and d^* in (26) tends to $\frac{-m\log\Delta}{1/\Delta-1}$. Figure 4 shows numerical examples of choosing the optimized window size w.

D. First-Order Optimality

Finally, we show the asymptotic property of Spectral-CUSUM, the EDD ratio relative to the exact CUSUM, under the same ARL constraint; since the exact CUSUM is shown to be optimal and achieves the smallest EDD under a constant ARL constraint as shown in the classic results [30] and [31]. Note that for the "equalizer" argument to work, thus we can obtain ARL and EDD expressions for Spectral CUSUM, we only need the drift term d to be related to δ_∞ through (20). In the following Theorem, we show the resulted EDD under optimal window w^* matches the order of EDD of the Exact CUSUM.

Theorem 2 (Asymptotic Optimality of Spectral-CUSUM): Given an ARL lower bound γ, the ratio between the EDD of the Spectral-CUSUM and the EDD of the exact CUSUM satisfies:

$$\frac{\mathbb{E}_0[\mathcal{T}_C]}{\mathbb{E}_0[T_C]} = \mathcal{O}(1) + \frac{\mathcal{I}_0 w}{2\log\gamma}\left(1 + \mathcal{O}\left(\frac{1}{\log\gamma}\right)\right).$$

When $\gamma \to \infty$, for fixed m, n, σ^2 bounded away from zero, and $w = \sqrt{\log\gamma}$, this ratio $\mathbb{E}_0[\mathcal{T}_C]/\mathbb{E}_0[T_C]$ tends to an absolute constant, and then the Spectral-CUSUM is asymptotically first-order optimum.

V. Efficient Computation by Subspace Tracking

In this section, we present an efficient algorithm to keep track of the underlying community structure even if we do not assume the form of the graph structure. This can be treated as a complementary approach to the proposed Spectral-CUSUM method when we have low confidence or lack of pre-change observations. Inspired by the GROUSE algorithm [39], which implements stochastic gradient descent on the Grassmann manifold to update subspaces at each time slot, we design a subspace tracking algorithm to update the estimated subspace, denoted as $\widehat{Q}_t$, each time a new graph observation G_t arrives. Here we consider graph observation instead of vector observation v_t on nodes. Similar to (7), the graph observation at time t can be represented using vector observation as $G_t = v_t v_t^\top$.

If we treat each observation as a static graph and apply the spectral clustering method in [40] at each time, we would get a

sequence of $\widehat{Q}_t$ independently, which is quite time-consuming and not guaranteed to converge if considering noisy cases. Instead, we can perform an updating procedure each time on $Q \in \mathbb{R}^{n\times m}$ on a Grassmann manifold. The Grassmannian denoted as $\mathrm{Gr}(m,n)$ is a space that contains all m-dimensional linear subspaces of the n-dimensional vector space. As a compact Grassmann manifold, its geodesics can be computed as indicated in [41]. Our target matrix Q can then be represented as a point in the Grassmann manifold. Thus the optimization problem becomes finding optimized $\widehat{Q}$ such that:

$$\widehat{Q} = \arg\min_{Q\in\mathbb{R}^{n\times m}} \sum_t \mathrm{tr}(Q^\top G_t Q),\ s.t.\ Q^\top Q = I.$$

We consider the problem for every time slot and define function $f_t(Q) = \mathrm{tr}(Q^\top G_t Q)$ and the derivative of f_t with respect to Q is [42]:

$$\frac{df_t}{dQ} = \frac{d(\mathrm{tr}(Q^\top G_t Q))}{dQ} = (G_t + G_t^\top)Q. \tag{27}$$

Then we use Equation (2.70) in [41] to get the gradient of function $f_t(Q)$ on Grassmann manifold from (27):

$$\nabla f_t = (I - QQ^\top)\frac{df_t}{dQ} = (I - QQ^\top)(G_t + G_t^\top)Q.$$

Gradient descent algorithm along a Grassmann manifold is given by equation (2.65) in [41], proving that it is a function of the singular values and vectors of ∇f_t, so suppose we have got the reduced Singular Value Decomposition (rSVD) of $-\nabla f_t = U\Sigma V^\top$ where only the top-k eigenvalues and eigenvectors are kept so that the computational cost is much reduced, we can write the updating function with a step size η as:

$$Q(\eta) = \begin{pmatrix} QV & U \end{pmatrix} \begin{pmatrix} \cos\ \Sigma\eta \\ \sin\ \Sigma\eta \end{pmatrix} V^\top. \tag{28}$$

Here we update Q with a step size η to get closer to the local minimum on the Grassmann manifold. The complete algorithm is shown in Algorithm 2.

Algorithm 2 Subspace Tracking for Spectral-CUSUM

Input: Weighted adjacency matrix at time t denoted as G_t, the total number of iterations T, number of communities m, a set of step sizes η_t.

Output: Subspace Representation Q of the graph.

1: Initialize Q randomly, introduce y as a T-dimensional vector with all entries equal to 0.
2: **for** $t = 1, \ldots, T$ **do**
3: Observe current adjacency matrix G_t.
4: Compute $\nabla f_t = (I - QQ^\top)(G_t + G_t^\top)Q$.
5: Compute SVD of $-\nabla f_t = U\Sigma V^\top$.
6: Update Q using (28) with step size η_t.
7: Update $y(t) = \mathrm{tr}(Q^\top G_t Q)$
8: Run CUSUM detection procedure on y and get detection statistic S.
9: **return** S

Choice of Step Size. For our problem, constant step size and decreasing step size can be efficient. The constant step is slower initially, but it is more stable to detect changes in community structures. In our numerical experiments, we take constant step size $\eta = 0.01$.

Complexity. In practice, we notice that in this algorithm, we have to perform SVD for each iteration, which is quite time-consuming. To accelerate the whole algorithm, we use the incremental SVD algorithm [43] in each iteration.

VI. Numerical Experiments

In this section, we compare our method with the state-of-the-art and discuss their numerical results on both synthetic and real data sets. Since a graph can be considered as a discrete approximation to a manifold [44], we also show that our model can achieve promising performance on dynamic manifold data.

A. Methods for Comparison

In our experiments, we compare our method with four other baseline approaches, including (1) Generalized Likelihood Ratio procedure based on vectorized data (`Vectorized GLR`) (2) Hotelling's T-squared CUSUM (`Hotelling`); (3) Single eigenvector procedure (`SC`, $m = 1$); (4) Rank-m Subspace Tracking of Section V (`SGD`); and (5) Exact CUSUM as a sanity check. The detailed explanation of these baseline methods is as follows:

(1) *Generalized Likelihood Ratio (GLR) procedure based on vectorized data.* This baseline method completely ignores the topology properties of the adjacency graph G_t and vectorizes it as $g_t = \mathrm{vec}(G_t)$ such that the previous hypothesis test (2) becomes:

$$\begin{aligned} H_0 &: g_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_{n^2}) \quad t = 1, 2, \ldots, \tau \\ H_1 &: g_t \overset{\text{iid}}{\sim} \mathcal{N}(h, \sigma^2 I_{n^2}) \quad t = \tau+1, \tau+2, \ldots \end{aligned} \tag{29}$$

where $h = \mathrm{vec}(AA^\top)$. Then classical result from [30], [45] gives the optimal stopping criteria for GLR procedure:

$$T_{\mathrm{GLR}} = \inf\left\{t : \max_{t-w<k<t} \frac{(\sum_{i=k+1}^t \|g_i\|)^2}{t-k} > b\right\}. \tag{30}$$

(2) *Hotelling's T-squared CUSUM.* We also compare our method with Hotelling's T-squared statistic, which was introduced in [46]. The way to implement this is to calculate a pre-change sample mean $\hat{\mu}_0$ and sample covariance $\widehat{\Sigma}_0$, then keep track of the current sample mean $\bar{\mu}_{t-w,t}$ via a sliding window. We also need to pre-define the drift parameter d^{H_2}, which can be obtained from historical data. Thus the detection statistic is given by:

$$S_t^{H_2} = (S_{t-1}^{H_2})^+ + (\bar{\mu}_{t-w,t} - \hat{\mu}_0)^\top \widehat{\Sigma}_0^{-1}(\bar{\mu}_{t-w,t} - \hat{\mu}_0) - d^{H_2}.$$

Then the optimal stopping time for Hotelling's statistics is given by:

$$T^{H_2} = \inf\left\{t > 0 : S_t^{H_2} \geq b\right\}. \tag{31}$$

(3) *Single eigenvector procedure of Spectral-CUSUM.* This baseline method is almost the same as the proposed Spectral-CUSUM with the only difference of fixing community size $m = 1$. We use it to illustrate the importance of finding the optimal community size.

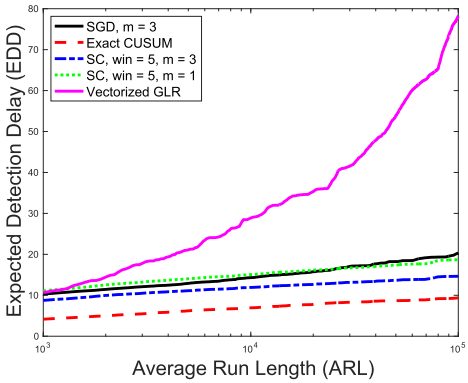


Fig. 3. Comparison of the single eigenvector procedure and Exact-CUSUM procedure for emergence problem. Community size fixed and σ is set to be 5.

(4) *Rank-m Subspace Tracking.* This is proposed in Section V as a numerical alternative method for our Spectral CUSUM. So we list it as another baseline.

For synthetic data, we study the relationship between ARL and EDD. For real data, we study EDD only since we cannot implement simulation to decide the average run length if the history data under pre-change distribution is not sufficient.

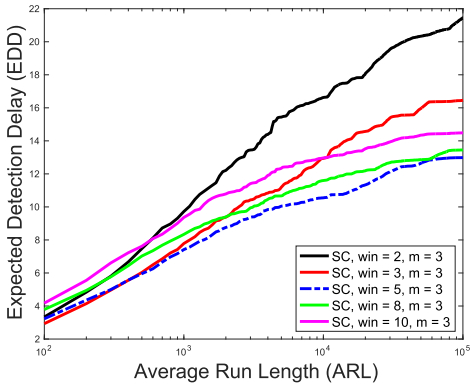
B. Synthetic Experiments

We devise three synthetic experiments to study the performance of our method on the emergence and the switching membership problems.

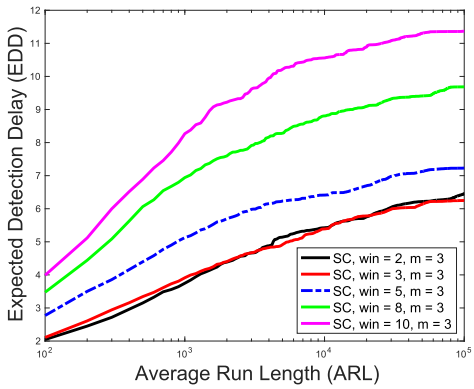
Synthetic data 1 is used for evaluating our method of detecting emerging communities. In this experiment, we set $\sigma = 5$ and assume a dynamic graph that contains 50 nodes without any community structure in the beginning. When the change occurs at time $t =$, three communities are formed, containing 10, 10, and 15 nodes, respectively. In other words, we have $m = 0$ before t and $m^* = 3$ after t. The optimal window size can be found according to (25), where $w = 5$.

The experimental result shows that our method significantly outperforms other baselines. Figure 3 presents the ARL of all the methods on the synthetic data 1 with different choices of community density and window size. As can be observed, our method (blue line) works much better (the lower, the better) than the baselines and is closest to the sanity check method Exact-CUSUM. A comparison of how community size would affect the choice of optimal window size is also shown in the figure 4(b) where we increase the size of each emerging community by 5, referred to as larger emerging communities. It can be inferred that the change becomes easier to detect in larger communities. Thus, the optimal window size decreases correspondingly. In figure 5, we show how different noise levels could infect the general detection delay, which leads to the conclusion that a larger noise level leads to a longer average detection delay. Moreover, when ARL is relatively small, a large noise level would not have much effect on the EDD. However, if the ARL is larger, the EDD will be increased significantly for higher noise levels.

Synthetic data 2 is used for evaluating our method of detecting switching membership. For the design of the switching membership experiment, the number of nodes is 50, and the



(a) Smaller emerging communities



(b) Larger emerging communities

Fig. 4. ARL vs. EDD plot for optimized window size for Spectral-CUSUM with smaller (a) and larger (b) emerging communities.

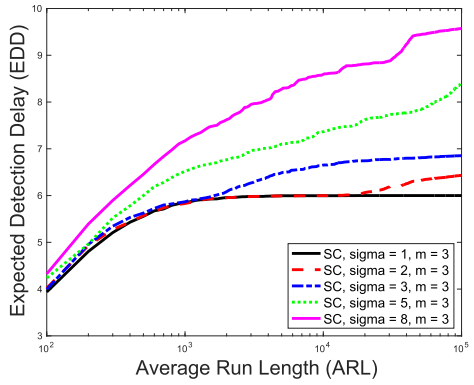


Fig. 5. Comparison of the different noise levels for emergence problem. Community size is fixed, and values of σ are set to be 1,2,3,5,8, respectively.

size of the three communities remains the same: 10, 10, and 15 nodes. However, the membership for all three communities will change. A comparison between SGD, Exact-CUSUM, and Spectral-CUSUM is shown in Figure 6. We can see that the single eigenvector procedure ($m = 1$) performs much worse than Spectral-CUSUM using the correct potential community size ($m = 3$). The performance gain of spectral-CUSUM in the switching membership setup is larger than that in the emerging subspace problems compared with the Exact-CUSUM procedure. Moreover, the EDD of SGD is growing quite fast with the increase of ARL, which indicates that it is noise-sensitive.

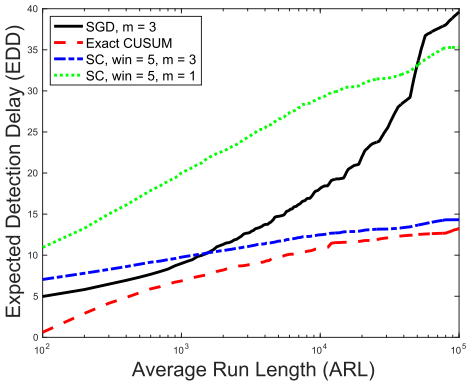


Fig. 6. Comparison of the single eigenvector procedure and Exact-CUSUM procedure for switching membership problem. Community Size fixed to be 3. The total number of nodes in the network is set to 50. Before the change, each community has a size of 10, and after the change, each community's size increases to 15 by absorbing nodes that do not belong to any community before the change happens.

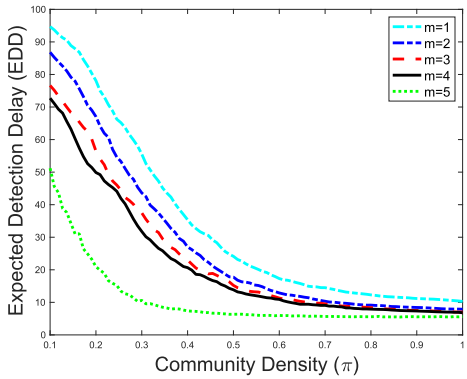
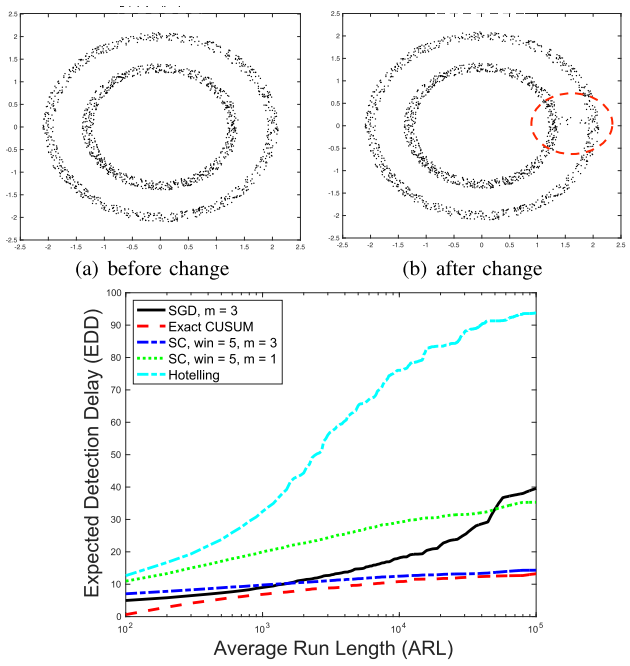


Fig. 7. Relationship between the community density and EDD under different settings of the number of communities $m = \{1, 2 \ldots, 5\}$. for emergence community problem. Smaller m leads to higher EDD, and a larger community density π after the change would have also decreased the EDD.

Here we do not compare with the vectorized GLR procedure since it does not work in the switching membership problem. The reason is that the switching membership change does not lead to the increase or decrease of average edge weights; thus, the change cannot be detected without considering the graph topology.

Synthetic data 3 is used to study the impact of network topology. We also explore the influence of community density by assuming that for community i, the probability of forming an edge within the community is p_i. Then by changing the value of p_i under an emergence community scenario, the density of the community can change. From Figure 7 we see that a denser community structure after the change leads to a more accurate detection. However, we can also observe from the drastically decreasing curve that the detection can be performed more accurately if the community density $p_i \geq 0.4$. This provides an approach to discovering a community through the Spectral CUSUM method. In addition, we can see that the detection power increases with a larger m and the increase in community density. This result is consistent with our intuition since we can always collect more useful information when $m \leq k$, and the denser the community is, the more signal that reveals the structure would be much stronger, leading to quicker change detection.



(a) before change (b) after change

(c) ARL VS EDD plot

Fig. 8. A case that two separate rings have some overlapping nodes by a bridge after the change: (a) Before the change, there are two rings structures that form two communities; (b) After the change, two communities remain the same, but some of their members join another community, leading to the overlapping of two communities. (c) The ARL VS EDD results show that Spectral-CUSUM when $m = 4$ outperforms other procedures under larger ARL, but SGD is better for smaller ARL.

C. Detecting Changes of Manifold Structure

This section shows that our detection algorithm also works for detecting structural changes in the manifold. We first consider a one-dimensional manifold on two-dimensional rings. Before the change happens at time $t = 200$, the two rings are separately located, while after the change, there forms a bridge between the two rings, as shown in Figure 8.

We use the Isomap [47] to calculate the adjacency graph, which is sparse. Given the adjacency matrix, we set the threshold to eliminate all the edges whose distances run below the threshold. Thus, inside one community, any pair of members can be reached via a few steps. Using the Spectral-CUSUM, we show that, in this case, we can detect the change quite quickly in figure 8(c).

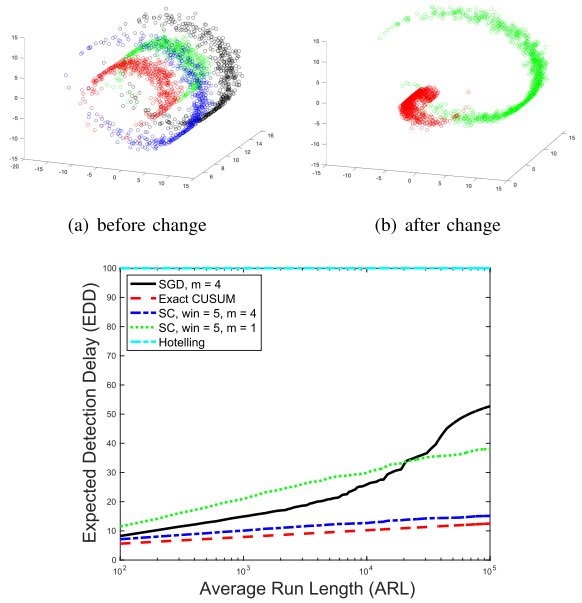
Another classical manifold is Swiss Roll data [48], which is a two-dimensional structure lying in a three-dimensional space. The experiment is designed such that four Swiss rolls merge into two Swiss rolls after the change, as shown in Figure 9. We aim to use this to show that our method can deal with various data topological structure changes based on a choice of similarity measure. We see from the result that when m equals its potential community size, the performance is quite close to the optimal Exact-CUSUM.

D. Yellowstone Seismic Sensor Network Data

We further consider a seismic sensor network data set adapted from [49]. The sensors are placed in different locations to measure signals around the Old Faithful Geyser in

TABLE I
DETECTION DELAYS UNDER DIFFERENT # COMMUNITIES (m) FOR SEISMIC DATA

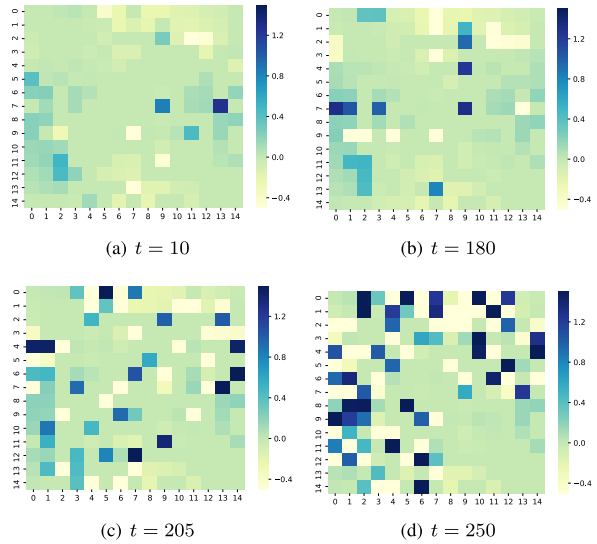
# communities (m)	1	2	3	4	5	6	7	8	9
Detection Delay	23	20	17	4	3	3	2	2	2



(a) before change (b) after change

(c) ARL VS EDD plot

Fig. 9. A case that four Swiss rolls merge into two communities after the change: (a) Before the change, there are four swiss roll structures which form four different communities; (b) After the change, two communities disappear and merge into the other communities. Note that in this experiment, Hotelling's T-squared statistics cannot detect the change at all.

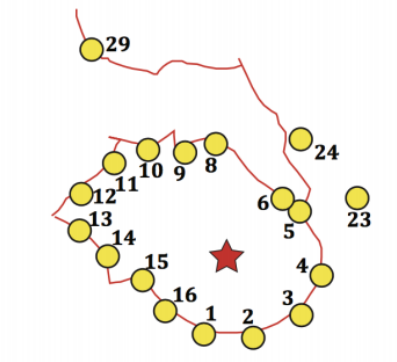


(a) $t = 10$ (b) $t = 180$

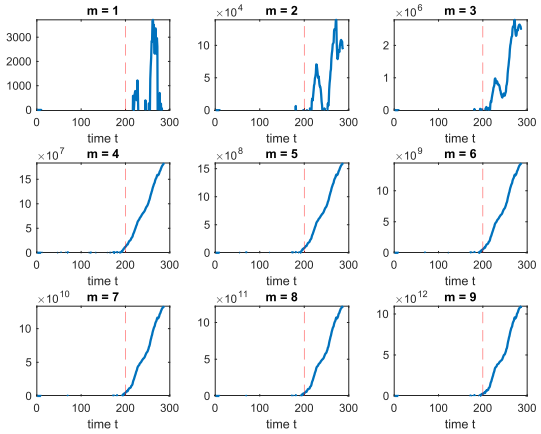
(c) $t = 205$ (d) $t = 250$

Fig. 10. Correlation matrix for seismic sensors at different times $t = 10, 180, 205, 250$. We can see clearly from the above matrices that there's a big change before and after $t = 200$, which is the ground-truth change-point in the seismic network, causing an emergence community scenario.

Yellowstone National Park. The total number of sensors is 19 in this case, but we removed four sensor signals since they failed to work during data collection. We then observe a sequence of sensor signals and translate each one into a dynamic cross-correlation graph. At the very beginning, the cross-correlation between each pair of sensors is low, which



(a) Seismic topology visualization



(b) Detection statistics

Fig. 11. The topology of seismic sensors and detection statistics for sequential seismic data with an outburst of community structure changes. In (a), the total number of 19 seismic sensors roughly forms a circle, and signals can be observed on each sensor [50]. Thus the community would be formulated when an earthquake happens, making several sensors correlated. In (b), we applied our Spectral-CUSUM statistics on sensor signals, and we know the ground-truth earthquake happens at time 200. The results demonstrate the ability of the algorithm to detect such changes very quickly.

means they are not related. The emergence of community happens in the middle of the sequence. Such change will cause the infected sensors to generate similar signals, which lead to a higher correlation magnitude between them. In contrast, an unaffected sensor still generates random noises, thus having a low correlation with other sensors. Consequently, a community containing all the affected sensors emerges after the change happens. We visualize the correlation matrix for better understanding in Figure 10. In this case, the true changepoint time is roughly known, corresponding to the geyser eruption time.

We apply our Spectral-CUSUM procedure to the dataset with choices of different potential community sizes m. Figure 11 (a) shows the sensor locations, and the results of detection statistics using Spectral-CUSUM procedure are shown in Figure 11 (b). It can be seen clearly that setting

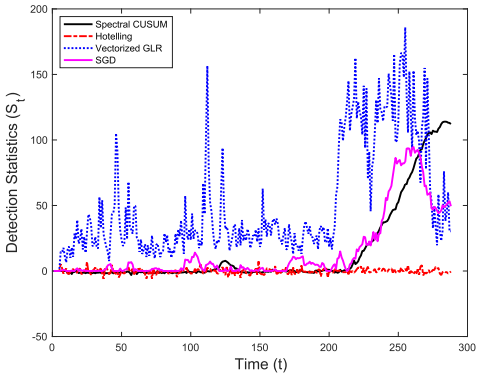


Fig. 12. Comparison of our Spectral-CUSUM procedure with other baseline methods, including Hotelling's statistics, SGD, and vectorized GLR.

potential community size $m \geq 4$ gives better performance on detecting change at around $t = 200$, which corresponds to a geyser eruption time – treated as a true change-point time. The result shows that the event can be detected without false alarms when $m \geq 4$, which reveals that this seismic network's true underlying community size is about 4. The detection delay is shown in Table I. This shows that our detection statistics can be useful when the underlying graph structure is unknown.

In addition, we compare our Spectral-CUSUM with other methods on this real data. We can see from Figure 12 that Spectral-CUSUM and SGD have the most obvious abrupt change around the true change-point time 200. However, vectorized GLR and Hotelling's T-squared statistics can not detect such a change.

VII. Conclusion

We present a novel Spectral-CUSUM procedure for detecting underlying community changes through noisy observations. We provide the first-order asymptotic optimality of Spectral-CUSUM under the optimal choice of parameters. We also present an efficient online computation procedure to evaluate the Spectral-CUSUM statistic without remembering all data based on subspace tracking. Experimental results on both synthetic and real-world data demonstrate the superior performance of our procedure compared with alternative methods.

Appendix

Deriving the Bounds for Terms in (IV-A)*:* Recall $\rho_i = 1/(\sigma^2 + |C_i|)$. First, the Frobenius norm for the second term in the covariance matrix of the estimation error:

$$\begin{aligned}
&\frac{\rho_i/\sigma^2}{(\rho_i - 1/\sigma^2)^2}\|I - UU^\top\|_F \\
&= \frac{\sigma^2/\rho_i}{(\sigma^2 - 1/\rho_i)^2}\|I - UU^\top\|_F \\
&= \frac{\sigma^2(\sigma^2 + |C_i|)}{|C_i|^2}\|I - UU^\top\|_F \\
&= \frac{\sigma^2}{|C_i|}\left(\frac{\sigma^2}{|C_i|} + 1\right)\sqrt{n-m},
\end{aligned}$$

where we have used $\|I - UU^\top\|_F^2 = \text{tr}[(I - UU^\top)(I - UU^\top)] = \text{tr}(I - UU^\top) = n - m$ since U is a n-by-m orthonomal matrix and $I - UU^\top$ is a projection matrix. Assume $\sigma^2/|C_i| < \epsilon$, $\epsilon > 0$, we have the last equation above is upper bounded by $\epsilon(1+\epsilon)\sqrt{n-m}$.

On the other hand, it can be shown that the square Frobenius norm of the first term $\sum_{k=1,k\neq i}^{m} \frac{\rho_i\rho_k}{(\rho_i-\rho_k)^2} u_k u_k^\top$ is given by (and further bounded by)

$$\begin{aligned}
&\sum_{k=1,k\neq i}^{m} \frac{\rho_i^2\rho_k^2}{(\rho_i-\rho_k)^4} \\
=&\sum_{k\neq i} \frac{\frac{1}{\rho_i^2}\frac{1}{\rho_k^2}}{\left(\frac{1}{\rho_k} - \frac{1}{\rho_i}\right)^4} \\
=&\sum_{k\neq i} \frac{(\sigma^2 + |C_i|)^2(\sigma^2 + |C_k|)^2}{(|C_k| - |C_i|)^4} \\
=&\sum_{k\neq i} \frac{\left(\frac{\sigma^2}{|C_i|} + 1\right)^2\left(\frac{\sigma^2}{|C_k|} + 1\right)^2}{\left(\sqrt{\frac{|C_k|}{|C_i|}} - \sqrt{\frac{|C_i|}{|C_k|}}\right)^4} \\
\geq&\frac{m-1}{\left(1 + \theta - \frac{1}{1+\theta}\right)^4},
\end{aligned} \tag{32}$$

where the last inequality is based on Assumption IV.1(2).

Thus, to have the first term dominate the second term, we need

$$\epsilon(1+\epsilon) < \eta \cdot \sqrt{\frac{m-1}{n-m}} \cdot \frac{1}{\left(1 + \theta - \frac{1}{1+\theta}\right)^2},$$

where $\eta \in (0,1)$ controls the relative magnitude of the terms, which gives us the condition

$$\epsilon < \left(\frac{1}{4} + \eta \cdot \sqrt{\frac{m-1}{n-m}} \cdot \frac{1}{\left(1 + \theta - \frac{1}{1+\theta}\right)^2}\right)^{1/2} - \frac{1}{2}.$$

Together with $\frac{\sigma^2}{|C_i|} < \epsilon$, we can derive the necessary upper bound. Note that when θ is relatively small, this bound can be easily satisfied. □

Proof of Lemma 1: Since $\widehat{A}_t$ is estimated using data from $t+1$ to $t+w$, it is *independent* from v_t. This independence property allows for the straightforward computation of the two expectations in Lemma 1 and contributes towards the proper selection of drift d. Recall we denote $\hat{\Lambda}_t = \text{diag}(\hat{\rho}_{t1}, \ldots, \hat{\rho}_{tm})$ as the smallest m eigenvalues of the sample covariance matrix, and $\widehat{U}_t$ as the corresponding eigenvectors. Note that under the pre-change distribution we can write:

$$\begin{aligned}
&\mathbb{E}_\infty[v_t^\top \widehat{A}_t \widehat{A}_t^\top v_t] \\
=&\mathbb{E}_\infty[v_t^\top \widehat{U}_t \hat{\Lambda}_t^{-1} \widehat{U}_t^\top v_t] = \mathbb{E}_\infty[v_t^\top (\sum_{i=1}^{m} \frac{1}{\hat{\rho}_{ti}} \hat{u}_{ti}\hat{u}_{ti}^\top) v_t] \\
=&\sum_{i=1}^{m} \mathbb{E}_\infty[\frac{1}{\hat{\rho}_{ti}}(\hat{u}_{ti}^\top v_t)^2] = \sum_{i=1}^{m} \mathbb{E}_\infty[\frac{1}{\hat{\rho}_{ti}} \hat{u}_{ti}^\top \mathbb{E}_\infty[v_t v_t^\top]\hat{u}_{ti}] \\
=&\frac{1}{\sigma^2}\sum_{i=1}^{m} \mathbb{E}_\infty[\frac{1}{\hat{\rho}_{ti}} \hat{u}_{ti}^\top \hat{u}_{ti}] = \frac{1}{\sigma^2}\sum_{i=1}^{m} \mathbb{E}_\infty[\frac{1}{\hat{\rho}_{ti}}].
\end{aligned}$$

Note that under the pre-change measure, the estimated $\widehat{U}_t$ and $(\widehat{\Lambda}_t)^{-1}$ are eigenvectors and eigenvalues for the sample

covariance matrix constructed by samples generated from $\mathcal{N}(0,(1/\sigma^2)I)$. We approximate the above expectation as $\mathbb{E}_\infty[1/\hat{\rho}_{ti}] \approx \sigma^2$ using the eigenvalues of the ground truth sample covariance matrix. Note that such approximation is valid when the number of samples (i.e., the sliding window size w) is large enough, under such cases, the sample eigenvalues will highly concentrate around the true eigenvalues. As we will show later, the optimal window size w is indeed sufficiently large (in the order of $\sqrt{\log\gamma}$) in the asymptotic regime we consider. Following the discussions above, we write:

$$\mathbb{E}_\infty[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t] \approx m. \tag{33}$$

Similarly, we can derive the results for post-change distribution:

$$\begin{aligned}
&\mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]\\
=&\mathbb{E}_0[v_t^\top \widehat{U}_t\Lambda_t^{-1}\widehat{U}_t^\top v_t] = \mathbb{E}_0[v_t^\top(\sum_{i=1}^m \frac{1}{\hat{\rho}_{ti}}\hat{u}_{ti}\hat{u}_{ti}^\top)v_t]\\
=&\sum_{i=1}^m \mathbb{E}_0[\frac{1}{\hat{\rho}_{ti}}(\hat{u}_{ti}^\top v_t)^2] = \sum_{i=1}^m \mathbb{E}_0[\frac{1}{\hat{\rho}_{ti}}\hat{u}_{ti}^\top \mathbb{E}_0[v_t v_t^\top]\hat{u}_{ti}]\\
=&\sum_{i=1}^m \mathbb{E}_0[\frac{1}{\hat{\rho}_{ti}}\hat{u}_{ti}^\top(\sum_{j=1}^n \rho_j u_j u_j^\top)\hat{u}_{ti}]\\
=&\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}_0[\frac{\rho_j}{\hat{\rho}_{ti}}(\hat{u}_{ti}^\top u_j)^2],
\end{aligned}$$

where $1/\hat{\rho}_1,\ldots,1/\hat{\rho}_m$ are the estimated eigenvalues of $\widehat{A}\widehat{A}^\top$ and $\rho_1,\ldots,\rho_n$ are the eigenvalues of the true covariance matrix $(AA^T+\sigma^2 I)^{-1}$ so that it can be written as $\mathbb{E}_0[v_t v_t^\top] = \sum_{j=1}^n \rho_j u_j u_j^\top$. As stated in Theorem 1, the estimated eigenvalue and eigenvector from the sample covariance matrix are independent, and Theorem 1 characterizes the estimation error $e_{it} = \hat{\varphi}_{ti} - u_i$, which satisfies $\sqrt{w}e_{it} \xrightarrow{d} \mathcal{N}\left(0, \sum_{k=1,k\neq i}^m \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2}u_k u_k^\top\right)$. Asymptotically, we have $\mathbb{E}_0[u_i^\top e_{it}] = u_i^\top \mathbb{E}_0[e_{it}] \to 0$ and

$$\mathbb{P}(u_i^\top e_{it}e_{it}^\top u_i > \epsilon) \le \frac{\mathbb{E}_0[u_i^\top e_{it}e_{it}^\top u_i]}{\epsilon} = \frac{u_i^\top \mathrm{Cov}(e_{it})u_i}{\epsilon} \to 0,$$

thus u_i is perpendicular to the error term e_{it} with probability 1.

Since the estimated eigenvector is normalized to be unit-norm, we have the following:

$$\hat{u}_{it} = \frac{\hat{\varphi}_{ti}}{\|\hat{\varphi}_{ti}\|} = \frac{u_i + e_{it}}{\|u_i + e_{it}\|}.$$

First of all, due to the orthogonality of different eigenvectors, we have:

$$\sum_{i=1}^m\sum_{j=1}^n \mathbb{E}_0[\frac{\rho_j}{\hat{\rho}_{ti}}(\hat{u}_{ti}^\top u_j)^2] = \sum_{i=1}^m \mathbb{E}_0[\frac{\rho_i}{\hat{\rho}_{ti}}(\hat{u}_{ti}^\top u_i)^2],$$

where u_i, $i=1,\ldots,m$, are the eigenvectors corresponding to the smallest m eigenvalues of the true covariance matrix. Then we examine the term $\rho_i/\hat{\rho}_{ti}$, note that the true eigenvalue of the covariance matrix is $\rho_i = 1/(|C_i| + \sigma^2) = 1/(\lambda_i+\sigma^2)$ and the estimated eigenvalues of $(\widehat{A}\widehat{A}^\top)^{-1}$ is approximately $1/\lambda_i$. Thus we have $\mathbb{E}_0[\rho_i/\hat{\rho}_{ti}] = \lambda_i/(\lambda_i+\sigma^2)$, and of course such approximation is under the asymptotic case where the window size w is sufficiently large.

Combining these together, we have

$$\begin{aligned}
&\mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]\\
=&\sum_{i=1}^m \mathbb{E}_0[\frac{\rho_i}{\hat{\rho}_{ti}}(\hat{u}_{ti}^\top u_i)^2] = \sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\mathbb{E}_0\left[\frac{(\|u_i\|^2 + e_{it}^\top u_i)^2}{\|u_i+e_{it}\|^2}\right]\\
=&\sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\mathbb{E}_0\left[\frac{(1+e_{it}^\top u_i)^2}{1+\|e_{it}\|^2}\right] = \sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\mathbb{E}_0\left[\frac{1}{1+\|e_{it}\|^2}\right]\\
=&\sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\mathbb{E}_0\left[\frac{1-\|e_{it}\|^4+\|e_{it}\|^4}{1+\|e_{it}\|^2}\right]\\
=&\sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\mathbb{E}_0\left[\frac{(1+\|e_{it}\|^2)(1-\|e_{it}\|^2)+\|e_{it}\|^4}{1+\|e_{it}\|^2}\right]\\
=&\sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\left\{1-\mathbb{E}_0[\|e_{it}\|^2] + \mathbb{E}_0\left[\frac{\|e_{it}\|^4}{1+\|e_{it}\|^2}\right]\right\}.
\end{aligned}$$

The second equality is by substituting $\hat{u}_{it} = (u_i + e_{it})/\|u_i+e_{it}\|$, and the third equality is due to the fact that u_i is an unit-norm vector and u_i is perpendicular to e_{it} with probability 1. For the two expectations above, using Gaussian approximation from Theorem 1, we have

$$\mathbb{E}_0[\|e_{it}\|^2] = \frac{1}{w}\sum_{k=1,k\neq i}^m \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2} = \frac{B_i}{w},$$

and to estimate $\mathbb{E}_0\big[\|e_{it}\|^4/(1+\|e_{it}\|^2)\big]$, we provide the following lemma:

Lemma 4: For a multivariate Gaussian random variable $x \sim \mathcal{N}(0,\Sigma)$ where $\Sigma_{ij} = \sigma_{ij}^2$, we have

$$\mathbb{E}(\|x\|_2^4) \le 3(\mathrm{tr}(\Sigma))^2.$$

We can derive this lemma using the properties of uni-variate Gaussian distribution where $x_i \sim \mathcal{N}(0,\sigma_{ii}^2)$ $\mathbb{E}(x_i^4) = 3\sigma_{ii}^4$ and Cauchy-Schwarz inequality (details omitted). As a result, according to Lemma 4, we can conclude that random variable e_{it} that follows a Gaussian distribution have $\mathbb{E}_0[\|e_{it}\|^4] = \mathbb{E}_0[(\|e_{it}\|^2)^2] \le 3[\mathrm{tr}(\mathrm{Cov}(e_{it}))]^2 = 3B_i^2/w^2$. Then we have:

$$\begin{aligned}
0 \le& \mathbb{E}_0\left[\frac{\|e_{it}\|^4}{1+\|e_{it}\|^2}\right]\\
\le& \mathbb{E}_0[\|e_{it}\|^4] \le \frac{3}{w^2}\left(\sum_{k=1,k\neq i}^m \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2}\right)^2 = \frac{3B_i^2}{w^2}.
\end{aligned}$$

Therefore, we get the desired approximation to the post-change expectation:

$$\begin{aligned}
\widetilde{D} =& \sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\left(1-\frac{B_i}{w}\right) \le \mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]\\
\le& \sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\left(1-\frac{B_i}{w}+\frac{3B_i^2}{w^2}\right) = D.
\end{aligned}$$

From above, we see that when the window size w becomes large, $\widetilde{D}/D \to 1$, we then have $\mathbb{E}_0[v_t^\top \widehat{A}_t\widehat{A}_t^\top v_t]$ can be well approximated by D. □

Proof of Lemma 2: By definition of KL divergence for emerging subspace case,

$$\begin{aligned}\mathcal{I}_0 &= \mathbb{E}_0\left[\log\frac{f_0(v_t)}{f_\infty(v_t)}\right]\\ &= \mathbb{E}_0\left[-\frac{1}{2}v_t^\top AA^\top v_t + \frac{1}{2}\log\frac{\det(AA^\top+\sigma^2 I)}{\sigma^{2n}}\right]\\ &= -\frac{1}{2}\mathbb{E}_0\left[v_t^\top AA^\top v_t\right] + \frac{1}{2}\log\frac{\det(AA^\top+\sigma^2 I)}{\sigma^{2n}}.\end{aligned}$$

Since A is not a full-rank matrix, the eigendecomposition of $AA^\top = U\Lambda U^\top$ has dimension $U \in \mathbb{R}^{n\times m}$ and $\Lambda \in \mathbb{R}^{m\times m}$. By adding basis ($\tilde{U} \in \mathbb{R}^{(n\times(n-m))}$) from nullspace to make U' to be a square matrix, we have the following:

$$U' = \begin{bmatrix}U & \tilde{U}\end{bmatrix} \in \mathbb{R}^{n\times n}, \Lambda' = \begin{bmatrix}\lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_m & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0\end{bmatrix} \in \mathbb{R}^{n\times n},$$

where U' is an orthogonal matrix of rank n where $U'^\top U' = I$ and $AA^\top = U'\Lambda' U'^\top$. Then the first term of $\mathcal{I}_0$ can be derived as:

$$\begin{aligned}&\mathbb{E}_0\left[v_t^\top AA^\top v_t\right]\\ =&\mathbb{E}_0\left[\text{tr}(AA^\top v_t v_t^\top)\right] = \text{tr}\left[AA^\top \mathbb{E}_0(vv^\top)\right]\\ =&\text{tr}\left[AA^\top(AA^\top+\sigma^2 I)^{-1}\right]\\ =&\text{tr}\left[U'\Lambda' U'^\top(U'(\Lambda'+\sigma^2 I)^{-1}U'^\top)\right]\\ =&\text{tr}\left[U'\Lambda'(\Lambda'+\sigma^2 I)^{-1}U'^\top\right] = \text{tr}\left[\Lambda'(\Lambda'+\sigma^2 I)^{-1}U'^\top U'\right]\\ =&\text{tr}\left[\Lambda'(\Lambda'+\sigma^2 I)^{-1}\right] = \sum_{i=1}^m \frac{\lambda_i}{\sigma^2+\lambda_i}.\end{aligned}$$

Moreover, the second term becomes

$$\begin{aligned}\frac{1}{2}\log\frac{\det(AA^\top+\sigma^2 I)}{\sigma^{2n}} &= \frac{1}{2}\sum_{i=1}^m \log\left(\frac{\lambda_i}{\sigma^2}+1\right)\\ &= -\frac{1}{2}\sum_{i=1}^m \log\left(1-\frac{\lambda_i}{\sigma^2+\lambda_i}\right).\end{aligned}$$

Thus we have shown that:

$$\mathcal{I}_0 = -\frac{1}{2}\sum_{i=1}^m h\left(\frac{\lambda_i}{\sigma^2+\lambda_i}\right),$$

where $h(x) = x + \log(1-x)$. □

Proof of Lemma 3: Following previous results, we define $\psi(w) = \mathbb{E}_0(\mathcal{T}_C)$ in (24) as a function of w and try to find the w^* that minimizes $\psi(w)$. Recall we have

$$\begin{aligned}\psi(w) &= \frac{2\log\gamma\big(1+o(1)\big)}{-2\delta_\infty \mathbb{E}_0[v_t^\top \widehat{U}_t\widehat{\Lambda}_t^{-1}\widehat{U}_t^\top v_t] + m\log(1+2\delta_\infty)} + w\\ &= \frac{2\log\gamma\big(1+o(1)\big)}{-2\delta_\infty D + m\log(1+2\delta_\infty)} + w.\end{aligned}$$

Substitute D into $\psi(w)$ gives us the expression for the EDD as:

$$\psi(w) = \frac{2\log\gamma}{D-m+m\log(m/D)} + w = \frac{2\log\gamma}{mg(D/m)} + w.$$

Recall we have denoted $\Delta = \frac{1}{m}\sum_{i=1}^m \frac{\lambda_i}{\sigma^2+\lambda_i}$, thus we have

$$\begin{aligned}D &= \sum_{i=1}^m \frac{\lambda_i}{\sigma^2+\lambda_i} - \sum_{i=1}^m \frac{\lambda_i B_i}{\sigma^2+\lambda_i}\frac{1}{w} + \sum_{i=1}^m \frac{3B_i^2\lambda_i}{\sigma^2+\lambda^2}\frac{1}{w^2}\\ &= m\Delta - \sum_{i=1}^m \frac{\lambda_i B_i}{\sigma^2+\lambda_i}\frac{1}{w} + \sum_{i=1}^m \frac{3B_i^2\lambda_i}{\sigma^2+\lambda^2}\frac{1}{w^2}.\end{aligned}$$

Note that $m\Delta \in (0,m)$ and it is close to close to 0 when σ^2 is large and close to m when σ^2 is small. Let

$$B_i' = \frac{\lambda_i B_i}{\sigma^2+\lambda_i}; \quad C_i' = \frac{3B_i^2\lambda_i}{\sigma^2+\lambda^2}.$$

We have

$$D = m\Delta - \frac{\sum B_i'}{w} + \frac{\sum C_i'}{w^2}.$$

Then

$$\psi(w) = w + \frac{2\log\gamma}{m(\Delta-1) - \frac{\sum B_i'}{w} + \frac{\sum C_i'}{w^2} - m\log(\Delta - \frac{\sum B_i'}{mw} + \frac{\sum C_i'}{mw^2})}.$$

Note that

$$\Delta - \frac{\sum B_i'}{mw} + \frac{\sum C_i'}{mw^2} = \Delta(1 - \frac{\sum B_i'}{\Delta mw} + \frac{\sum C_i'}{\Delta mw^2}).$$

Apply the Taylor expansion for $\log(1-x)$, we have

$$\log\left(1 - \frac{\sum B_i'}{\Delta mw} + \frac{\sum C_i'}{\Delta mw^2}\right) = -\frac{\sum B_i'}{\Delta mw} + \frac{\sum C_i'}{\Delta mw^2} + o(1/w).$$

Substitute into the EDD $\psi(w)$ we can obtain that:

$$\begin{aligned}\psi(w) = w + (2\log\gamma)\Big/\Big[&m(\Delta-1) - m\log(\Delta)\\ &- \frac{1}{w}(1-\frac{1}{\Delta})(\sum_i B_i') + \frac{1}{w^2}(1-\frac{1}{\Delta}))(\sum_i C_i')\Big],\end{aligned}$$

and note that the denominator will be dominated by the term $m(\Delta-1) - m\log(\Delta)$. Let the first-order derivative of $\psi(w)$ equal to 0 and ignore $o(1/w)$ terms, we obtain

$$\begin{aligned}\psi'(w) &= 1 + \frac{2(\log\gamma)(\frac{1}{\Delta}-1)(\sum B_i')}{[m(\Delta-1) - m\log(\Delta) - (\frac{1}{\Delta}-1)(\sum B_i')/w]^2 \cdot w^2}\\ &\approx 1 - \frac{2(\log\gamma)(\frac{1}{\Delta}-1)(\sum B_i')}{[m(\Delta-1) - m\log(\Delta)]^2 \cdot w^2}.\end{aligned}$$

The derivation above is approximated by ignoring $o(1/w)$ term in the denominator in the second step. It can be seen clearly that when $\psi'(w^*) = 0$ we get the minimum value for $\psi(w)$. Thus we have:

$$w^* = \frac{\sqrt{2(\log\gamma)(\frac{1}{\Delta}-1)(\sum B_i')}}{m(\Delta-1) - m\log(\Delta)}$$

$$= \frac{\sqrt{2(\log\gamma)(\frac{1}{\Delta}-1)(\sum \frac{\lambda_i B_i}{\sigma^2+\lambda_i})}}{m(\Delta-1-\log(\Delta))}$$
$$= \frac{\sqrt{2(\log\gamma)(\frac{1}{\Delta}-1)(\sum_{i=1}^m \frac{\lambda_i B_i}{\sigma^2+\lambda_i})}}{mg\,(\Delta)}.$$

Recall $g(x) = x - 1 - \log(x)$. Note that the denominator is always positive since $g(x) > 0$ for all $x \in (0,1)$ and $\Delta \in (0,1)$. □

Proof of Theorem 2: To explain the $o(1)$ terms in (15), note that the result in [38] states $\mathbb{E}_\infty[T_C] = e^b[1-(b+1)e^{-b}]/\mathcal{I}_\infty$, and $\mathbb{E}_0[T_C] = b[1+(e^{-b}-1)b^{-1}]/\mathcal{I}_0$, and thus the two $o(1)$ terms are on the order of be^{-b} and b^{-1} respectively, or equivalently, the two $o(1)$ terms are on the order of $(\log\gamma)/\gamma$ and $1/\log\gamma$ respectively. Now using the more precise version of the EDD for the exact CUSUM (16) and the Spectral CUSUM (24)

$$\mathbb{E}_0[T_C] = \frac{b(1+\mathcal{O}(b^{-1}))}{\mathcal{I}_0}. \quad \text{(exact CUSUM)}$$
$$\mathbb{E}_0(\mathcal{T}_C) = \frac{2b\big(1+\mathcal{O}(b^{-1})\big)}{mg(D/m)} + w. \quad \text{(Spectral CUSUM)}$$

Recall $g(x) = x - 1 - \log(x)$.

Since, to achieve ARL constraint γ, the threshold b for both procedures are on the order of $\log\gamma$, we obtain the ratio:

$$\frac{\mathbb{E}_0[\mathcal{T}_C]}{\mathbb{E}_0[T_C]} = \frac{2\mathcal{I}_0}{mg(D/m)} + \frac{w\mathcal{I}_0}{b(1+\mathcal{O}(b^{-1})} = \frac{2\mathcal{I}_0/m}{g(D/m)} + \frac{w\mathcal{I}_0}{b}(1+\mathcal{O}(b^{-1})),$$

using Taylor expansion $1/(1+x) = 1 - x + x^2/2\cdots$. Note that

$$\frac{2\mathcal{I}_0}{m} = -\frac{1}{m}\sum_{i=1}^m h\left(\frac{\lambda_i}{\sigma^2+\lambda_i}\right) = -\Delta - \frac{1}{m}\sum_{i=1}^m \log\left(\frac{\sigma^2}{\sigma^2+\lambda_i}\right),$$

where we recall that $h(x) = x + \log(1-x)$, $\Delta = \frac{1}{m}\sum_{i=1}^m \frac{\lambda_i}{\sigma^2+\lambda_i}$. Meanwhile we can write

$$D = \sum_{i=1}^m \frac{\lambda_i}{\lambda_i+\sigma^2}\left(1 - \frac{B_i}{w} + \frac{3B_i^2}{w^2}\right)$$
$$= m\Delta - \sum_{i=1}^m \frac{\lambda_i B_i}{\sigma^2+\lambda_i}\frac{1}{w} + \sum_{i=1}^m \frac{3B_i^2\lambda_i}{\sigma^2+\lambda^2}\frac{1}{w^2},$$

and thus

$$g(D/m) = g\left(\Delta - \sum_{i=1}^m \frac{\lambda_i B_i}{\sigma^2+\lambda_i}\frac{1}{mw} + \sum_{i=1}^m \frac{3B_i^2\lambda_i}{\sigma^2+\lambda^2}\frac{1}{mw^2}\right).$$

When w is sufficiently large (e.g., in our setting $w = \sqrt{\log\gamma}$, $\gamma \to \infty$), due to monotonicity and continuity of g, we have $g(D/m) \to g(\Delta) = \Delta - 1 - \log(\Delta)$ and

$$\frac{2\mathcal{I}_0/m}{g(D/m)} \to \frac{\Delta + \frac{1}{m}\sum_{i=1}^m \log\left(\frac{\sigma^2}{\sigma^2+\lambda_i}\right)}{1-\Delta+\log(\Delta)} = \mathcal{O}(1), \qquad (34)$$

(which does not depend on w or γ) since $g(\Delta) > 0$ and $\Delta \in (0,1)$ for σ^2 bounded away from 0. Combining with b is on the order of $\log\gamma$, w is in the order of $\sqrt{\log\gamma}$, we are done.

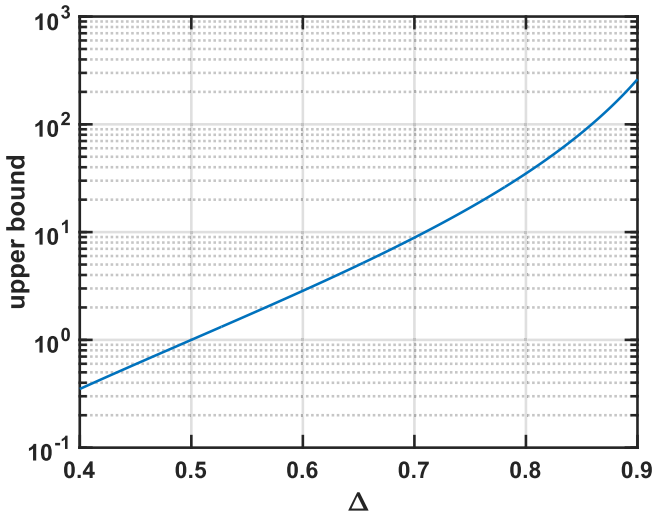


Fig. 13. Plot of the asymptotic upper bound $\frac{\Delta+\log(1-\Delta)}{1-\Delta+\log\Delta}$ as a function Δ, for $\Delta \in (0.4, 0.9)$. Note that it is reasonable for Δ bounded away from 1 (i.e., when σ^2 is bounded away from 0.)

We can further evaluate the limiting upper bound, using the concavity of the logarithm function, (34) can be further bounded by

$$\frac{\Delta + \frac{1}{m}\sum_{i=1}^m \log\left(\frac{\sigma^2}{\sigma^2+\lambda_i}\right)}{1-\Delta+\log(\Delta)} \le \frac{\Delta+\log(1-\Delta)}{1-\Delta+\log\Delta},$$

and the figure above plots the upper bound for $\Delta \in (0.4, 0.9)$. □

REFERENCES

[1] M. Faulkner, M. Olson, R. Chandy, J. Krause, K. M. Chandy, and A. Krause, "The next big one: Detecting earthquakes and other rare events from community-based sensors," in *Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2011, pp. 13–24.

[2] E. Ahmed, A. Clark, and G. Mohay, "A novel sliding window based change detection algorithm for asymmetric traffic," in *Proc. IFIP Int. Conf. Netw. Parallel Comput.*, Oct. 2008, pp. 168–175.

[3] M. Berger, L. M. Seversky, and D. S. Brown, "Classifying swarm behavior via compressive subspace learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 5328–5335.

[4] P. Basuchowdhuri, S. Sikdar, V. Nagarajan, K. Mishra, S. Gupta, and S. Majumder, "Fast detection of community structures using graph traversal in social networks," *Knowl. Inf. Syst.*, vol. 59, no. 1, pp. 1–31, Apr. 2019.

[5] H. Keshavarz, C. Scott, and X. Nguyen, "Optimal change point detection in Gaussian processes," *J. Stat. Planning Inference*, vol. 193, pp. 151–178, Feb. 2018.

[6] C. E. Johnson, A. Lindh, and B. Hirshorn, "Robust regional phase association," United States Geol. Surv., Reston, VA, USA, Tech. Rep. 94-621, 1997.

[7] D. Vere-Jones, "Stochastic models for earthquake occurrence," *J. Roy. Stat. Soc., B, Methodol.*, vol. 32, no. 1, pp. 1–45, Jan. 1970.

[8] T. Omi, Y. Ogata, Y. Hirata, and K. Aihara, "Forecasting large aftershocks within one day after the main shock," *Sci. Rep.*, vol. 3, no. 1, pp. 1–7, Jul. 2013.

[9] H.-H. Huang, F.-C. Lin, B. Schmandt, J. Farrell, R. B. Smith, and V. C. Tsai, "The Yellowstone magmatic system from the mantle plume to the upper crust," *Science*, vol. 348, no. 6236, pp. 773–776, May 2015.

[10] L. Peel and A. Clauset, "Detecting change points in the large-scale structure of evolving networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[11] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks," in *Proc. 9th WebKDD 1st SNA-KDD Workshop Web Mining Social Netw. Anal.*, Aug. 2007, pp. 16–25.

[12] A. Talwalkar, S. Kumar, and H. Rowley, "Large-scale manifold learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[13] E. Abbe, "Community detection and stochastic block models: Recent developments," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6446–6531, Jan. 2017.
[14] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Math. J.*, vol. 23, no. 2, pp. 298–305, 1973.
[15] A. Pothen, H. D. Simon, and K. P. Liou, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM J. Matrix Anal. Appl.*, vol. 11, no. 3, pp. 430–452, 1990.
[16] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.
[17] B. W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Syst. Tech. J.*, vol. 49, no. 2, pp. 291–307, 2014.
[18] C. Pizzuti, "GA-Net: A genetic algorithm for community detection in social networks," in *Proc. Int. Conf. Parallel Problem Solving Nature*. Cham, Switzerland: Springer, 2008, pp. 1081–1090.
[19] Y. Wang, A. Chakrabarti, D. Sivakoff, and S. Parthasarathy, "Fast change point detection on dynamic social networks," 2017, *arXiv:1705.07325*.
[20] D. Eswaran, C. Faloutsos, S. Guha, and N. Mishra, "SpotLight: Detecting anomalies in streaming graphs," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1378–1386.
[21] S. Huang, Y. Hitti, G. Rabusseau, and R. Rabbany, "Laplacian change point detection for dynamic graphs," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 349–358.
[22] I. U. Hewapathirana, D. Lee, E. Moltchanova, and J. McLeod, "Change detection in noisy dynamic networks: A spectral embedding approach," *Social Netw. Anal. Mining*, vol. 10, no. 1, pp. 1–22, Dec. 2020.
[23] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, Mar. 2007.
[24] H. Liu and L. Wang, "TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models," *Electron. J. Statist.*, vol. 11, no. 1, pp. 241–294, Jan. 2017.
[25] H. Keshavarz, G. Michailidis, and Y. Atchadé, "Sequential change-point detection in high-dimensional Gaussian graphical models," *J. Mach. Learn. Res.*, vol. 21, no. 82, pp. 1–47, 2020.
[26] Y. Xie, J. Huang, and R. Willett, "Change-point detection for high-dimensional time series with missing data," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 12–27, 2012.
[27] J. Yang, J. McAuley, and J. Leskovec, "Community detection in networks with node attributes," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 1151–1156.
[28] P.-L. Loh and M. J. Wainwright, "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses," *Ann. Statist.*, vol. 41, no. 6, pp. 3022–3049, Dec. 2013.
[29] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, nos. 1–2, pp. 100–115, 1954.
[30] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Stat.*, vol. 42, pp. 1897–1908, Dec. 1971.
[31] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Stat.*, vol. 14, no. 4, pp. 1379–1387, 1986.
[32] T. Oskiper and H. V. Poor, "Quickest detection of a random signal in background noise using a sensor array," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 1, Dec. 2005, Art. no. 360150.
[33] D. Egea-Roca, G. Seco-Granados, and J. A. López-Salcedo, "Comprehensive overview of quickest detection theory and its application to GNSS threat detection," *Gyroscopy Navigat.*, vol. 8, no. 1, pp. 1–14, Jan. 2017.
[34] L. Xie, G. V. Moustakides, and Y. Xie, "Window-limited CUSUM for sequential change detection," 2022, *arXiv:2206.06777*.
[35] L. Xie, G. V. Moustakides, and Y. Xie, "First-order optimal sequential subspace change-point detection," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Nov. 2018, pp. 111–115.
[36] L. Xie, Y. Xie, and G. V. Moustakides, "Sequential subspace change point detection," *Sequential Anal.*, vol. 39, no. 3, pp. 307–335, Jul. 2020.
[37] T. W. Anderson, "Asymptotic theory for principal component analysis," *Ann. Math. Statist.*, vol. 34, no. 1, pp. 122–148, 1963.
[38] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*. Springer, 1985.
[39] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2010, pp. 704–711.
[40] U. Von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Jul. 2007.
[41] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. 2, pp. 303–353, Apr. 1998.
[42] K. B. Petersen et al., "The matrix cookbook," *Tech. Univ. Denmark*, vol. 7, no. 15, p. 510, 2008.
[43] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Incremental singular value decomposition algorithms for highly scalable recommender systems," in *Proc. 5th Int. Conf. Comput. Inf. Sci.*, vol. 1, 2002, pp. 8–27.
[44] R. L. Bishop and R. J. Crittenden, *Geometry of Manifolds*. New York, NY, USA: Academic, 2011.
[45] T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2917–2929, Nov. 1998.
[46] H. Hotelling, "The generalization of student's ratio," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 54–65.
[47] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, p. 7, 2002.
[48] C. Moolenbeek and E. J. Ruitenberg, "The 'Swiss roll': A simple technique for histological studies of the rodent intestine," *Lab. Animals*, vol. 15, no. 1, pp. 57–60, Jan. 1981.
[49] X. He, Y. Xie, S.-M. Wu, and F.-C. Lin, "Sequential graph scanning statistic for change-point detection," in *Proc. 52nd Asilomar Conf. Signals, Syst., Comput.*, Oct. 2018, pp. 1317–1321.
[50] S. Wu, K. M. Ward, J. Farrell, F. Lin, M. Karplus, and R. B. Smith, "Anatomy of old faithful from subsurface seismic imaging of the Yellowstone upper geyser basin," *Geophys. Res. Lett.*, vol. 44, no. 20, pp. 10–240, Oct. 2017.

Minghe Zhang received the B.S. degree in electronic engineering from Tsinghua University in 2017 and the M.S. degree in electrical and computer engineering in 2019. He is currently pursuing the Ph.D. degree in machine learning with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology. His research interests include statistical machine learning and network change detection.

Liyan Xie (Student Member, IEEE) received the B.Sc. degree in statistics from the University of Science and Technology of China in 2016 and the Ph.D. degree in industrial engineering from the Georgia Institute of Technology in 2021. She is currently an Assistant Professor with the School of Data Science, The Chinese University of Hong Kong, Shenzhen. Her research interests include intersection between statistics and optimization, with a primary focus on sequential change detection and its applications in sensor networks and healthcare. She was a recipient of the 2020 ARC Student Fellowship and the IDEaS-TRIAD Research Scholarships from the Transdisciplinary Research Institute for Advancing Data Science (TRIAD) at Georgia Tech. She was awarded the finalist of the INFORMS QSR Student Paper Competition and the runner-up for the INFORMS Computing Society Student Paper Prize in 2019.

Yao Xie (Member, IEEE) received the Ph.D. degree in electrical engineering (mathematics) from Stanford University. She was a Research Scientist with Duke University. She is currently an Associate Professor and the Harold R. and Mary Anne Nash Early Career Professor with the H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, where she is also the Associate Director of the Machine Learning Center. Her research interests include the intersection of statistics, machine learning, and optimization in providing theoretical guarantees and developing computationally efficient and statistically powerful methods for problems motivated by real-world applications. She received the National Science Foundation (NSF) CAREER Award in 2017, the INFORMS Wagner Prize Finalist in 2021, and the INFORMS Gaver Early Career Award for Excellence in Operations Research in 2022. She is currently an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON SIGNAL PROCESSING, *Journal of the American Statistical Association, Theory and Methods*, *Sequential Analysis: Design Methods and Applications*, and *INFORMS Journal on Data Science*. She is the Area Chair of NeurIPS in 2021, 2022, 2023, and ICML in 2023.