

Optimal Change Point Detection and Localization in Sparse Dynamic Networks

Daren Wang¹, Yi Yu², and Alessandro Rinaldo³

¹Department of Statistics, University of Chicago

²School of Mathematics, University of Bristol

³Department of Statistics and Data Science, Carnegie Mellon University

September 26, 2018

Abstract

We study the problem of change point detection and localization in dynamic networks. We assume that we observe a sequence of independent adjacency matrices of given size, each corresponding to one realization from an unknown inhomogeneous Bernoulli model. The underlying distribution of the adjacency matrices may change over a subset of the time points, called change points. Our task is to recover with high accuracy the unknown number and positions of the change points. Our generic model setting allows for all the model parameters to change with the total number of time points, including the network size, the minimal spacing between consecutive change points, the magnitude of the smallest change and the degree of sparsity of the networks. We first identify an impossible region in the space of the model parameters such that no change point estimator is provably consistent if the data are generated according to parameters falling in that region. We propose a computationally simple novel algorithm for network change point localization, called Network Binary Segmentation, which relies on weighted averages of the adjacency matrices. We show that Network Binary Segmentation is consistent over a range of the model parameters that nearly cover the complement of the impossibility region, thus demonstrating the existence of a phase transition for the problem at hand. Next, we devise a more sophisticated algorithm based on singular value thresholding, called Local Refinement, that delivers more accurate estimates of the change point locations. We show that, under appropriate conditions, Local Refinement guarantees a minimax optimal rate for network change point localization while remaining computationally feasible.

Keywords: Change point detection; Low-rank networks; Stochastic block model; Minimax optimality.

1 Introduction

The analysis of network is a fundamental task in statistics due to the increasing popularity of network data generated from various scientific areas, the social sciences, emerging industries, as well as everyday life. Over the last decade, most of the advances in the area of statistical network analysis have revolved around *static network models*, where the properties of the data generating process are inferred from a single realization of the network. For this type of problems, a large body of results of computational, methodological and theoretical nature exist.

In contrast to basic premise of the static network modeling framework, many network data sets consist instead of multiple network realizations indexed by time, so that both the number of nodes and the connectivity structure of the network exhibit time-varying features. Such a *dynamic network modeling* setting is naturally more complex and challenging, as it is necessary to additionally formalize and model the underlying temporal dynamic. While there is a vast body of work on dynamic network models (see, e.g., [Barabási and Albert, 1999](#)) in the broader scientific literature, theoretical results on such models are comparatively scarce in the statistical literature, with many of the contributions being fairly recent (see [Section 1.1](#) below for some literature review).

In this article we study a simple problem for network dynamics in discrete time, in which the set of nodes is fixed but the edge probabilities are time-varying. Specifically, we assume that we observe a sequence of T independent and possibly sparse networks of constant size whose distributions may change at $K < T$ fixed but unknown time points, or change points. We impose minimal restrictions on the number and locations of the change points and especially on the nature of the distributional changes that may occur at those times. In particular, most popular static network models can fit into our framework. The task we set out to solve is to detect whether any such change has taken place, and to accurately estimate the time or location of the corresponding change point. Importantly, we are not interested in estimating the underlying distributions, but only the location of their changes. As our analysis will reveal, although we only consider a fairly straightforward form of network dynamic, the associated inference problem is rather subtle and far from trivial.

To set up the problem, we assume a sequence of T independent adjacency matrices of size n , each from a possibly sparse inhomogeneous Bernoulli network model, defined next.

Definition 1 (Inhomogeneous Bernoulli networks). *A network with node set $\{1, \dots, n\}$ is an inhomogeneous Bernoulli network if its adjacency matrix $A \in \mathbb{R}^{n \times n}$ satisfies*

$$A_{ij} = A_{ji} = \begin{cases} 1, & \text{nodes } i \text{ and } j \text{ are connected by an edge,} \\ 0, & \text{otherwise;} \end{cases}$$

and $\{A_{ij}, i < j\}$ are independent Bernoulli random variables with $\mathbb{E}(A_{ij}) = \Theta_{ij}$.

In addition, let $\rho = \|\Theta\|_\infty \leq 1$, where $\|\cdot\|_\infty$ denotes the entrywise maximum norm of a matrix. If $\rho = o(1)$, then we call the corresponding network a sparse network.

Remark 1. *Definition 1 covers a wide range of models for undirected networks, including Erdős–Rényi random graph ([Erdős and Rényi, 1959](#)), stochastic block model ([Holland et al., 1983](#)), degree corrected block model ([Karrer and Newman, 2011](#)) and random dot product model ([Young and Scheinerman, 2007](#)), etc. It is worth pointing out that although we are only considering undirected networks, our results extend straightforwardly to directed networks, i.e. asymmetric adjacency matrices. Additionally, for technical convenience, we are allowing self-loops, even though networks with no loops can be easily accommodated; see [Section 3.2](#) below. Finally, we take notice that the inhomogeneous Bernoulli network is specified by up to $\frac{n(n+1)}{2}$ parameters, which correspond to the number of observable edges.*

We further assume that the probability distributions of the networks change only over an unknown subset of the time points, called change points. At this point, we impose no restrictions on the type and magnitude of the distributional changes occurring at the change points, the minimal spacing among two consecutive change points and the sparsity of the networks. We formalize our setting below.

Assumption 1 (Model setting). Let $\{A(1), \dots, A(T)\} \subset \mathbb{R}^{n \times n}$ be a collection of adjacency matrices of independent inhomogeneous Bernoulli networks with means $\Theta(1), \dots, \Theta(T)$ satisfying the following properties.

1. The sparsity parameter $\rho = \max_{t=1, \dots, T} \|\Theta(t)\|_\infty$ is such that

$$\rho n \geq \log(n). \quad (1)$$

2. There exists a sequence $\eta_0 < \eta_1 < \dots < \eta_{K+1}$ of time points, called change points, with $\eta_0 = 0$ and $\eta_{K+1} = T$, such that

$$\Theta(\eta_k + 1) = \Theta(\eta_k + 2) = \dots = \Theta(\eta_{k+1}), \text{ for any } k = 0, \dots, K.$$

3. The minimal spacing between two consecutive change points satisfies

$$\min_{k=1, \dots, K+1} \{\eta_k - \eta_{k-1}\} \geq \Delta > 0.$$

Notice that, necessarily, $\Delta \leq T$.

4. The magnitudes of the changes in the data generating distribution are such such that

$$\|\Theta(\eta_k) - \Theta(\eta_{k-1})\|_F = \kappa_k > 0, \text{ for any } k = 1, \dots, K + 1,$$

and

$$\min_{k=1, \dots, K+1} \kappa_k = n\rho\kappa_0 > 0,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Remark 2. In (1) we require $\rho n \geq \log(n)$. In Appendix B, we use $p = n^2$ and assume $\rho\sqrt{p} \geq \log(p)$. We admit the discrepancy regarding a constant due to the notational convenience.

Remark 3 (Definition of change points). It can be seen from Assumption 1 part 2 that the changes occur at time points $\{\eta_k + 1\}_{k=0}^K$, but we refer to $\{\eta_k\}_{k=0}^K$ as change points in this paper for convenience, considering the fact that we define $\eta_0 = 0$ while the observations start at $\eta_0 + 1$.

Remark 4 (The normalized minimal magnitude of signal jump). In Assumption 1 part 1, we define the entrywise sparsity parameter ρ , which is upper bounded by 1. When allowing $\rho \rightarrow 0$, as $n \rightarrow \infty$, it is convenient to normalize the change size, which is defined in terms of Frobenius norm, both by the sparsity parameter and the size of the network. We therefore define $\kappa_0 = (n\rho)^{-1} \min_{k=1, \dots, K+1} \kappa_k$. With this normalization, κ_0 is a scale-free parameter in $(0, 1]$.

The model described above is defined by the parameters Δ , κ_0 , n and ρ , which are allowed to change as T grows unbounded¹. We refer to any relationship among all the model parameters $(\Delta, \kappa_0, n, \rho)$ and T that holds as $T \rightarrow \infty$ as a **scaling**.

¹The number of change points K also may change with T , but since the minimal spacing parameter Δ is related to K by the inequality

$$K \leq \frac{T}{\Delta},$$

we will capture such dependence only through Δ .

We are concerned with the problem of estimating the unknown number and unknown locations of the change points based on one observation of a sequence $\{A(1), \dots, A(T)\}$ of adjacency matrices satisfying the above assumptions. More precisely, for a given scaling of the model parameters, we seek to produce estimators of (η_1, \dots, η_K) of the form

$$(A(1), \dots, A(T)) \mapsto (\hat{\eta}_1, \dots, \hat{\eta}_{\hat{K}})$$

such that, with probability tending to 1 as $T \rightarrow \infty$,

$$\hat{K} = K \quad \text{and} \quad \max_{k=1, \dots, \hat{K}} |\hat{\eta}_k - \eta_k| \leq \epsilon, \quad (2)$$

where $\epsilon = \epsilon(T, \Delta, \kappa_0, \rho, n)$. It is important to emphasize that the limiting probability (in T) of the event in (2) – and the value of the limit itself, if it exists – depends on the choice of the scaling. For instance, it is intuitively clear that a scaling in which all parameters decrease with T will lead to a sequence of change point problems of increasing difficulty.

We will refer to the ratio $\frac{\epsilon}{T}$ as the **localization rate** of the estimator and we will deem such estimator **consistent** if, as $T \rightarrow \infty$, its localization rate vanishes, i.e. if

$$\lim_{T \rightarrow \infty} \frac{\epsilon}{T} = 0.$$

Our main goal is to derive conditions on the scaling of the model parameters that guarantee the feasibility of consistent estimation of the change points and to derive computationally-feasible estimators that are consistent and in fact optimal, in the sense of achieving the minimax localization rate. Throughout, we will specify any scaling regime among the parameters by expressing them as functions of quantity

$$\sqrt{\rho} \kappa_0, \quad (3)$$

which can be considered as a uniform lower bound on the signal-to-noise ratio for any network change point model satisfying Assumption 1. Indeed, the above quantity is the minimal magnitude of the signal jump, namely $\kappa_0 n \rho$, divided by $n \sqrt{\rho}$, which is an upper bound on the total variance of the entries of A .

Our contributions are as follows.

- We first demonstrate the existence of a phase transition for the problem at hand by giving nearly matching necessary and sufficient conditions on the scaling of the model parameters and T for consistent estimation of the change points. Specifically, under the low signal-to-noise scaling

$$\rho \kappa_0^2 \lesssim \frac{1}{n \Delta}, \quad (4)$$

no algorithm is guaranteed to be consistent (in the minimax sense: there exists a change point problem setting compatible with the above assumption such that any algorithm will have a localization rate uniformly bounded away from 0). On the other hand, if for any $\xi > 0$ ²,

$$\rho \kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta} \frac{1}{n}, \quad (5)$$

²Note that ξ is allowed to be zero, if one assumes n diverges with T or $\Delta = o(T)$.

we demonstrate a computationally-efficient procedure, called Network Binary Segmentation (NBS) (see Algorithm 1 below) that is provably consistent. The procedure combines sample splitting, which has been proved effective by Wang and Samworth (2016) and by Wang et al. (2017) in high-dimensional change point problems, with the randomized search strategy implemented in the wild binary segmentation (WBS) algorithm of Fryzlewicz (2014). To show the consistency of the NBS we have generalized in non-trivial ways the analysis in Venkatraman (1992) to allow for vector- and matrix-valued CUSUM statistics; we believe that such generalization may be applied to other change point detection problems and, therefore, may be of independent interest.

NBS is consistent under nearly the weakest possible conditions, since it leads to a vanishing localization rate under the scaling (5) which, save for a $\log^{2+2\xi}(T)$ term, matches the phase transition boundary in (4). Remarkably, no structural assumptions on the distributions of the networks themselves are used. Indeed, in deriving the bound (4), we construct a worst-case class of distributions consisting of dynamic networks satisfying stochastic block models. This reveals that, under the scaling in which NBS is analyzed, imposing extra network structural assumptions do not necessarily lead to easier change point detection problems. This is in stark contrast with many other network problems, such as graphon estimation, clustering and testing, where some structural conditions on the edge probabilities are always necessary. For instance, Gao et al. (2015) showed that, when the number of communities r in a stochastic block model is of order n , the minimax lower bound under the normalized mean squared error loss for graphon estimators is of order 1.

- In our second set of results, we seek to investigate conditions under which structural assumptions do help with our change-point localization task. Towards that end, we introduce additional assumptions on the model defined in Assumption 1 by requiring that each difference $\Theta(\eta_k) - \Theta(\eta_{k-1})$, $k = 1, \dots, K + 1$ be a matrix of rank at most r , which is also allowed to change with T . Such low rank condition is relative mild and is satisfied by many instances of the stochastic block model. Then, with this assumption in place and under the slightly stronger parameter scaling

$$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T) r}{\Delta n}, \quad (6)$$

we are able to devise a computationally-efficient and consistent change points estimator with localization error of the order

$$\epsilon \lesssim \frac{\log^2(T)}{\kappa_0^2 n^2 \rho} \quad (7)$$

The proposed procedure takes as input the estimates of the change point locations from any reasonable (not necessarily consistent nor optimal) estimator, including NBS, and further improves their accuracy to deliver the above localization rate. At its core, the LR algorithm relies on exactly K (this, we recall, being number of change points) separate applications of the universal singular value thresholding procedure of Chatterjee (2015). Furthermore, we show that the localization rate afforded by the LR algorithm, given in (7), is in fact minimax rate-optimal, save for the $\log^{2+2\xi}(T)$ factor. Interestingly, the expression of the rate (7) is essentially identical to the optimal localization rate for covariance and mean change point estimation, adjusted for the differences in the model settings (e.g. Wang et al., 2017).

- We apply the LR algorithm to the problem of change point detection for sequence of networks from stochastic block models and derive optimal localization rates. For networks without self-loops – a commonly feature of network models – a technical complication arises in treating the expected adjacency matrix from a stochastic model as a low-rank matrix. When the network has no self-loops, then the diagonal entries of the expected adjacency matrix are set to be zero, which in general would prevent the low-rank assumption. In this case, we show that with very mild additional assumptions, we are still able to recover the nearly optimal localization rate (7) in this problem as well as borrow tools and ideas from several areas, including change point detection, network analysis and graphon estimation.

It is important to mention that our analysis of the NBS and LR algorithms uses different techniques and tools than the ones previously used by Wang and Samworth (2016) and Wang et al. (2017) in change-point localization problems for matrices and vectors, many of which have been newly developed or extended from the recent literature on statistical network analysis. Similarly, the NBS algorithm builds upon the WBS procedure developed and studied by Fryzlewicz (2014) for univariate mean change point estimation. However, the nature and settings of the network change point estimation problem tackled here are significantly different, and in fact more complex; and, consequently, so is our analysis.

We also want to point out that, under the model assumption of being stochastic block models and the scaling (6), other graphon-based algorithms (e.g. Gao et al., 2015) will also lead to optimal rates. In fact, these algorithms can be shown to produce the optimal rate (7) under the scaling

$$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T) (1 + r^2/n)}{\Delta n}, \quad (8)$$

which is slightly weaker than the one we consider, i.e. (6). However, such algorithms are NP-hard or computationally-unfeasible and, for this reason, we abstain from pursuing such analysis. Note that if in addition to (8), one further assumes $r \lesssim \sqrt{n}$, which is commonly assumed in the community-structured network models, then (8) is essentially the same as (6).

The rest of this paper is organized as follows: related literature and general notation needed in this paper are collected in Sections 1.1 and 1.2 respectively; in Section 2, we propose a consistent change point detection method for inhomogeneous Bernoulli networks under conditions which will be proved to be minimax optimal; for more realistic but constrained models, e.g. stochastic block models, in Section 3 we will propose network change point detection methods which are not only consistent, but also optimal in terms of localization rates; detailed proofs will be deferred to the Appendices.

1.1 Related work

Dynamic network is a topical area which is intensely studied across different disciplines. The relevant papers listed in this section are by no means exhaustive. Readers may refer to Carrington et al. (2005); Goldenberg et al. (2010); Boccaletti et al. (2014); Kolaczyk (2017) for more comprehensive reviews.

In terms of the invariant quantities, most of the existing work focus on a fixed set of nodes across time, but there are also exceptions. For instance, Barabási and Albert (1999) allowed for time-varying nodes and edges, Crane (2015) assumed infinite population at every time point and allowed for random observations at different time points, to name but a few. In terms of the network

models imposed for every time point, [Snijders \(2002\)](#) explored dynamic exponential random graph models, [Tang et al. \(2013\)](#) studied dynamic version of random dot product models, [Ho et al. \(2011\)](#) extended the mixed membership models to a dynamic one, [Xu and Zheng \(2009\)](#); [Sewell and Chen \(2015\)](#) among others considered dynamic latent space models, and dynamic stochastic block models have also been extensively studied.

Among the work on dynamic stochastic block models, [Xu \(2015\)](#) proposed a stochastic block transition model using a hidden Markov-type approach; [Xu and Hero \(2014\)](#) proposed to track dynamic stochastic block models using Gaussian approximation and an extended Kalman filter algorithm; [Matias and Miele \(2017\)](#) integrated a Markov chain determined group labels evolving process; [Pensky and Zhang \(2017\)](#) exploited kernel-based smoothing techniques dealing with the evolving block structures; [Bhattacharyya and Chatterjee \(2017\)](#) focused on time-varying stochastic block model and variants thereof with time-independent community labels, applied spectral clustering on an averaged version of adjacency matrices, and achieved consistent community detection. [Wang et al. \(2014\)](#) used two types of scan statistics investigating change point detection on time-varying stochastic block model sequences, emphasizing testing connectivity matrices changes. [Cribben and Yu \(2017\)](#) proposed an eigen-space based statistics testing the community structures changes in stochastic block model sequences. [Liu et al. \(2018\)](#) proposed a loss function based on the eigen-space to track the changes of the community structures in stochastic block model sequences. Both [Cribben and Yu \(2017\)](#) and [Liu et al. \(2018\)](#) have roots in subspace tracking in signal processing, but both lack theoretical justifications. [Chu and Chen \(2017\)](#) proposed a test statistics for general data type including network sequences, and their method focuses on the testing perspective.

Another related area is the general change point detection area, which is by no means new but which has been going through a renaissance due to a massive upsurge of complex data sets and advanced statistical methods in the last decade or so. Multiple change point detection problems in univariate time series have been revisited by many researchers from different angles (e.g. [Harchaoui and Lévy-Leduc, 2010](#); [Frick et al., 2014](#); [Fryzlewicz, 2014](#); [Lin et al., 2017](#)). High-dimensional time series mean change point detection problems are the obvious but by no means straightforward extension of the univariate case. Papers on this topic include [Horváth and Hušková \(2012\)](#), [Cho \(2015\)](#), [Wang and Samworth \(2016\)](#), among many others. Besides mean change point detection, efforts have also been made in covariance change point detection problems (e.g. [Aue et al., 2009](#); [Barigozzi et al., 2016](#); [Wang et al., 2017](#)).

1.2 Notation

For any $A \in \mathbb{R}^{n \times n}$, let A_{ij} be the (i, j) th entry of A , A_{i*} and A_{*j} the i th row and j th column of A . Let $\kappa_i(A)$ be the i th eigenvalue of A with ordering $|\kappa_1(A)| \geq |\kappa_2(A)| \geq \dots \geq |\kappa_n(A)|$, and $\|A\|_{\text{op}} = |\kappa_1(A)|$ be the operator norm of A . Let $\|A\|_{\infty} = \max_{1 \leq i, j \leq n} |A_{ij}|$ be the entrywise maximum norm. In addition, for any $B \in \mathbb{R}^{n \times n}$, let $(A, B) = \sum_{1 \leq i, j \leq n} A_{ij} B_{ij}$ be the inner product of A and B in the matrix space, and $\|A\|_{\text{F}} = \sqrt{(A, A)}$ be the Frobenius norm of A . For any vector $v \in \mathbb{R}^p$, let v_i be the i th entry of v , $\|v\|$ and $\|v\|_{\infty}$ be the ℓ_2 - and entrywise maximum norms of v , respectively. For any set S , let S^c be its complement.

For any positive functions of n , namely $f(n)$ and $g(n)$, denote $f(n) \lesssim g(n)$, if there exist constants $C > 0$ and n_0 such that $f(n) \leq Cg(n)$ for any $n \geq n_0$; denote $f(n) \gtrsim g(n)$, if $g(n) \lesssim f(n)$; and denote $f(n) \asymp g(n)$, if $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$.

Definition 2 (CUSUM statistics). For a collection of any type of data $\{X(t)\}_{t=1}^T$, any pair of time points $(s, e) \subset \{0, \dots, T\}$ with $s < e - 1$, and any time point $t = s + 1, \dots, e - 1$, let the CUSUM statistics be

$$\tilde{X}^{s,e}(t) = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^t X_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^e X_i.$$

Definition 2 can be applied to sequences of scalars, vectors or adjacency matrices. As an example, for a collection of adjacency matrices $\{A(t)\}_{t=1}^T$, time points $1 \leq s < t < e \leq T$, we have

$$\mathbb{E}(\tilde{A}^{s,e}(t)) = \tilde{\Theta}^{s,e}(t).$$

2 Consistent Estimation: Phase Transition

Below we characterize the conditions under which consistent estimation of the change point locations for the model described in Assumption 1 is feasible. Specifically, we derive a phase transition in the space of the model parameters that separates parameter scalings for which there exists some algorithm with vanishing localization rate is possible from the ones for which no estimator is consistent.

We begin with an information-theoretic lower bound, which demonstrates that, if

$$\rho\kappa_0^2 \lesssim \frac{1}{n\Delta}, \quad (9)$$

then no consistent estimator of the change points can exist. The proof constructs two sequences of mixtures of stochastic block models with two communities of all possible sizes that cannot be reliably discriminated under the above scaling, and then employs the convex version of Le Cam's Lemma (see, e.g. Yu, 1997) to conclude that any change point estimator must have a localization rate bounded away from zero. As a by-product of our lower bound construction, we also see that imposing additional structural assumptions on the edge probabilities (such as that of a stochastic block model with a bounded number of communities) does not necessarily lead to a consistent estimator under the scaling in (9). The details of the construction can be found in Appendix A.

Lemma 1. Let $\{A(t)\}_{t=1}^T$ be a sequence of independent inhomogeneous Bernoulli networks satisfying Assumption 1 with $K = 1$ (i.e. there exists one and only one change point). Let $P_{\kappa_0, \Delta, n, \rho}^T$ denote the corresponding joint distribution. Consider the class of distributions

$$\mathcal{P} = \left\{ P_{\kappa_0, \Delta, n, \rho}^T : \rho\kappa_0^2 \leq \frac{1}{33n\Delta}, \Delta \leq T/3, \rho \leq 1/2, \kappa_0 \leq 1 \right\},$$

and, for each $P \in \mathcal{P}$, let $\eta(P) \in \{1, \dots, T\}$ denote the location of the corresponding change point.

For each $\zeta \leq 1/33$, denote

$$\mathcal{P}_\zeta = \left\{ P_{\kappa_0, \Delta, n, \rho}^T : \Delta = \min \left\{ \left\lfloor \frac{\zeta}{n\rho\kappa_0^2} \right\rfloor, \lfloor T/3 \rfloor \right\}, \rho \leq 1/2, \kappa_0 \leq 1 \right\}.$$

It holds that

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}_\zeta} \mathbb{E}_P(|\hat{\eta} - \eta(P)|) \geq T/4,$$

where the infimum is over all the possible estimators of the change point location.

Remark 5. *The lower bound construction used in the proof of Lemma 1 only considers change point models for networks for which $\Delta = \frac{1}{33n\rho\kappa_0^2}$.*

In our next result, we show that parameter scalings of the form given in (9) are essentially the only ones for which consistent change point estimation is infeasible, thus proving the existence of a phase transition in the space of parameters. In particular, we will derive an algorithm (see Algorithm 1 below) that will return a consistent estimator provided that for any $\xi > 0$,

$$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{n\Delta}. \quad (10)$$

Note that the above bound does not quite match the scaling (9) because of the extra $\log^{2+2\xi}(T)$ term. But aside from this logarithmic gap, our results cover all the possible scalings.

We will formally state the condition (10) as an assumption.

Assumption 2. *There exists a sufficiently large $C_\alpha > 0$ such that for any $\xi > 0$,*

$$\kappa_0\sqrt{\rho} \geq C_\alpha\sqrt{\frac{1}{n\Delta}}\log^{1+\xi}(T).$$

Remark 6. *Note that if we further assume $n \rightarrow \infty$ as $T \rightarrow \infty$, or assume $\Delta = o(T)$, then we only need to assume*

$$\kappa_0\sqrt{\rho} \geq C_\alpha\sqrt{\frac{1}{n\Delta}}\log(T),$$

i.e. $\xi = 0$. In the rest of this paper, we will stick to the case where $\xi > 0$.

To appreciate how Assumption 2 is compatible with a broad range of change point scenarios for networks and is therefore fairly mild, we highlight the following two extreme cases.

- Assume a non-sparse setting (i.e. $\rho = 1$). If the minimal spacing $\Delta \asymp \log^{2+2\xi}(T)$, which, at least in the univariate mean change point case (see, e.g., Yao and Au, 1989), corresponding to the minimax optimal spacing up to an extra $\log^{2+2\xi}(T)$ term, then it is required that $n\kappa_0 \asymp n^{1/2}$, and therefore polynomially increasing in n . This means that the edge probabilities need to change for at least \sqrt{n} order many nodes.
- On the other hand, in the sparse setting where ρ is allowed to vanish with n as in Equation (1), if $\Delta \asymp T$ (so that the number of change points is bounded), then Assumption 2 only requires κ_0 to be at least of the order

$$\frac{\log^{1+\xi}(T)\log^{-1/2}(n)}{\sqrt{T}}.$$

Thus κ_0 is allowed to vanish with T , even for fixed n .

We now introduce the procedure Network Binary Segmentation (NBS), detailed below in Algorithm 1, for consistent estimation under nearly the worst possible scaling of Assumption 2.

NBS is a novel algorithm that builds on the traditional machinery developed for the univariate mean change point detection problem. The cornerstones of NBS are the CUSUM statistics $\tilde{A}^{sm,em}(t)$ and $\tilde{B}^{sm,em}(t)$ (see Definition 1). However, instead of searching for the maximum CUSUM statistics directly, as it is traditionally done in the binary segmentation and its more modern variants (see, e.g.

Algorithm 1 Network Binary Segmentation. $\text{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau_1, \delta)$

INPUT: Two independent samples $\{A(t)\}_{t=1}^T, \{B(t)\}_{t=1}^T \in \mathbb{R}^{n \times n}, \tau_1, 1/2 > \delta > 0$.

```

for  $m = 1, \dots, M$  do
   $[s'_m, e'_m] \leftarrow [s, e] \cap [\alpha_m, \beta_m]$ 
   $(s_m, e_m) \leftarrow [s'_m + \delta(e'_m - s'_m), e'_m - \delta(e'_m - s'_m)]$ 
  if  $e_m - s_m \geq 1$  then
     $b_m \leftarrow \arg \max_{t=s_m+1, \dots, e_m-1} (\tilde{A}^{s_m, e_m}(t), \tilde{B}^{s_m, e_m}(t))$ 
     $a_m \leftarrow (\tilde{A}^{s_m, e_m}(b_m), \tilde{B}^{s_m, e_m}(b_m))$ 
  else
     $a_m \leftarrow -1$ 
  end if
end for
 $m^* \leftarrow \arg \max_{m=1, \dots, M} a_m$ 
if  $a_{m^*} > \tau_1$  then
  add  $b_{m^*}$  to the set of estimated change points
   $\text{NBS}((s, b_{m^*}), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau_1, \delta)$ 
   $\text{NBS}((b_{m^*} + 1, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau_1, \delta)$ 
end if

```

OUTPUT: The set of estimated change points.

Vostrikova, 1981; Fryzlewicz, 2014; Wang and Samworth, 2016), NBS maximizes the inner product of two CUSUM statistics based on two independent samples. This is due to the fact that each entry of the adjacency matrix is a Bernoulli random variable, and for any Bernoulli random variable X , it holds that $X^2 = X$. This results in that $\|\tilde{A}^{s_m, e_m}(t)\|_{\mathbb{F}}^2$ cannot serve as a good estimator of $\|\tilde{\Theta}^{s_m, e_m}(t)\|_{\mathbb{F}}^2$.

In addition, we introduce a tuning parameter δ to avoid false positives, which is similar to δ used in Algorithm 3 in Wang et al. (2017), and also similar to β used in Wang and Samworth (2016).

An interesting and possibly surprising feature of the NBS algorithm is that it merely relies on network CUSUM statistics – weighted sample averages of adjacency matrices (see Definition 2) – and does not deploy any network or graphon estimation procedures, which are computationally-costly. Though the NBS is not attempting to estimate any network parameters at all, it is still able achieve consistent network change point detection for a fairly large class of models in a fast fashion, as the next result shows.

Theorem 1. *Let Assumptions 1 and 2 hold and let $\{(\alpha_m, \beta_m)\}_{m=1}^M \subset (0, T)$ be a collection of intervals whose end points are drawn independently and uniformly from $\{1, \dots, T\}$ and such that $\max_{m=1, \dots, M} (\beta_m - \alpha_m) \leq C_R \Delta$ for an absolute constant $C_R > 0$. Set*

$$\epsilon = C_1 \log(T) \left(\frac{\sqrt{\Delta}}{\kappa_0 n \rho} + \frac{\log^{1/2}(T)}{\kappa_0^2 n \rho} \right), \quad (11)$$

for an absolute constant $C_1 > 0$.

Suppose there exists sufficiently small $0 < c_2 < 1$ such that the input parameters τ and δ of Algorithm 1 satisfy

$$C_\beta \rho n \log^{3/2}(T) < \tau < c_2 \kappa_0^2 n^2 \rho^2 \Delta, \quad \delta = \frac{1}{32} \min \left\{ \frac{1}{2}, \frac{C_1}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{C_1}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right\}. \quad (12)$$

Then the collection of the estimated change points $\mathcal{B} = \{\hat{\eta}_k\}_{k=1}^{\hat{K}}$ returned by NBS with input parameters $(0, T)$, $\{(\alpha_m, \beta_m)\}_{m=1}^M$, τ and δ is such that

$$\mathbb{P} \left\{ \hat{K} = K; \max_{k=1, \dots, K} |\eta_k - \hat{\eta}_k| \leq \epsilon \right\} \geq 1 - \exp \left(\log \frac{T}{\Delta} - M \frac{\Delta^2}{16T^2} \right) - (6T^{3-c_T} + 2T^{3-c}) \quad (13)$$

for some absolute constants $c, c_T > 3$ and any $n, T \geq 2$.

Remark 7 (Input parameters τ and δ). *The lower and upper bounds of τ detailed in (12) are the upper and lower bounds of a_{m^*} in the cases where there is no change point and there exists at least one change point in the induction step, respectively. These are derived in **Steps 1** and **2**. The parameter δ is used in **Steps 1** and **3** generating working interval $[s_m, e_m]$. The upper bound of δ is properly chosen based on the magnitude of ϵ , see (25).*

Remark 8 (Tracking the constants in Theorem 1). *It can be seen from the proof that we have the following hierarchy of the constants involved in Theorem 1. Firstly, c and c_T are chosen such that (13) tends to 0 as $T \rightarrow \infty$. Consequently, C_β is chosen according to Lemma 7. The constant c_2 therefore depends on C_β and C_α . A sufficiently large C_α can guarantee that $c_2 > 0$ exists. Finally, a sufficiently large $C_1 > 0$ is chosen and depends on all the aforementioned constants and C_R .*

To see how Theorem 1 implies that NBS is consistent, we plug in the inequalities

$$\sqrt{\rho} \kappa_0 \geq \frac{C_\alpha \log^{1+\xi}(T)}{\sqrt{n} \Delta} \quad \text{and} \quad \rho \geq \frac{\log(n)}{n},$$

stemming from Assumption 2 and Assumption 1 respectively, to bound the localization rate of NBS as follows:

$$\begin{aligned} \frac{\epsilon}{T} &= C_1 \log(T) \left(\frac{\sqrt{\Delta}}{\kappa_0 n \rho} + \frac{\log^{1/2}(T)}{\kappa_0^2 n \rho} \right) \frac{1}{T} \\ &\leq C_1 \left(\frac{\Delta}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right) \frac{1}{T}. \end{aligned} \quad (14)$$

As $T \rightarrow \infty$ (with all the remaining parameters also possibly changing in accordance to any scaling compatible with Assumption 2), the last expression satisfies

$$C_1 \left(\frac{\Delta}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right) \frac{1}{T} \asymp \frac{\Delta}{T} \left(\frac{1}{\log^{1/2}(n) \log^\xi(T)} \vee \frac{1}{\log^{1/2+2\xi}(T)} \right) \rightarrow 0, \quad (15)$$

since $\frac{\Delta}{T} \leq 1$. Interestingly, the above quantity remains vanishing even when $n \asymp 1$, which is achieved by introducing $\xi > 0$ in Assumption 2, so that consistent localization is possible even when the number of nodes is bounded. Of course, this is in striking contrast with the problem of

consistent estimation of the edge probabilities – or of the underlying graphon, if such a model is postulated – which at the minimum requires $n \rightarrow \infty$.

We emphasize that, while Theorem 1 shows that the NBS algorithm is consistent, the bound (11) on the localization rate it implies is slow, possibly only poly-logarithmically in T (in the worst case in which $n \gtrsim 1$). We do not make any claim to the optimality of such bound. This is somewhat unexpected, since NBS shares many features with Algorithm 3 of Wang et al. (2017), which leads to minimax optimal localization rates for change point estimation for covariance matrices under the worst possible scaling of the parameters for that problem. This difference is partially due to the fact that in this paper, the signal jumps are characterized in Frobenius norm, while operator norm is used in Wang et al. (2017).

What is more, a seemingly puzzling feature of the bound (11) is that it is decreasing in Δ , which is rather odd at first sight, since the smaller the value of the minimal spacing between consecutive change points the harder it should be to estimate the location of the change points. This is because the time interval considered in Theorem 1 is $[0, T]$. A standard way to analyze the error bound in the change point detection literature is to rescale the time interval so that the data are sampled on $[0, 1]$. Therefore one can normalize the localization error ϵ in (11) by T . By doing so, one can see that

$$\epsilon/T \leq \epsilon/\Delta = C_1 \log(T) \left(\frac{1}{\kappa_0 n \rho \sqrt{\Delta}} + \frac{\log^{1/2}(T)}{\kappa_0^2 n \rho \Delta} \right),$$

which indicates that the scaled localization rate decreases as Δ increases.

Proof of Theorem 1. As the random intervals $\{(\alpha_m, \beta_m)\}_{m=1}^M$ are generated independently from the data, we will assume throughout the proof that the event \mathcal{M} defined in (60) in Appendix E holds. By Lemma 25, the probability of the complementary event is smaller than

$$\exp \left\{ \log \left(\frac{T}{\Delta} \right) - \frac{M \Delta^2}{16 T^2} \right\},$$

which vanishes provided that

$$M \gtrsim (T/\Delta)^2 \log(T/\Delta).$$

For $1 \leq s < t < e \leq T$, we consider the event

$$\mathcal{A}(s, t, e) = \left\{ \left| (\tilde{A}^{s,e}(t), \tilde{B}^{s,e}(t)) - \|\tilde{\Theta}^{s,e}(t)\|_{\mathbb{F}}^2 \right| \leq C_\beta \log(T) \left(\|\tilde{\Theta}^{s,e}(t)\|_{\mathbb{F}} + \log^{1/2}(T) \rho n \right) \right\}. \quad (16)$$

Due to Lemma 7, it holds that $\mathbb{P}(\mathcal{A}(s, t, e)^c) \leq 6T^{-c_T} + 2T^{-c}$ for some $c, c_T > 3$, and, by a union bound argument,

$$\mathbb{P}(\mathcal{A}) = \mathbb{P} \left(\bigcup_{1 \leq s \leq t \leq e \leq T} \mathcal{A}(s, t, e) \right) \geq 1 - (6T^{3-c_T} + 2T^{3-c}).$$

All the analysis in the rest of this proof is conducted in the event $\mathcal{A} \cap \mathcal{M}$.

The general strategy of the proof is to utilize a standard induction-like argument that is commonly used in proving the consistency of change point estimators; see, e.g. Fryzlewicz (2014), Wang and Samworth (2016) and Wang et al. (2017). Of course the specific details and technicalities of this argument are new and challenging in our problem. In a nutshell, we will show that, in

the event $\mathcal{A} \cap \mathcal{M}$ and assuming that the algorithm has not made any mistakes so far in the detection and localization of change points, the procedure will also correctly identify any undetected change point and estimate its location within an error of ϵ , if such an undetected change point exists. Towards that end, it suffices to consider any generic time interval $(s, e) \subset (0, T)$ that satisfies

$$\eta_{r-1} \leq s \leq \eta_r \leq \dots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq -1$$

and

$$\max\{\min\{\eta_r - s, s - \eta_{r-1}\}, \min\{\eta_{r+q+1} - e, e - \eta_{r+q}\}\} \leq \epsilon,$$

where $q = -1$ indicates that there is no change point contained in (s, e) and ϵ is given in (11).

Observe that

$$\epsilon = C_1 \log(T) \left(\frac{\sqrt{\Delta}}{\kappa_0 n \rho} + \frac{\log^{1/2}(T)}{\kappa_0^2 n \rho} \right) \leq C_1 \left(\frac{\Delta}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right), \quad (17)$$

where the inequality follows from Assumption 1 part 1. and Assumption 2. Therefore, using the previous bound,

$$\epsilon \leq 2C_1 \Delta \max \left\{ \frac{1}{C_\alpha \log^{1/2}(n) \log^\xi(T)}, \frac{1}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right\} \leq \Delta/4,$$

by appropriately assuming C_α to be large enough. It, therefore, has to be the case that, for any change point $\eta_p \in (0, T)$, either $|\eta_p - s| \leq \epsilon$ or $|\eta_p - s| \geq \Delta - \epsilon \geq 3\Delta/4$. This means that $\min\{|\eta_p - e|, |\eta_p - s|\} \leq \epsilon$ indicates that η_p is a change point that has been previously detected and estimated within an error of magnitude ϵ in the previous induction step, even if $\eta_p \in (s, e)$. Below we will say that a change point η_p in $[s, e]$ is undetected if $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$.

In order to complete the induction step, it suffices to show that $\text{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$ (i) will not detect any new change point in (s, e) if all the change points in that interval have been previously detected, and (ii) will find a point b in (s, e) such that $|\eta_p - b| \leq \epsilon$ if there exists at least one undetected change point in (s, e) .

Step 1. Suppose that there does not exist any undetected change points within (s, e) . Then, for any $(s'_m, e'_m) = (\alpha_m, \beta_m) \cap (s, e)$, one of the following situations must hold:

- (a) there is no change point within (s'_m, e'_m) ;
- (b) there exists only one change point η_r within (s'_m, e'_m) and $\min\{\eta_r - s'_m, e'_m - \eta_r\} \leq \epsilon$ or
- (c) there exist two change points η_r, η_{r+1} within (s'_m, e'_m) and $\max\{\eta_r - s'_m, e'_m - \eta_{r+1}\} \leq \epsilon$.

We will analyze situation (c) only, as the other two cases are similar and in fact simpler. Observe that if (c) holds, then by (17) and (12),

$$\epsilon \leq \delta \Delta \leq \delta(e'_m - s'_m),$$

where the second inequality is fulfilled by choosing a sufficiently large C_α and because, by assumption, $\delta \leq 1/4$. Therefore, the interval

$$[s_m, e_m] = [s'_m + \delta(e'_m - s'_m), e_m - \delta(e'_m - s'_m)],$$

contains no change points. To see this, notice that, in the event of \mathcal{A} , $\tilde{\Theta}^{s_m, e_m}(t) = 0$ for all $t \in (s_m, e_m)$, as there is no change point in $[s_m, e_m]$. Furthermore, by Lemma 7, there exists a large enough constant $C_\beta > 0$ such that

$$\max_{s_m \leq t \leq e_m} (\tilde{A}^{s_m, e_m}(t), \tilde{B}^{s_m, e_m}(t)) \leq C_\beta \rho n \log^{3/2}(T).$$

Thus, with the input parameter τ satisfying

$$\tau \geq C_\beta \rho n \log^{3/2}(T),$$

we conclude that $\text{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$ will always correctly reject the existence of undetected change points.

Step 2. Suppose now that there exists a change point $\eta_p \in (s, e)$ such that $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$. Let a_m, b_m and m^* be defined as in $\text{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$. In the event \mathcal{M} , for any $\eta_p \in (s, e)$ such that $\min\{\eta_p - s, e - \eta_p\} \geq 3\Delta/4$, there exists an interval $[s'_m, e'_m]$ containing only one change point η_p such that

$$\eta_p - 3\Delta/4 \leq s'_m \leq \eta_p - \Delta/2 \quad \text{and} \quad \eta_p + \Delta/2 \leq e'_m \leq \eta_p + 3\Delta/4.$$

Therefore, if $[s_m, e_m] = [s'_m + \delta(e'_m - s'_m), e'_m - \delta(e'_m - s'_m)]$, then, since $\delta \leq 1/4$, one has that

$$\eta_p - \Delta 3/4 \leq s_m \leq \eta_p - \Delta/8 \quad \text{and} \quad \eta_p + \Delta/8 \leq e_m \leq \eta_p + \Delta 3/4. \quad (18)$$

Next, in the event \mathcal{A} , it holds that

$$(\tilde{A}^{s_m, e_m}(\eta_p), \tilde{B}^{s_m, e_m}(\eta_p)) \geq \|\tilde{\Theta}^{s_m, e_m}(\eta_p)\|_{\mathbb{F}}^2 - C_\beta \log(T) (\log^{1/2}(T) \rho n + \|\tilde{\Theta}^{s_m, e_m}(\eta_p)\|_{\mathbb{F}}).$$

It then follows from Lemma 18 that

$$\|\tilde{\Theta}^{s_m, e_m}(\eta_p)\|_{\mathbb{F}}^2 = \frac{(\eta_p - s_m)(e_m - \eta_p)}{e_m - s_m} \kappa_p^2 \geq \min\{e_m - \eta_p, \eta_p - s_m\} \kappa_p^2 \geq \kappa_p^2 \Delta/8,$$

where the last inequality stems from Equation (18). Thus, due to Assumptions 1 part 1. and 2, we conclude that

$$\kappa_p^2 \Delta/16 \geq \kappa_0^2 n^2 \rho^2 \Delta/16 \geq C_\alpha^2/16 n \rho \log^{2+2\xi}(T) > C_\beta n \rho \log^{3/2}(T),$$

and

$$\kappa_p \sqrt{\Delta}/4 \geq \kappa_0 n \rho \sqrt{\Delta}/4 \geq C_\alpha/4 \sqrt{n \rho} \log^{1+\xi}(T) \geq C_\alpha/4 \log^{1/2}(n) \log^{1+\xi}(T) > 2C_\beta \log(T). \quad (19)$$

Therefore, when $n, T \geq 2$ and

$$C_\beta < \min \left\{ 8^{-1} C_\alpha \log^\xi(T) \log^{1/2}(n), C_\alpha^2/16 \log^{1/2+2\xi}(T) \right\},$$

for some $c_2 > 0$, we arrive at

$$(\tilde{A}^{s_m, e_m}(\eta_p), \tilde{B}^{s_m, e_m}(\eta_p)) \geq c_2 \kappa_p^2 \Delta.$$

By definition of m^* , one then obtain the inequality

$$a_{m^*} = (\tilde{A}^{s_{m^*}, e_{m^*}}(b_{m^*}), \tilde{B}^{s_{m^*}, e_{m^*}}(b_{m^*})) \geq c_2(\kappa_{\max}^{s,e})^2 \Delta, \quad (20)$$

where $\kappa_{\max}^{s,e} = \max\{\kappa_k : \min\{\eta_p - s, e - \eta_p\} \geq 3\Delta/4\}$. Thus, with input parameter τ satisfying

$$\tau < c_2 \kappa_0^2 n^2 \rho^2 \Delta,$$

NBS can consistently detect the existence of undetected change points.

Step 3. Assume next that there exists at least one undetected change point $\eta_p \in (s, e)$ such that $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$. Let a_m, b_m and m^* be defined as in $\text{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$.

To complete the induction step and therefore the proof, it suffices to show that there exists a (necessarily undetected) change point $\eta_p \in [s_{m^*}, e_{m^*}]$ such that

$$\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4 \quad (21)$$

and that

$$|b_{m^*} - \eta_p| \leq \epsilon. \quad (22)$$

In this step we will prove that (21) holds. Denote

$$[s_{m^*}, e_{m^*}] = [s'_{m^*} + \delta(e'_{m^*} - s'_{m^*}), e_{m^*} - \delta(e'_{m^*} - s'_{m^*})].$$

Suppose for the sake of contradiction that

$$\max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2 < c_2(\kappa_{\max}^{s,e})^2 \Delta/2. \quad (23)$$

Then

$$\begin{aligned} & \max_{s_{m^*} \leq t \leq e_{m^*}} (\tilde{A}^{s_{m^*}, e_{m^*}}(t), \tilde{B}^{s_{m^*}, e_{m^*}}(t)) \\ & \leq \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2 + C_\beta \log(T) (\log^{1/2}(T) \rho n + \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}), \\ & \leq c_2(\kappa_{\max}^{s,e})^2 \Delta/2 + C_\beta \log^{3/2}(T) \rho n + C_\beta \log(T) \sqrt{c_2/2} \kappa_{\max}^{s,e} \sqrt{\Delta} \\ & < c_2(\kappa_{\max}^{s,e})^2 \Delta/2 + c_2(\kappa_{\max}^{s,e})^2 \Delta/4 + c_2(\kappa_{\max}^{s,e})^2 \Delta/4 = c_2(\kappa_{\max}^{s,e})^2 \Delta, \end{aligned}$$

where the first inequality is due to the definition of the event \mathcal{A} , the second inequality follows from (23) and the third inequality from Assumption 2, for an appropriately large C_α . This contradicts (20). Therefore

$$\max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2 \geq c_2(\kappa_{\max}^{s,e})^2 \Delta/2. \quad (24)$$

Observe that if $[s_{m^*}, e_{m^*}]$ contains two change points, then $e_{m^*} - s_{m^*} \geq \Delta$ and if $[s_{m^*}, e_{m^*}]$ contains one change point η , then it has to be the case that $\min\{\eta - s_{m^*}, e_{m^*} - \eta\} \geq c_2 \Delta/2$, as otherwise by Lemma 18,

$$\max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2 = \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(\eta)\|_{\mathbb{F}}^2 \leq c_2(\kappa_{\max}^{s,e})^2 \Delta/2,$$

which contradicts (24).

Therefore since $e_{m^*} - s_{m^*} \geq c_2 \Delta / 2 = \Delta / 32$, the bound (17) implies that

$$\epsilon \leq C_1 \left(\frac{\Delta}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right) \leq \delta(e'_{m^*} - s'_{m^*}), \quad (25)$$

where the second inequality follows if C_α is sufficiently large. By a similar argument as in **Step 1**, $[s_{m^*}, e_{m^*}]$ contains no detected change points. Observe that by (20), $[s_{m^*}, e_{m^*}]$ contains at least one undetected change point.

Step 4. In the final step of the proof we will show that (22) occurs. To that end, we will apply Lemma 8. Let

$$\lambda = \max_{s_{m^*} \leq t \leq e_{m^*}} |(\tilde{A}^{s_{m^*}, e_{m^*}}(t), \tilde{B}^{s_{m^*}, e_{m^*}}(t)) - \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2|. \quad (26)$$

Observe that (24) and (19) imply that

$$c_3 \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2 / 2 > C_\beta \log(T) \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}},$$

and

$$c_3 \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2 / 2 > C_\beta \log^{3/2}(T) \rho n,$$

for a sufficiently large $c_3 > 0$. Then, due to the definition of the event \mathcal{A} ,

$$\lambda \leq C_\beta \log(T) \left(\log^{1/2}(T) \rho n + \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}} \right) \leq c_3 \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2. \quad (27)$$

Since (40) follows from (20), (41) follows from (26), and (42) follows from (27), all the conditions in Lemma 8 hold. Lemma 8 implies that there exists η_p being an undetected change point within $[s, e]$ such that

$$|\eta_k - b| \leq \frac{C_3 \Delta \lambda}{\|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(\eta_k)\|_{\mathbb{F}}^2} \quad \text{and} \quad \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(\eta_k)\|_{\mathbb{F}}^2 \geq c' \max_{s_{m^*} \leq t \leq e_{m^*}} \|\tilde{\Theta}^{s_{m^*}, e_{m^*}}(t)\|_{\mathbb{F}}^2.$$

and this combining with (24) provides that

$$|\eta_k - b| \leq \frac{2C_3 C_\beta \log^{3/2}(T)}{c_2 (c')^2 \kappa_0^2 n \rho} + \frac{\sqrt{2} C_3 C_\beta \sqrt{\Delta} \log(T)}{c' \sqrt{c_2} \kappa_0 n \rho} \leq C_1 \log(T) \left(\frac{\log^{1/2}(T)}{\kappa_0^2 n \rho} + \frac{\sqrt{\Delta}}{\kappa_0 n \rho} \right),$$

where $C_1 > \frac{2C_3 C_\beta}{c_2 (c')^2} + \frac{\sqrt{2} C_3 C_\beta}{c' \sqrt{c_2}}$ and $c' < 2 \log(2) C_\beta / c_3$. This completes the induction. \square

3 Optimal Localization

In the previous section we saw how the NBS algorithm (see Algorithm 1) can consistently estimate the locations of the change points for the dynamic network model of Assumption 1 under nearly any scaling for which this task is feasible, albeit not at a fast rate and possibly not in an optimal manner. In this section, we are to show that under stronger, but still fairly general, conditions on both the model and the scaling, a two-step algorithm that builds on NBS to refine the estimators of the locations of the change points, will achieve a minimax optimal localization rate. The additional step beyond NBS is named LR (local refinement) and is detailed in Algorithm 3.

Algorithm 2 USVT(A, τ_2, τ_3)

INPUT: Symmetric matrix $A \in \mathbb{R}^{n \times n}$, $\tau_2, \tau_3 > 0$.

$(\kappa_i(A), v_i) \leftarrow$ the i th eigen-pair of A , with $|\kappa_1(A)| \geq \dots \geq |\kappa_n(A)|$

$A' \leftarrow \sum_{i: |\kappa_i(A)| \geq \tau_2} \kappa_i(A) v_i v_i^\top$

USVT(A, τ_2, τ_3) $\leftarrow (A''_{ij})$ with

$$(A'')_{ij} \leftarrow \begin{cases} (A')_{ij}, & \text{if } |(A')_{ij}| \leq \tau_3 \\ \text{sign}((A')_{ij})\tau_3, & \text{if } |(A')_{ij}| > \tau_3 \end{cases}$$

OUTPUT: USVT(A, τ_2, τ_3).

Algorithm 3 Local Refinement

INPUT: $\{A(t)\}_{t=1}^T, \{B(t)\}_{t=1}^T \in \mathbb{R}^{n \times n}$, $\tau_2, \tau_3, \delta > 0$, $\{\nu_k\}_{k=1}^K \subset \{1, \dots, T-1\}$, $\nu_0 = 0$, $\nu_{K+1} = T$.

for $k = 1, \dots, K$ **do**

$[s, e] \leftarrow [\nu_{k-1} + \delta(\nu_k - \nu_{k-1}), \nu_{k+1} - \delta(\nu_{k+1} - \nu_k)]$

$\tilde{\Delta}_k \leftarrow \sqrt{\frac{(e-\nu_k)(\nu_k-s)}{e-s}}$

$\tilde{\Theta}_k \leftarrow \text{USVT}(\tilde{B}^{s,e}(\nu_k), \tau_2, \tau_3 \tilde{\Delta}_k)$

$b_k \leftarrow \text{argmax}_{s \leq t \leq e} (A^{s,e}(t), \tilde{\Theta}_k)$

end for

OUTPUT: $\{b_k\}_{k=1}^K$.

The LR algorithm takes as input two identically distributed sequences of networks fulfilling Assumption 1 (obtained for instance by sample splitting), along with sequence $\{\nu_k\}_{k=1}^K$ of consistent estimators of the locations of the jumps, such as the one returned by NBS. Such estimators may be obtained from any one of the two sequences of networks that are also an input of LR (or may be obtained using other methods). The procedure inspects all the triplets of consecutive change point estimators one at a time (with the time points 0 and T as two dummy change points, for notational consistency). For each such triplet, LR utilizes the universal singular value thresholding (USVT) algorithm (Chatterjee, 2015) to construct a more accurate estimator of a local CUSUM matrix of the expected adjacency matrix at the middle point estimator. This estimator is in turn used to probe nearby locations in order to refine the original estimator of the location of the middle change point location. This results in a provably more precise estimator of that location. The conditions under which LR improves upon NBS are stronger are the ones that guarantee consistency of the latter, and are imposed in order to ensure that the USVT procedure is effective (see, e.g. Xu, 2017). We formalized them next.

Assumption 3. Let $\{\Theta(t)\}_{t=1}^T$ be defined as in Assumption 1. For all $k = 1, \dots, K$,

$$\text{rank}(\Theta(\eta_k) - \Theta(\eta_{k-1})) \leq r,$$

for some $r > 0$. Furthermore, for a sufficiently large $C_\alpha > 0$ and any $\xi > 0$,

$$\kappa_0 \sqrt{\rho} \geq C_\alpha \frac{\log^{1+\xi}(T)}{\sqrt{\Delta}} \sqrt{\frac{r}{n}}. \quad (28)$$

The first condition in the above assumption is about the model itself and requires that, in addition to all the properties listed in Assumption 1, the difference between any two different consecutive expected adjacency matrices to be of low rank. On the other hand while, the second condition imposes additional restrictions on the scalings of the model parameters compared to Assumption 2 used for the consistency of the NBS. The parameter r controlling the maximal rank of the difference of consecutive expected adjacency matrices is, like all the other parameters, allowed to change with T (provided that the scaling condition (28) is satisfied). We will revisit this assumption regarding ξ after Assumption 4.

Remark 9. *Assumption 3 is compatible with a broad range of parameter scalings. Focusing on the rank parameter, we highlight two extreme cases.*

- *When $r \asymp 1$, the scaling (28) is the same as that in Assumption 2, which is optimal.*
- *On the other hand, if the change points are far from each others so that $\Delta \asymp T$ and again $\kappa_0 \sqrt{\rho} \asymp n^{-1/2}$, then as long as $r \lesssim T \log^{-(2+2\xi)}(T)$, then Assumption 3 holds. This includes the situation where $T \log^{-(2+2\xi)}(T) \geq n$, which essentially leaves the order of magnitude of r unconstrained.*

Remark 10. *Note that if matrices A and B are of ranks r_1 and r_2 , then the difference $A - B$ is of rank at most $r_1 + r_2$. Therefore Assumption 3 can be viewed as an additional assumption on the rank of each network involved.*

3.1 Upper and lower bounds

The input of Local Refinement includes a sequence of change point estimators satisfying mild conditions and not necessarily consistent. As shown in Theorem 1 above, such a sequence may be obtained by applying the NBS algorithm, which, with high probability as T increases, estimates the correct number K of change points and their locations at an error ϵ (see (15)) satisfying

$$\epsilon \lesssim \Delta \left(\frac{1}{\log^{1/2}(n) \log^\xi(T)} \vee \frac{1}{\log^{1/2+2\xi}(T)} \right).$$

While other procedures besides NBS may return a consistent sequence of estimators of the change point locations at a different, possibly faster, rate, we will assume throughout that the LR algorithm is provided a preliminary sequence of K change point estimates from NBS satisfying the above error bound. We formalize this as an assumption below.

Assumption 4. *Let $\mathcal{B} = \{\nu_k\}_{k=1}^K \subset \{1, \dots, T\}$ be a collection of time points. Suppose that*

$$\max_{k=1, \dots, K} |\nu_k - \eta_k| = \epsilon_0 < \Delta/6.$$

Remark 11. *Recall that for Theorem 1, we have*

$$\epsilon = C_1 \log(T) \left(\frac{\sqrt{\Delta}}{\kappa_0 n \rho} + \frac{\log^{1/2}(T)}{\kappa_0^2 n \rho} \right) \leq C_1 \left(\frac{\Delta}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right).$$

Therefore, there exists a large enough $T_0 \in \mathbb{Z}_+$, such that for all $T \geq T_0$, it holds that

$$\frac{C_1}{C_\alpha \log^{1/2}(n) \log^\xi(T)} + \frac{C_1}{C_\alpha^2 \log^{1/2+2\xi}(T)} < 1/6,$$

which leads to that $\epsilon < \Delta/6$ and therefore with probability tending to 1, the outputs of the NBS satisfy Assumption 4.

Remark 12 (Recall Assumption 3). To be consistent with Assumption 2, we introduce ξ in Assumption 3, which is essentially unnecessary, but to cover the case when $r \asymp n \asymp 1$, $\Delta \asymp T$, the outputs of NBS still satisfy Assumption 4 automatically.

Theorem 2. Let Assumptions 1, 3 and 4 hold. For a large enough absolute constant $C_a > 0$ suppose that $\delta = 1/2$,

$$\tau_2 = (3/4)(C\sqrt{n\rho} + C_\epsilon \log(T)) \quad \text{and} \quad \tau_3 = \rho,$$

where $C > 64 \times 2^{1/4e^2}$ and $C_\epsilon > 12$. Then the collection of the estimated change points $\mathcal{B} = \{\hat{\eta}_k\}_{k=1}^K$ returned by LR with input parameters of $(0, T)$, $\{\nu_k\}_{k=1}^K$, τ_2 , τ_3 and δ satisfies

$$\mathbb{P}\left\{\max_{k=1, \dots, K} |\eta_k - \hat{\eta}_k| \leq \epsilon\right\} \geq 1 - 2T^{3-3C_\epsilon/4} - 4T^{3-3C_3^2/8},$$

where $\epsilon = C_2 \log^2(T) \kappa_0^{-2} n^{-2} \rho^{-1}$, $C_3 > 2\sqrt{2}$, and $C_2 > 0$ depending on C_α , C and C_3 .

Remark 13 (Input parameters τ_2 , τ_3 and δ). As also remarked in Xu (2017), the parameter τ_2 serves as a cutoff of the upper bound of the operator norm difference between the sample and population version matrices of interest. The choice we adopt here can be found in the large probability event \mathcal{A} defined in Step 3. The parameter τ_3 is an entrywise cutoff, and for more details see the arguments after (30). The parameter δ is chosen to be 1/2, which is arbitrarily chosen for convenience. It depends on the choice of the upper bound of ϵ_0 defined in Assumption 4.

Remark 14 (Tracking the constants in Theorem 2). It can be seen from the proof that we have the following hierarchy of the constants involved in Theorem 2. Firstly, C and C_ϵ are chosen to guarantee that $2T^{3-3C_\epsilon/4} \rightarrow 0$. Secondly, C_3 is chosen such that $4T^{3-3C_3^2/8} \rightarrow 0$. Finally $C_2 > 0$ depends on all the aforementioned constants and C_α .

Proof. Note that Algorithm 3 is parallelizable in the sense that we can deal with each $k \in \{1, \dots, K\}$ in parallel. For convenience, we have broken down the proof in five steps, each of which is applied to every $k \in \{1, \dots, K\}$. Before proceeding to the details, we have an overview of all steps.

In Step 1, we are to show that each working interval (s, e) contains one and only one true change point, and the two endpoints are well separated; Step 2 shows that the population CUSUM statistics within each working interval has good performances; the reasoning of the choices of the parameters in Algorithms 2 and 3, and the good performances of the sampler CUSUM statistics in large probability events, will be detailed in Step 3; additional probability controls regarding data splitting are demonstrated in Step 4; and finally to show the localization rates, we are to transfer the network CUSUM statistics into a univariate case in Step 5.

Step 1. By Assumption 4, $\eta_k \in [\nu_{k-1}, \nu_{k+1}]$ and

$$\begin{aligned}\eta_k - \nu_{k-1} &\geq \eta_k - \eta_{k-1} - |\eta_{k-1} - \nu_{k-1}| \geq \Delta - \epsilon_0 \geq 5\Delta/6, \\ \nu_{k+1} - \eta_k &\geq \eta_{k+1} - \eta_k - |\eta_{k+1} - \nu_{k+1}| \geq \Delta - \epsilon_0 \geq 5\Delta/6.\end{aligned}$$

Similar calculations show also that

$$\min\{\nu_k - \nu_{k-1}, \nu_{k+1} - \nu_k\} \geq \Delta - 2\epsilon_0 \geq 2\Delta/3.$$

Therefore with $\delta = 1/2$, it holds that

$$\delta \min\{\nu_k - \nu_{k-1}, \nu_{k+1} - \nu_k\} \geq \epsilon_0.$$

As a result, the interval

$$[s, e] = [\nu_{k-1} + \delta(\nu_k - \nu_{k-1}), \nu_{k+1} - \delta(\nu_{k+1} - \nu_k)]$$

contains only one change point η_k . We have that

$$\nu_k - s = (1 - \delta)(\nu_k - \nu_{k-1}) \geq (1 - \delta)2\Delta/3 = \Delta/3,$$

and $e - \nu_k \geq \Delta/3$. Therefore, $\min\{e - \nu_k, \nu_k - s\} \geq \Delta/3$.

Step 2. Let $\Lambda(k) = \Theta(\eta_k) - \Theta(\eta_{k-1})$. Then, by Lemma 18,

$$\|\tilde{\Theta}^{s,e}(t)\|_{\mathbb{F}}^2 = \begin{cases} \frac{t-s}{(e-s)(e-t)}(e - \eta_k)^2 \|\Lambda(k)\|_{\mathbb{F}}^2, & t \leq \eta_k, \\ \frac{e-t}{(e-s)(t-s)}(\eta_k - s)^2 \|\Lambda(k)\|_{\mathbb{F}}^2, & t \geq \eta_k. \end{cases}$$

Next, we set

$$\tilde{\Delta}_k = \sqrt{\frac{(\nu_k - s)(e - \nu_k)}{e - s}}$$

and, without loss of generality, we may assume that $\nu_k \leq \eta_k$. Since

$$\tilde{\Delta}_k \geq \min\{\nu_k - s, e - \nu_k\}/2 \geq \Delta/6,$$

we obtain that

$$\begin{aligned}\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\mathbb{F}}^2 &= \frac{\nu_k - s}{(e - s)(e - \nu_k)}(e - \eta_k)^2 \|\Lambda(k)\|_{\mathbb{F}}^2 \\ &= \tilde{\Delta}_k^2 \left(\frac{e - \eta_k}{e - \nu_k}\right)^2 \kappa_k^2 \\ &= \tilde{\Delta}_k^2 \left(1 - \frac{\eta_k - \nu_k}{e - \nu_k}\right)^2 \kappa_k^2 \\ &\geq \frac{\Delta}{6} \left(1 - \frac{\epsilon_0}{\Delta/3}\right)^2 \kappa_k^2 \\ &\geq \Delta \kappa_k^2 / 24.\end{aligned}\tag{29}$$

Step 3. We next apply Lemma 9 by letting $\varepsilon = C_\varepsilon \log(T)$, with $C_\varepsilon > 12$. Define the event

$$\mathcal{A} = \left\{ \sup_{1 \leq s \leq t \leq e \leq n} \|\tilde{A}^{s,e}(t) - \tilde{\Theta}^{s,e}(t)\|_{\text{op}} \leq C\sqrt{n\rho} + C_\varepsilon \log(T) \right\},$$

where $C > 64 \times 2^{1/4e^2}$. Due to Lemma 9, we have $\mathbb{P}(\mathcal{A}) \geq 1 - 2T^{3-C_\varepsilon/4}$.

We then apply Lemma 12. Set $\tau_2 = (3/4)(C\sqrt{n\rho} + C_\varepsilon \log(T))$, and define

$$\mathcal{B} = \left\{ \sup_{1 \leq s \leq t \leq e \leq n} \|\text{USVT}(\tilde{A}^{s,e}(t), \tau_2, \infty) - \tilde{\Theta}^{s,e}(t)\|_{\text{F}} \leq 3\sqrt{r}(C\sqrt{n\rho} + C_\varepsilon \log(T)) \right\}.$$

In order to apply Lemma 12, let $A = \tilde{A}^{s,e}(t)$, $B = \tilde{\Theta}^{s,e}(t)$, $\tau = \tau_2$ and $\delta = 1/3$. We then have $\mathbb{P}(\mathcal{B}) \geq 1 - 2T^{3-C_\varepsilon/4}$.

Let

$$\hat{A}^{s,e}(\nu_k) = \text{USVT}(\tilde{A}^{s,e}(\nu_k), \tau_2, \tau_3 \tilde{\Delta}_k). \quad (30)$$

Since $\nu_k \leq \eta_k$, for any $i, j = 1, \dots, n$, it holds that

$$\tilde{\Theta}_{ij}^{s,e}(\nu_k) = \sqrt{\frac{\nu_k - s}{(e-s)(e-\nu_k)}}(e - \eta_k) \Lambda_{ij}(k) \leq \tilde{\Delta}_k \rho \frac{e - \eta_k}{e - \nu_k} \leq \tilde{\Delta}_k \rho = \tilde{\Delta}_k \tau_3.$$

In the event \mathcal{B} ,

$$\|\hat{A}^{s,e}(\nu_k) - \tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}} \leq \|\text{USVT}(\tilde{A}^{s,e}(\nu_k), \tau_2, \infty) - \tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}} \leq 3\sqrt{r}(C\sqrt{n\rho} + C_\varepsilon \log(T)).$$

By the triangle inequality and Assumption 3, we have that

$$\|\hat{A}^{s,e}(\nu_k)\|_{\text{F}} \geq \|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}} - 3\sqrt{r}(C\sqrt{n\rho} + C_\varepsilon \log(T)) \geq c'_1 \sqrt{\Delta} \kappa_k, \quad (31)$$

where

$$c'_1 \leq 1/\sqrt{24} - \frac{3C}{C_\alpha \log^{1+\xi}(2)} - \frac{3C_\varepsilon}{C_\alpha \log^{1/2+\xi}(2)},$$

for any $n, T \geq 2$. As a consequence,

$$\begin{aligned} 2 \left(\frac{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}}{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}}, \frac{\|\hat{A}^{s,e}(\nu_k)\|_{\text{F}}}{\|\hat{A}^{s,e}(\nu_k)\|_{\text{F}}} \right) &= 2 - \left\| \frac{\tilde{\Theta}^{s,e}(\nu_k)}{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}} - \frac{\hat{A}^{s,e}(\nu_k)}{\|\hat{A}^{s,e}(\nu_k)\|_{\text{F}}} \right\|_{\text{F}}^2 \\ &\geq 2 - 4 \left(\frac{\|\tilde{\Theta}^{s,e}(\nu_k) - \hat{A}^{s,e}(\nu_k)\|_{\text{F}}}{\max \left\{ \|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}, \|\hat{A}^{s,e}(\nu_k)\|_{\text{F}} \right\}} \right)^2 \\ &\geq 2 - \frac{9r(C\sqrt{n\rho} + C_\varepsilon \log(T))^2}{(c'_1)^2 \kappa_k^2 \Delta} \\ &\geq 1, \end{aligned}$$

where the second inequality follows from the definition of the event \mathcal{B} and from (29), while the last inequality follows from Assumption 3, choosing a sufficiently large C_α . Therefore,

$$(\tilde{\Theta}^{s,e}(\nu_k), \hat{A}^{s,e}(\nu_k)) / \|\hat{A}^{s,e}(\nu_k)\|_{\text{F}} \geq \|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}} / 2 \geq (4\sqrt{6})^{-1} \sqrt{\Delta} \kappa_k, \quad (32)$$

where in the last inequality we have used again (29).

Step 4. Since $\{B(t)\}_{t=1}^T$ is independent of $\{A(t)\}_{t=1}^T$, the distribution of $\{B(t)\}_{t=1}^T$ does not change in the event \mathcal{B} . Observe that, from (30),

$$\|\widehat{A}^{s,e}(\nu_k)\|_\infty \leq \widetilde{\Delta}_k \tau_3 = \widetilde{\Delta}_k \rho.$$

In combination with (31), the previous inequality implies that

$$(e-s)^{-1/2} \|\widehat{A}^{s,e}(\nu_k)\|_\infty / \|\widehat{A}^{s,e}(\nu_k)\|_F \leq \frac{\rho}{c'_1 \sqrt{\Delta} \kappa_k}.$$

Using this bound along with Lemma 5, we obtain that, for any $\varepsilon > 0$,

$$\mathbb{P} \left(\left| \frac{1}{\sqrt{e-s}} \sum_{t=s+1}^e (\Theta(t) - B(t), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F) \right| \geq \varepsilon \right) \leq 2 \exp \left(\frac{-3/2\varepsilon^2}{3\rho + \varepsilon\rho / (c'_1 \kappa_k \sqrt{\Delta})} \right).$$

Setting $\varepsilon = C\sqrt{\rho} \log(T)$, with $C > 2\sqrt{2}$, we finally obtain the probabilistic bound

$$\mathbb{P} \left(\left| \frac{1}{\sqrt{e-s}} \sum_{t=s}^e (\Theta(t) - B(t), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F) \right| \geq C\sqrt{\rho} \log(T) \right) \leq 2T^{-3C^2/8}. \quad (33)$$

Similar arguments also show that

$$\mathbb{P} \left(\left| (\widetilde{\Theta}^{s,e}(t) - \widetilde{B}^{s,e}(t), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F) \right| \geq C\sqrt{\rho} \log(T) \right) \leq 2T^{-3C^2/8}. \quad (34)$$

Step 5. Consider the one dimensional time series $y(t) = (B(t), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F)$. Conditionally on $\{A(t)\}_{t=1}^T$, in the event \mathcal{B} , it holds that

$$t \in [s, e] \mapsto f(t) := \mathbb{E}(y(t)) = (\Theta(t), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F)$$

is a piecewise constant function with only one change point, namely η_k . Due to (32), it holds that

$$|\widetilde{f}^{s,e}(\eta_k)| = |(\widetilde{\Theta}^{s,e}(\eta_k), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F)| \geq |(\widetilde{\Theta}^{s,e}(\nu_k), \widehat{A}^{s,e}(\nu_k) / \|\widehat{A}^{s,e}(\nu_k)\|_F)| \geq (4\sqrt{6})^{-1} \sqrt{\Delta} \kappa_k,$$

and, by (33) and (34),

$$\mathbb{P} \left(\sup_{s \leq t \leq e} \left| \frac{1}{\sqrt{e-s}} \sum_{t=s}^e (x(t) - f(t)) \right| \geq C\sqrt{\rho} \log(T) \right) \leq 2T^{-c}$$

and

$$\mathbb{P} \left(\sup_{s \leq t \leq e} |\widetilde{x}^{s,e}(t) - \widetilde{f}^{s,e}(t)| \geq C\sqrt{\rho} \log(T) \right) \leq 2T^{-c},$$

where $c = 3(C^2/8 - 1) > 0$. We then apply Lemma 12 in Wang et al. (2017) by setting $\lambda = C\sqrt{\rho} \log(T)$. It follows that $b_k = \arg \max_{s \leq t \leq e} |\widetilde{x}^{s,e}(t)|$ is an undetected change point such that, for a large enough constant $C_2 > 0$,

$$|b_k - \eta_k| \leq C_2 \frac{\rho (\log T)^2}{\kappa_k^2}.$$

□

Let $\{A(t)\}_{t=1}^T$ satisfy Assumption 1 with only one change point and let $P_{\kappa_0, \Delta, n, \rho}^T$ denote the corresponding joint distribution.

Lemma 2. *Let $\{A(t)\}_{t=1}^T$ be a sequence of independent inhomogeneous Bernoulli networks satisfying Assumption 1 with $K = 1$ (i.e. there exists one and only one change point). Let $P_{\kappa_0, \Delta, n, \rho}^T$ denote the corresponding joint distribution. Consider the class of distributions*

$$\mathcal{Q} = \{P_{\kappa_0, \Delta, n, \rho}^T : \kappa_0 \leq 1/2, \rho \leq 1/2\}.$$

It holds that

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{Q}} \mathbb{E}_P(|\hat{\eta} - \eta|) \geq \max\{c\kappa_0^{-2}n^{-2}\rho, 1/2\}.$$

Note that the family of distributions \mathcal{Q} includes a wide range of changes. The constant $1/2$ is arbitrary here and can be replaced by any constant between 0 and 1. Recall the definition of jump size κ_0 and for simplicity denote the two graphons before and after the change point by Θ_1 and Θ_2 . The jump size $\kappa_0 n \rho = \|\Theta_1 - \Theta_2\|_F \leq n\rho/2$ means every entry is allowed to change.

Lemma 2 shows that the output of LR with high probability has a minimax optimal localization rate.

3.2 Sparse stochastic block model

Theorem 2 has shown that for network models with rank constraints, we demonstrate optimal localization for network change point detection. Low-rank network models include a wide range of common network models, e.g. Erdős–Rényi random graph model (Erdős and Rényi, 1959), stochastic block model (e.g. Holland et al., 1983) and random dot product model (Young and Scheinerman, 2007). However, when one uses the above mentioned models, one usually assumes there are no self-loops in the networks, i.e., the diagonal entries of the adjacency matrices are always 0, which violates the low-rank assumption. In order to fill in the gap regarding the diagonal entries, we use stochastic block model as an example in this subsection, to show that the optimal change point detection still possesses same performances even if we assume there are no self-loops in the networks.

Note that the difference of the population adjacency matrices of two stochastic block model networks still preserves block structures, while the difference possessing a block structure does not imply that it is the difference from two block structure networks. For completeness, we include a definition on sparse stochastic block model in Definition 3. In fact we tackle a more general setting with assumptions listed in Assumption 5.

Definition 3 (Sparse Stochastic Block Model). *A network is from a sparse stochastic block model with size n , sparsity parameter ρ , membership matrix $Z \in \mathbb{R}^{n \times s}$ and connectivity matrix $Q \in [0, 1]^{r \times r}$ if the corresponding adjacency matrix satisfies*

$$\mathbb{E}(A) = \rho Z Q Z^\top - \text{diag}(\rho Z Q Z^\top).$$

The membership matrix Z consists of rows, each of which has one and only one entry being 1 and has all the entries being 0; moreover, Z is a column full rank matrix, i.e. $\text{rank}(Z) = r$. The sparsity parameter $\rho \in [0, 1]$ potentially depends on n .

Assumption 5. Let $\{A(t)\}_{t=1}^T \in \mathbb{R}^{n \times n}$ be a sequence of independent adjacency matrices with $\Theta(t) = \mathbb{E}(A(t))$ satisfying Assumption 1.

In addition, assume for all $k \in \{1, \dots, K\}$, that

$$\Theta(\eta_k) - \Theta(\eta_{k-1}) = \Lambda(k) - \text{diag}(\Lambda(k)),$$

where $\Lambda(k) = Z_k Q_k Z_k^\top$, where Z_k is a membership matrix such that $\text{rank}(Z_k) \leq r$, and Q_k is a connectivity matrix.

There exists a sufficiently large $C_\alpha > 0$ and any $\xi > 0$ such that

$$\kappa_0 \sqrt{\rho} \geq C_\alpha \frac{\log^{1+\xi}(T)}{\sqrt{\Delta}} \sqrt{\frac{r}{n}}.$$

Observe that under Assumption 5, $\Theta(\eta_k) - \Theta(\eta_{k-1})$ is not a necessarily a low-rank matrix and therefore Assumption 3 may not hold. In the standard graphon estimation of stochastic block model, one has

$$\|\text{diag}(\Lambda(k))\|_{\text{op}} \leq \rho, \text{ and } \|\text{diag}(\Lambda(k))\|_{\text{F}} \leq \rho \sqrt{n},$$

which means one should expect to see that adding or subtracting the diagonal elements has little effect on the estimation rates.

However, in the change point setting, this is not the case. Observe that if $[s, e]$ contains one change point η_k , then for $t \in (s, e)$,

$$\tilde{\Theta}^{s,e}(t) = \begin{cases} \sqrt{\frac{t-s}{(e-s)(e-t)}}(e - \eta_k)(\Lambda(k) - \text{diag}(\Lambda(k))), & t \leq \eta_k, \\ \sqrt{\frac{e-t}{(e-s)(t-s)}}(\eta_k - s)(\Lambda(k) - \text{diag}(\Lambda(k))), & t \geq \eta_k. \end{cases}$$

In particular, at $t = \eta_k$,

$$\left\| \sqrt{\frac{(t - \eta_k)(e - \eta_k)}{(e - s)}} \text{diag}(\Lambda(k)) \right\|_{\text{op}} \lesssim \sqrt{\min\{e - \eta_k, \eta_k - s\}} \rho,$$

which depends on the spacing between change points. We therefore consider the following assumption.

Assumption 6. Assume $\|\Lambda(k)\|_{\text{F}} \geq C_\Lambda \|\text{diag}(\Lambda(k))\|_{\text{F}}$, where $C_\Lambda > 0$ is a large enough constant.

We conjecture that C_Λ can be optimized by anything greater than 1 if the constant in Lemma 14 is optimized accordingly. If $\Lambda(k)$ is a diagonally-dominant matrix, then it is unclear how to estimate $\Lambda(k)$ because in the no-self-loop networks, the diagonals of the adjacency matrices are always 0.

Theorem 3. In Theorem 2, if Assumption 3 is replaced by Assumption 5 and Assumption 6, then the same conclusion still holds.

Proof. In the proof of Theorem 2, note that arguments in **Steps 1** and **2** still hold under Assumptions in this theorem, and arguments in **Steps 4** and **5** will still hold if the conclusions in **Step 3** still holds.

Let $[s, e]$ be defined as that in the proof of Theorem 2. We apply Lemma 9 by letting $\varepsilon = C_\varepsilon \log(T)$, with $C_\varepsilon > 12$. Define the event

$$\mathcal{A}' = \left\{ \sup_{1 \leq s \leq t \leq e \leq n} \|\tilde{A}^{s,e}(t) - \tilde{\Theta}^{s,e}(t)\|_{\text{op}} \leq C \sqrt{n\rho} + C_\varepsilon \log(T) \right\},$$

where $C > 64 \times 2^{1/4e^2}$. Due to Lemma 9, we have $\mathbb{P}(\mathcal{A}') \geq 1 - 2T^{3-C_\varepsilon/4}$.

By Lemma 14, in the event \mathcal{A}' , it holds that

$$\begin{aligned} \mathcal{B}' &= \left\{ \sup_{1 \leq s \leq t \leq e \leq n} \|\text{USVT}(\tilde{A}^{s,e}(t), \tau_2, \infty) - \tilde{\Lambda}^{s,e}(t)\|_{\text{F}}^2 \right. \\ &\quad \left. \leq 9r(C\sqrt{n\rho} + C_\varepsilon \log(T))^2 + 512\|\text{diag}(\tilde{\Lambda}^{s,e}(\nu_k))\|_{\text{F}}^2 \right\}, \end{aligned}$$

where we choose $\delta = 1/3$ in Lemma 14 for convenience.

Let

$$\hat{A}^{s,e}(\nu_k) = \text{USVT}(\tilde{A}^{s,e}(\nu_k), \tau_2, \tilde{\Delta}_k \tau_3).$$

Observe that since $\nu_k \leq \eta_k$ and $\|\tilde{\Lambda}^{s,e}(\nu_k)\|_{\infty} \leq \tilde{\Delta}_k \tau_3$, in the event \mathcal{B}' ,

$$\begin{aligned} \|\hat{A}^{s,e}(\nu_k) - \tilde{\Lambda}^{s,e}(\nu_k)\|_{\text{F}} &\leq \|\text{USVT}(\tilde{A}^{s,e}(\nu_k), \tau_2, \infty) - \tilde{\Lambda}^{s,e}(\nu_k)\|_{\text{F}} \\ &\leq 3\sqrt{r}(C\sqrt{n\rho} + C_\varepsilon \log(T)) + 16\sqrt{2}\|\text{diag}(\tilde{\Lambda}^{s,e}(\nu_k))\|_{\text{F}}. \end{aligned}$$

Since $[s, e]$ contains only one change point η_k , by Assumption 6 and Lemma 18,

$$\begin{aligned} \|\hat{A}^{s,e}(\nu_k)\|_{\text{F}} &\geq \|\tilde{\Lambda}^{s,e}(\nu_k)\|_{\text{F}} - 3\sqrt{r}(C\sqrt{n\rho} + C_\varepsilon \log(T)) - 16\sqrt{2}\|\text{diag}(\tilde{\Lambda}^{s,e}(\nu_k))\|_{\text{F}} \\ &\geq \frac{C_\Lambda - 16\sqrt{2}}{C_\Lambda + 1} \|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}} - 3\sqrt{r}(C\sqrt{n\rho} + C_\varepsilon \log(T)) \geq c'_1 \sqrt{\Delta} \kappa_k, \end{aligned} \quad (35)$$

with $c'_1 > 0$ by choosing proper constants. As a consequence,

$$\begin{aligned} &2 \left(\frac{\tilde{\Theta}^{s,e}(\nu_k)}{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}}, \frac{\hat{A}^{s,e}(\nu_k)}{\|\hat{A}^{s,e}(\nu_k)\|_{\text{F}}} \right) = 2 - \left\| \frac{\tilde{\Theta}^{s,e}(\nu_k)}{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}} - \frac{\hat{A}^{s,e}(\nu_k)}{\|\hat{A}^{s,e}(\nu_k)\|_{\text{F}}} \right\|_{\text{F}}^2 \\ &\geq 2 - 2 \left(\frac{\|\tilde{\Theta}^{s,e}(\nu_k) - \hat{A}^{s,e}(\nu_k)\|_{\text{F}}}{\max\{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}, \|\hat{A}^{s,e}(\nu_k)\|_{\text{F}}\}} \right)^2 \\ &\geq 2 - 2 \left(\frac{9r(C\sqrt{n\rho} + C_\varepsilon \log(T))^2}{(c'_1)^2 \kappa_k^2 \Delta} + \frac{513\|\text{diag}(\tilde{\Lambda}^{s,e}(\nu_k))\|_{\text{F}}}{\|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}}} \right) \geq 1, \end{aligned}$$

where the second inequality follows from (29) and event \mathcal{B}' and the last inequality follows from Assumption 5 and (35). Therefore

$$(\tilde{\Theta}^{s,e}(\nu_k), \hat{A}^{s,e}(\nu_k)) / \|\hat{A}^{s,e}(\nu_k)\|_{\text{F}} \geq 1/2 \|\tilde{\Theta}^{s,e}(\nu_k)\|_{\text{F}} \geq c'' \sqrt{\Delta} \kappa_k.$$

Thus all the conclusions in **Step 3** of the proof of Theorem 2 still hold. \square

4 Discussion

In this paper, we dealt with the sparse dynamic network change point detection problem. There have been papers dealing with problems under the same name, but we believe that there is none as general as the one in our paper. We proposed two algorithms based on CUSUM statistics. The first one, namely Network Binary Segmentation, is able to detect change points consistently under

an optimal scaling (off by a log-factor); and the second one called Local Refinement builds upon mild inputs, which can be provided by NBS. Under a slightly stronger but realistic assumption, we have shown that LR yields optimal change point estimators, in the sense of optimal localization rates. To justify the optimality, we have shown the phase transition in terms of scaling, and the lower bounds of the localization rates.

The algorithms proposed in the paper are computationally efficient. We would also like to emphasize that we have provided careful arguments to fill in the gap between the commonly assumed low-rank and no self-loops assumptions for stochastic block models.

For readability, we did not state all the detailed requirements for every single constant involved. However, all the constants except the ones in and using Lemma 6 can be traced in the statements in the lemmas in Appendices. The appearance of those in Lemma 6 is due to the use of $\|\cdot\|_{\psi_2}$ -norm. They are in fact can be extracted in Lemma 5.5 in [Vershynin \(2010\)](#).

References

- AUE, A., HÖMANN, S., HORVÁTH, L. and REIMHERR, M. (2009). Break detection in the covariance structure of multivariate nonlinear time series models. *The Annals of Statistics*, **37** 4046–4087.
- BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science*, **286** 509–512.
- BARIGOZZI, M., CHO, H. and FRYZLEWICZ, P. (2016). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *arXiv preprint arXiv: 1612.06928*.
- BHATTACHARYYA, S. and CHATTERJEE, S. (2017). Spectral clustering for dynamic stochastic block model.
- BOCCALETTI, S., BIANCONI, G., CRIADO, R., DEL GENIO, C. I., GÓMEZ-GARDENES, J., ROMANCE, M., SENDINA-NADAL, I., WANG, Z. and ZANIN, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, **544** 1–122.
- CARRINGTON, P. J., SCOTT, J. and WASSERMAN, S. (eds.) (2005). *Models and methods in social network analysis*, vol. 28. Cambridge University Press.
- CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, **43** 177–214.
- CHO, H. (2015). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics* in press.
- CHU, L. and CHEN, H. (2017). Asymptotic distribution-free change-point detection for modern data. *arXiv preprint*.
- CRANE, H. (2015). Time-varying network models. *Bernoulli*, **21** 1670–1696.
- CRIBBEN, I. and YU, Y. (2017). Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66** 607–627.
- ERDŐS, P. and RÉNYI, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, **6** 290–297.

- FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76** 495–580.
- FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, **42** 2243–2281.
- GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics*, **43** 2624–2652.
- GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Foundations and Trends $\text{\textcircled{R}}$ in Machine Learning* 129–233.
- HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of American Statistical Association*, **105** 1480–1493.
- HO, Q., SONG, L. and XING, E. (2011). Evolving cluster mixed-membership blockmodel for time-evolving networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 342–350.
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* 109–137.
- HORVÁTH, L. and HUŠKOVÁ, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis*, **33** 631–648.
- KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* 016107.
- KOLACZYK, E. D. (2017). *Topics at the Frontier of Statistics and Network Analysis:(re) visiting the Foundations*. Cambridge University Press.
- LIN, K., SHARPNACK, J. L., RINALDO, A. and TIBSHIRANI, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.). 6887–6896.
- LIU, F., CHOI, D., XIE, L. and ROEDER, K. (2018). Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences of the United States of America*.
- MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79** 1119–1141.
- PENSKY, M. and ZHANG, T. (2017). Spectral clustering in dynamic stochastic block model.
- SAMSON, P.-M. (2000). Concentration of measure inequalities for markov chains and ψ -mixing processes. **28**.
- SEGINER, Y. (2000). The expected norm of random matrices. *Combinatorics, Probability and Computing*, **9** 149–166.

- SEWELL, D. K. and CHEN, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, **110** 1646–1657.
- SNIJDERS, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, **3** 1–40.
- TANG, M., PARK, Y., LEE, N. H. and PRIEBE, C. E. (2013). Attribute fusion in a latent process model for time series of graphs. *IEEE Transactions on Signal Processing*, **61** 1721–1732.
- TOMOZEI, D.-C. and MASSOULIÉ, L. (2014). Distributed user profiling via spectral methods. *Stochastic Systems*, **4** 1–43.
- TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.
- VENKATRAMAN, E. S. (1992). *Consistency results in multiple change-point problems*. Ph.D. thesis.
- VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.
- VOSTRIKOVA, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, **259** 270–274.
- WANG, D., YU, Y. and RINALDO, A. (2017). Optimal covariance change point detection in high dimension. *arXiv preprint*.
- WANG, H., TANG, M., PARK, Y. and PRIEBE, C. E. (2014). Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, **62** 703–717.
- WANG, T. and SAMWORTH, R. J. (2016). High-dimensional changepoint estimation via sparse projection. *arXiv preprint arXiv:1606.06246*.
- XU, A. and ZHENG, X. (2009). Dynamic social network analysis using latent space model and an integrated clustering algorithm. In *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*. 620–625.
- XU, J. (2017). Rates of convergence of spectral methods for graphon estimation. *arXiv preprint*.
- XU, K. (2015). Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*. 1079–1087.
- XU, K. S. and HERO, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, **8** 552–562.
- YAO, Y.-C. and AU, S.-T. (1989). Least-squares estimation of a stop function. *Sankhyā: The Indian Journal of Statistics, Series A* 370–381.
- YOUNG, S. J. and SCHEINERMAN, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*. 138–149.
- YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*. Springer, 423–435.

A Proofs and auxiliary lemmas used for lower bounds

In this subsection, we provide proofs of Lemma 1 in Section 2 and Lemma 2 in Section 3.1, which provide the minimax lower bounds for detection and localization respectively. In addition, Lemmas 3 and 4 are used in the proofs for Lemmas 1 and 2.

Lemma 3. *Let $\Theta \in \mathbb{R}^{n \times n}$ such that $\Theta_{ij} = \rho$ for all $1 \leq i, j \leq n$, where $0 < \rho < 1/2$. Let A be an adjacency matrix of a inhomogeneous Bernoulli network with independent edges such that $\mathbb{E}(A) = \Theta$. For any $v_b, v_c \in [-\sqrt{\rho}, \sqrt{\rho}]^n$, let B and C be adjacency matrices of inhomogeneous Bernoulli networks with independent edges such that $\mathbb{E}(B) = v_b v_b^\top + \Theta$ and $\mathbb{E}(C) = v_c v_c^\top + \Theta$. Let P_A, P_B, P_C be the distributions of A, B and C . Then*

$$\mathbb{E}_{P_A} \left(\frac{dP_B dP_C}{dP_A dP_A} \right) \leq \exp \left(\frac{(v_b^\top v_c)^2}{\rho(1-\rho)} \right).$$

Let $A' = A - \text{diag}(A)$, $B' = B - \text{diag}(B)$ and $C' = C - \text{diag}(C)$. Then

$$\mathbb{E}_{P_{A'}} \left(\frac{dP_{B'} dP_{C'}}{dP_{A'} dP_{A'}} \right) \leq \exp \left(\frac{(v_b^\top v_c)^2}{\rho(1-\rho)} \right).$$

Proof. Let $\Gamma = v_b v_b^\top$ and $\Lambda = v_c v_c^\top$.

$$\begin{aligned} \mathbb{E}_{P_A} \left(\frac{dP_B dP_C}{dP_A dP_A} \right) &= \prod_{1 \leq i, j \leq n} \left(\frac{(\Gamma_{ij} + \rho)(\Lambda_{ij} + \rho)}{\rho} + \frac{(1 - \Gamma_{ij} - \rho)(1 - \Lambda_{ij} - \rho)}{(1 - \rho)} \right) \\ &= \prod_{1 \leq i, j \leq n} \left(1 + \frac{\Gamma_{ij} \Lambda_{ij}}{\rho(1 - \rho)} \right) \leq \prod_{1 \leq i, j \leq n} \exp \left(\frac{\Gamma_{ij} \Lambda_{ij}}{\rho(1 - \rho)} \right) = \exp \left(\frac{(\Gamma, \Lambda)}{\rho(1 - \rho)} \right) = \exp \left(\frac{(v_b^\top v_c)^2}{\rho(1 - \rho)} \right). \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E}_{P_{A'}} \left(\frac{dP_{B'} dP_{C'}}{dP_{A'} dP_{A'}} \right) &= \prod_{i \neq j} \left(\frac{(\Gamma_{ij} + \rho)(\Lambda_{ij} + \rho)}{\rho} + \frac{(1 - \Gamma_{ij} - \rho)(1 - \Lambda_{ij} - \rho)}{(1 - \rho)} \right) \\ &= \prod_{i \neq j} \left(1 + \frac{\Gamma_{ij} \Lambda_{ij}}{\rho(1 - \rho)} \right) \leq \prod_{1 \leq i, j \leq n} \left(1 + \frac{\Gamma_{ij} \Lambda_{ij}}{\rho(1 - \rho)} \right) \leq \prod_{1 \leq i, j \leq n} \exp \left(\frac{\Gamma_{ij} \Lambda_{ij}}{\rho(1 - \rho)} \right) = \exp \left(\frac{(v_b^\top v_c)^2}{\rho(1 - \rho)} \right), \end{aligned}$$

where the first inequality follows from the observation that $\Gamma_{ii} = (v_b)_i^2 \geq 0$ and $\Lambda_{ii} = (v_c)_i^2 > 0$. \square

Remark 15. *Let $\Theta_{ij} = \rho + (v v^\top)_{ij}$, where $v \in \{\pm\sqrt{\kappa_0 \rho}\}^n$, $0 < \rho < 1/2$ and $0 < \kappa_0 < 1$, then the community labels can be decided according to the vector $\text{sign}(v)$. More precisely let*

$$\mathcal{C}_1 = \{i : v_i > 0\}, \text{ and } \mathcal{C}_2 = \{i : v_i < 0\}.$$

The probability within \mathcal{C}_1 or \mathcal{C}_2 is $\rho(1 + \kappa_0)$. The probability between \mathcal{C}_1 and \mathcal{C}_2 is $\rho(1 - \kappa_0)$.

Lemma 4. *Let $\{A_t\}_{t=1}^\Delta, \{B_{v,t}\}_{t=1}^\Delta \subset \mathbb{R}^{n \times n}$ be two sequences of adjacency matrices of independent inhomogeneous Bernoulli networks with independent edges, satisfying*

$$\mathbb{E}(A_t)_{ij} = \rho, \text{ and } \mathbb{E}(B_{v,t})_{ij} = \rho + \kappa_0 \rho (v v^\top)_{ij},$$

where $\rho < 1/2$, $\kappa_0 < 1$, and $v \in \{1, -1\}^n$.

1. Let P_0^Δ denote the joint distribution of $\{A_t\}_{t=1}^\Delta$ and $P_{v,1}^\Delta$ denote the joint distribution of $\{B_{v,t}\}_{t=1}^\Delta$. Denote by

$$P_1^\Delta = \frac{1}{2^n} \sum_{v \in \{\pm 1\}^n} P_{v,1}^\Delta.$$

If $\Delta \kappa_0^2 n \rho \geq 1/33$, then $\chi^2(P_0^\Delta, P_1^\Delta) \leq 1/16$, where $\chi^2(\cdot, \cdot)$ is the chi-square divergence (e.g. Section 2.4 [Tsybakov, 2009](#)).

2. Let $A'_t = A_t - \text{diag}(A_t)$ and $B'_t = B_t - \text{diag}(B_t)$. Let Q_0^Δ denote the joint distribution of $\{A'_t\}_{t=1}^\Delta$ and $Q_{v,1}^\Delta$ denote the joint distribution of $\{B'_{v,t}\}_{t=1}^\Delta$. Denote by

$$Q_1^\Delta = \frac{1}{2^n} \sum_{v \in \{\pm 1\}^n} Q_{v,1}^\Delta.$$

If $\Delta \kappa_0^2 n \rho \geq 1/33$, then $\chi^2(Q_0^\Delta, Q_1^\Delta) \leq 1/16$.

Proof. We first prove 1. Observe that since $v \in \{1, -1\}^n$, $\|vv^\top\|_F = n$. Let $U, V \in \mathbb{R}^n$ be two independent random vectors with entries being independent Rademacher random variables. It holds that

$$\begin{aligned} \chi^2(P_1^\Delta, P_0^\Delta) + 1 &= \mathbb{E}_{P_0^\Delta} \left(\frac{dP_1^\Delta}{dP_0^\Delta} - 1 \right)^2 = \frac{1}{4^n} \sum_{u, v \in \{\pm 1\}^n} \mathbb{E}_{P_0^\Delta} \left(\frac{dP_{u,1}^\Delta}{dP_0^\Delta} \frac{dP_{v,1}^\Delta}{dP_0^\Delta} \right) \\ &= \frac{1}{4^n} \sum_{u, v \in \{\pm 1\}^n} \left\{ \mathbb{E}_{P_0} \left(\frac{dP_{u,1}}{dP_0} \frac{dP_{v,1}}{dP_0} \right) \right\}^\Delta \leq \frac{1}{4^n} \sum_{u, v \in \{\pm 1\}^n} \exp \left(\frac{\Delta \kappa_0^2 \rho (u, v)^2}{1 - \rho} \right) \\ &= \mathbb{E}_{U, V} \left\{ \exp \left(\frac{\Delta \kappa_0^2 \rho (U, V)^2}{1 - \rho} \right) \right\} = \mathbb{E}_V \left\{ \exp \left(\frac{\Delta \kappa_0^2 \rho (I, V)^2}{1 - \rho} \right) \right\}, \end{aligned} \quad (36)$$

where $I = (1, \dots, 1)^n$, P_0 is the distribution of A_t , $P_{v,1}$ and $P_{u,1}$ are the distributions of $B_{v,t}$ and $B_{u,t}$, respectively.

Let $\varepsilon_n = \frac{(I, V)^2}{n^2}$. Applying Hoeffding's inequality, we have that, for any $\lambda > 0$,

$$\mathbb{P}(\varepsilon_n \geq \lambda) \leq 2e^{-2n\lambda}. \quad (37)$$

Thus,

$$\begin{aligned} \chi^2(P_1^\Delta, P_0^\Delta) + 1 &\leq \mathbb{E} \left(\exp \left(\varepsilon_n \frac{\Delta \kappa_0^2 n^2 \rho}{1 - \rho} \right) \right) = \int_0^\infty \mathbb{P} \left\{ \exp \left(\varepsilon_n \frac{\Delta \kappa_0^2 n^2 \rho}{1 - \rho} \right) \geq u \right\} du \\ &\leq 1 + \int_1^\infty \mathbb{P} \left\{ \varepsilon_n \geq \log(u) \frac{1 - \rho}{\Delta \kappa_0^2 n^2 \rho} \right\} du \\ &\leq 1 + \int_1^\infty 2 \exp \left\{ -2 \log(u) \frac{(1 - \rho)}{\Delta \kappa_0^2 n \rho} \right\} du \\ &= 1 - \frac{2}{1 - \frac{2(1 - \rho)}{\kappa_0^2 n \rho \Delta}}, \end{aligned}$$

where the first inequality is proved in Lemma 3, the third inequality follows from (37) and the last identity holds if $\frac{2(1-\rho)}{\Delta\kappa_0^2 n\rho} > 1$. As the function $x \in (1, \infty) \mapsto 1 - \frac{2}{1-x}$ is strictly decreasing and converges from above to 1 as $x \rightarrow \infty$, the last display implies that, for $\frac{2(1-\rho)}{\kappa_0^2 n\rho\Delta}$ sufficiently large, the term $\chi^2(P_1^\Delta, P_0^\Delta)$ can be made arbitrarily close to 0. In particular, by taking $\frac{2(1-\rho)}{\Delta\kappa_0^2 n\rho} = 33$, which is equivalent to $\Delta\kappa_0^2 n\rho > 1/33$ under the condition that $\rho < 1/2$, the desired result follows.

As for **2**, by second part of Lemma 3, using the same calculations in (36), it can be shown that

$$\chi^2(Q_1^\Delta, Q_0^\Delta) + 1 \leq \mathbb{E}_V \left(\exp \left(\frac{\Delta\kappa_0^2 \rho}{1-\rho} (I, V)^2 \right) \right),$$

and the desired result follows. \square

Proof of Lemma 1. Let $J \in \mathbb{R}^{n \times n}$ be such that $J_{ij} = 1$, for all $i, j \in \{1, \dots, n\}$. For any vector $u \in \{-1, 1\}^n$, denote $B_u = \rho\kappa_0 uu^\top$.

Step 1. Let $P_{0,u}^T$ denote the joint distribution of $\{A(t)\}_{t=1}^T$ satisfying

$$\mathbb{E}(A(1)) = \dots = \mathbb{E}(A(\Delta)) = \rho J + B_u \quad \text{and} \quad \mathbb{E}(A(\Delta + 1)) = \dots = \mathbb{E}(A(T)) = \rho J.$$

Let $P_{1,u}^T$ denote the joint distribution of $\{A(t)\}_{t=1}^T$ satisfying

$$\mathbb{E}(A(1)) = \dots = \mathbb{E}(A(T - \Delta)) = \rho J \quad \text{and} \quad \mathbb{E}(A(T - \Delta + 1)) = \dots = \mathbb{E}(A(T)) = \rho J + B_u.$$

For $i = 0, 1$, let $P_i^T = \frac{1}{2^n} \sum_{u \in \{\pm 1\}^n} P_{i,u}^T$. Let $\eta(P_{i,u}^T)$ denote the location of the change point associated to the distribution $P_{i,u}^T$. Then since $\eta(P_{0,u}^T) = \Delta$ and $\eta(P_{1,u}^T) = T - \Delta$ for any $u \in \{\pm 1\}^n$, $|\eta(P_{0,u}^T) - \eta(P_{1,u}^T)| \geq T/3$. By Le Cam's Lemma (see, e.g. [Yu, 1997](#)),

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|\hat{\eta} - \eta|) \geq (T/3)(1 - d_{\text{TV}}(P_0^T, P_1^T)), \quad (38)$$

where $d_{\text{TV}}(P_0^T, P_1^T) = \frac{1}{2} \|P_0^T - P_1^T\|_1$.

Step 2. It follows from the same derivation in the proof of Lemma 3 in [Wang et al. \(2017\)](#), we have

$$\|P_0^T - P_1^T\|_1 \leq 2\|P_0^\Delta - P_1^\Delta\|_1,$$

where P_0^Δ is the joint distribution of $\{A(t)\}_{t=1}^\Delta$ where $\mathbb{E}(A(1)) = \dots = \mathbb{E}(A(\Delta)) = \rho J$, and $P_1^\Delta = \frac{1}{2^n} \sum_{u \in \{\pm 1\}^n} P_{1,u}^\Delta$, where $P_{1,u}^\Delta$ is the joint distribution of $\mathbb{E}(A(1)) = \dots = \mathbb{E}(A(\Delta)) = \rho J + B_u$. Thus (38) leads to

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|\hat{\eta} - \eta|) \geq (T/3)(1 - \|P_0^\Delta - P_1^\Delta\|_1).$$

By Lemma 4, if $\Delta\kappa_0^2 n\rho \geq 1/33$, then it follows from Equation (2.27) in [Tsybakov \(2009\)](#) that

$$\|P_0^\Delta - P_1^\Delta\|_1 \leq 1/4,$$

and this completes the proof. \square

Proof of Lemma 2. Let $\Theta(1), \Theta(2) \in \mathbb{R}^{n \times n}$ such that for all $i, j = 1, \dots, n$, $\Theta_{ij}(1) = \rho/2$ and that $\Theta_{ij}(2) = \rho/2 + \kappa_0\rho$. Since $\kappa_0 \leq 1/2$, it holds that $\|\Theta(2)\|_\infty \leq \rho$.

For $\delta > 0$ to be chosen later, let P_1^δ be the joint distribution of a collection of independent adjacency matrices $\{A(t)\}_{t=1}^T$ such that

$$\mathbb{E}(A(t)) = \begin{cases} \Theta(1), & \text{if } t \leq T/2 + \delta, \\ \Theta(2), & \text{if } t > T/2 + \delta. \end{cases}$$

Let P_2^δ be the joint distribution of a collection of independent adjacency matrices $\{B(t)\}_{t=1}^T$ such that

$$\mathbb{E}(B(t)) = \begin{cases} \Theta(1), & \text{if } t \leq T/2, \\ \Theta(2), & \text{if } t > T/2. \end{cases}$$

Then we have,

$$\begin{aligned} 2d_{TV}^2(P_1, P_2) &\leq KL(P_1, P_2) \\ &= \delta n^2 \left((\rho/2 + \kappa_0\rho) \log \left(\frac{\rho/2 + \kappa_0\rho}{\rho/2} \right) + (1 - \rho/2 - \kappa_0\rho) \log \left(\frac{1 - \rho/2 - \kappa_0\rho}{1 - \rho/2} \right) \right) \\ &\leq \delta n^2 \left((\rho/2 + \kappa_0\rho) \frac{\kappa_0\rho}{\rho/2} + (1 - \rho/2 - \kappa_0\rho) \frac{-\kappa_0\rho}{1 - \rho/2} \right) \\ &= \delta n^2 (\kappa_0\rho + 2\kappa_0^2\rho - \kappa_0\rho + \kappa_0^2\rho^2(1 - \rho/2)^{-1}) \leq 4\delta\kappa_0^2 n^2 \rho = 4\delta\kappa_0^2 n^2 \rho. \end{aligned}$$

Since

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|\hat{\eta} - \eta|) \geq \delta(1 - d_{TV}(P_1, P_2)),$$

taking $\delta = \frac{1}{8\kappa_0^2 n^2 \rho}$, we have

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|\hat{\eta} - \eta|) \geq \frac{1}{16\kappa_0^2 n^2 \rho}.$$

□

B Technical details in Section 2

Observe that in Section 2, no additional structure is imposed on the adjacency matrix. In addition, for two matrices $A, B \in \mathbb{R}^{n \times n}$, we have

$$(A, B) = \{\text{vec}(A)\}^\top \text{vec}(B),$$

where $\text{vec}(\cdot)$ is the vectorized version of a matrix by stacking the columns thereof. It, therefore, suffices to view A as a sparse Bernoulli vector with $p = n^2$ entries. The assumptions below are vector versions of Assumption 1. We include them here for brevity.

Assumption 7. Let $X(1), \dots, X(T) \in \mathbb{R}^p$ be independent random vectors with independent Bernoulli entries. Suppose that the i th coordinate $X_i(t)$ of $X(t)$ satisfies $\mathbb{E}(X_i(t)) = \mu_i(t)$ and that

$$\sup_{1 \leq t \leq T} \|\mu(t)\|_\infty \leq \rho.$$

Note that in fact if A is an adjacency matrix of an inhomogeneous Bernoulli network defined in Definition 1, then due to symmetry, there are in fact $p = n(n-1)/2$ independent entries. In this section, for notational simplicity, we let $p = n^2$ which has the same order as $n(n-1)/2$.

Assumption 8. Let $\{\eta_k\}_{k=0}^{K+1} \subset \{0, \dots, n\}$ be a collection of change points, such that $\eta_0 = 0$ and $\eta_{K+1} = T$ and that

$$\mu(\eta_k + 1) = \mu(\eta_k + 2) = \dots = \mu(\eta_{k+1}), \text{ for any } k = 0, \dots, K.$$

Assume the spacing Δ satisfy that

$$\inf_{k=1, \dots, K+1} \{\eta_k - \eta_{k-1}\} \geq \Delta > 0,$$

and the normalized jump size κ_0 satisfies

$$\inf_{k=1, \dots, K+1} \|\mu(\eta_k) - \mu(\eta_{k-1})\| = \inf_{k=1, \dots, K+1} \kappa_k \geq \kappa_0 \rho \sqrt{p} > 0.$$

B.1 Probability bounds

In this subsection, our task is to provide a probability bound for the event $\mathcal{A}(s, e, t)$ defined in eq. (16) to hold. The result is formally stated in Lemma 7, and necessary technical details are provided in Lemmas 5 and 6.

Suppose $\{w_t\}_{t=1}^T \subset \mathbb{R}$ satisfies

$$\sum_{t=1}^T w_t^2 = 1. \quad (39)$$

Lemma 5. Suppose Assumption 7 holds. Let $v \in \mathbb{R}^p$ and $\{w_t\}_{t=1}^T \subset \mathbb{R}$ satisfy (39). Then for any $\varepsilon > 0$, we have

$$\mathbb{P} \left(\left| \sum_{i=1}^p v_i \sum_{t=1}^T w_t (X_i(t) - \mu_i(t)) \right| \geq \varepsilon \right) \leq 2 \exp \left(- \frac{3/2 \varepsilon^2}{3\rho \|v\|_2^2 + \varepsilon \max_{i=1}^p |v_i| \max_{t=1}^T |w_t|} \right).$$

Proof. Observe that

$$\mathbb{E} \left(\sum_{i=1}^p v_i \sum_{t=1}^T w_t (X_i(t) - \mu_i(t)) \right)^2 = \sum_{i=1}^p \sum_{t=1}^T v_i^2 w_t^2 \mathbb{E} (X_i(t) - \mu_i(t))^2 \leq \rho \|v\|_2^2,$$

due to the independence assumption and the fact that $\sum_{t=1}^T w_t^2 = 1$, and that

$$\max_{\substack{t=1, \dots, T \\ i=1, \dots, p}} |w_t v_i (X_i(t) - \mu_i(t))| \leq \max_{i=1}^p |v_i| \max_{t=1}^T |w_t|,$$

since $X_i(t)$ is a Bernoulli random variable with mean $\mu_i(t)$. The desired result follows from Bernstein inequality. \square

Lemma 6. Assume that the collection $\{Y(t)\}_{t=1}^T$ satisfies Assumption 7. Let $v = \sum_{t=1}^T w_t(Y(t) - \mu(t)) \in \mathbb{R}^p$. Then there exists $C > 0$ depending on $c > 0$ such that

$$\mathbb{P}\left(\max_{1 \leq i \leq p} v_i \geq C\sqrt{\log(p) \vee \log(T)}\right) \leq T^{-c},$$

and

$$\mathbb{P}\left(\|v\| \geq C\sqrt{\log(p) \vee \log(T)} + \sqrt{\rho p}\right) \leq T^{-c},$$

Proof. For the first part observe that it follows from Lemma 5.9 in Vershynin (2010) that there exists some absolute constant $C_1 > 0$ such that

$$\|v_i\|_{\psi_2}^2 \leq C_1 \sum_{t=1}^T w_t^2 \|Y_i(t) - \mu_i(t)\|_{\psi_2}^2 \leq 2C_1,$$

where $\|\cdot\|_{\psi_2}$ is the Orlicz norm (e.g. Definition 5.7 in Vershynin, 2010), and the second inequality follows from $\|Y_i(t) - \mu_i(t)\|_{\psi_2}^2 \leq 2$ and $\sum_{t=1}^T w_t^2 = 1$. Therefore for each $i = 1, \dots, p$, v_i is sub-Gaussian and there exist a constant $c > 0$ and a large enough $C > 0$ depending on c and C_1 such that

$$\mathbb{P}\left(v_i \geq C\sqrt{\log(p) \vee \log(T)}\right) \leq (p \vee T)^{-c-1}.$$

Since

$$p(p \vee T)^{-c-1} \leq \begin{cases} T^{-c}, & p \leq T; \\ p^{-c} \leq T^{-c}, & p \geq T, \end{cases}$$

the desired result follows from a union bound argument.

For the second part, define $F(x_1, \dots, x_p) = \|x\|$ and $G_i(y_1, \dots, y_t) = \sum_{t=1}^T w_t(y_t - \mu_i(t))$, $i = 1, \dots, p$. Since both F and G_i for all i are one Lipschitz function, $\|v\|$ is a one Lipschitz function of $\{\{Y_i(t)\}_{i=1}^p\}_{t=1}^T$. It follows from the proof of Corollary 4 in Samson (2000) that, for any $\varepsilon > 0$,

$$\mathbb{P}(\|v\| > \mathbb{E}\|v\| + \varepsilon) \leq \exp(-\varepsilon^2/2).$$

Since $\mathbb{E}\|v\| \leq \sqrt{\sum_{i=1}^p \mathbb{E}(v_i^2)} \leq \sqrt{\rho p}$, the desired results follows by taking $\varepsilon = C\sqrt{\log(p) \vee \log(T)}$. \square

Lemma 7. Let $\{X(t)\}_{t=1}^T$ and $\{Y(t)\}_{t=1}^T$ be two independent copies both of which satisfying Assumption 7. Suppose in addition that

$$\rho\sqrt{p} \geq \log(p).$$

For $\{w_t\}_{t=1}^T$ satisfying $\sum_{t=1}^T w_t^2 = 1$, let $\tilde{X} = \sum_{t=1}^T w_t X(t)$, $\tilde{Y} = \sum_{t=1}^T w_t Y(t)$ and $\tilde{\mu} = \sum_{t=1}^T w_t \mu(t)$. There exists $C_\beta > 0$ depending on c and c_T such that

$$\mathbb{P}\left(\left|\sum_{i=1}^p \tilde{X}_i \tilde{Y}_i - \sum_{i=1}^p \tilde{\mu}_i^2\right| \geq C_\beta \log(T) \left(\|\tilde{\mu}\| + \log^{1/2}(T)\rho\sqrt{p}\right)\right) \leq 6T^{-c_T} + 2T^{-c},$$

where $C_\beta > \max\{4c_T/3, \sqrt{3c_T(C+1)^2 + C^2}\}$, and C, c are from Lemma 5.

Proof. Note that $\sum_{i=1}^p \tilde{X}_i \tilde{Y}_i - \sum_{i=1}^p \tilde{\mu}_i^2 = I + II + III$, where

$$\begin{aligned} I &= \sum_{i=1}^p (\tilde{X}_i - \tilde{\mu}_i)(\tilde{Y}_i - \tilde{\mu}_i), \\ II &= \sum_{i=1}^p (\tilde{X}_i - \tilde{\mu}_i)\tilde{\mu}_i, \\ III &= \sum_{i=1}^p (\tilde{Y}_i - \tilde{\mu}_i)\tilde{\mu}_i. \end{aligned}$$

It suffices to bound I and II , due to the fact that $\{X(t)\}_{t=1}^T$ and $\{Y(t)\}_{t=1}^T$ are iid.

As for I , for any $i = 1, \dots, p$, let $v_i = \sum_{t=1}^T w_t (Y_i(t) - \mu_i(t))$. Conditional on $\{Y(t)\}_{t=1}^T$, it follows from Lemma 5 that for any $\varepsilon > 0$, we have

$$\mathbb{P}_{X|Y} \left(\left| \sum_{i=1}^p v_i \sum_{t=1}^T w_t (X_i(t) - \mu_i(t)) \right| \geq \varepsilon \right) \leq 2 \exp \left(-\frac{3/2\varepsilon^2}{3\rho\|v\|^2 + \varepsilon \max_i |v_i|} \right),$$

due to the fact that $\max_t |w_t| \leq 1$. By Lemma 6, there exist $C, c > 0$ such that

$$\mathbb{P}_Y \left(\max_{i=1, \dots, p} |v_i| \geq C\sqrt{\log(p) \vee \log(T)} \right) \leq T^{-c},$$

and that

$$\mathbb{P}_Y \left(\|v\| \geq C\sqrt{\log(p) \vee \log(T)} + \sqrt{\rho p} \right) \leq T^{-c}.$$

Thus for any $\varepsilon > 0$, it holds that

$$\begin{aligned} & \mathbb{P}_{X,Y} \left(\left| \sum_{i=1}^p v_i \sum_{t=1}^T w_t (X_i(t) - \mu_i(t)) \right| \geq \varepsilon \right) \\ & \leq 2 \exp \left(-\frac{3/2\varepsilon^2}{3\rho(C\sqrt{\log(p) \vee \log(T)} + \sqrt{\rho p})^2 + C\varepsilon\sqrt{\log(p) \vee \log(T)}} \right) + 2T^{-c}. \end{aligned}$$

Since $\rho\sqrt{p} \geq \log(p)$, by taking $\varepsilon = C''\rho\sqrt{p}\log^{3/2}(T)$ for sufficiently large

$$C'' \geq \sqrt{3c_T(C+1)^2 + C^2},$$

it holds that

$$\mathbb{P}(|I| \geq C''\rho\sqrt{p}\log^{3/2}(T)) \leq 2T^{-c_T} + 2T^{-c}.$$

Observe that III is identically distributed as II . For II , observe that for $\varepsilon > 0$, it follows from Lemma 5,

$$\mathbb{P} \left(\left| \sum_{i=1}^p \tilde{\mu}_i \sum_{t=1}^T w_t (X_i(t) - \mu_i(t)) \right| \geq \varepsilon \right) \leq 2 \exp \left(-\frac{3/2\varepsilon^2}{3\rho\|\tilde{\mu}\|^2 + \varepsilon \max_i |\tilde{\mu}_i| \max_t |w_t|} \right).$$

Let $\varepsilon = C'\|\tilde{\mu}\| \log(T)$, with $C' > 4c_T/3$,

$$3\rho\|\tilde{\mu}\|^2 + \varepsilon \max_i |\tilde{\mu}_i| \max_t |w_t| \leq 3\rho\|\tilde{\mu}\|^2 + \varepsilon\rho \leq 3\|\tilde{\mu}\|^2 + \varepsilon \leq 3/(2c_T)\varepsilon^2 / \log(T).$$

Therefore $P(|II| \geq C'\|\tilde{\mu}\| \log(T)) \leq 2T^{-c_T}$.

□

B.2 Localization

This is the key lemma used in the proof of Theorem 1 to locate the change points. We deliberately present this lemma with seemingly low-level conditions, in order for us to directly check the conditions in the proof of Theorem 1.

Lemma 8. *Assume $\{X_t\}_{t=1}^T$ and $\{Y_t\}_{t=1}^T$ be two independent copies $\mathbb{E}(X_t) = \mathbb{E}(Y_t) = \mu(t)$ such that Assumption 8 holds.*

Let $[s_0, e_0]$ be any interval with $e_0 - s_0 \leq C_R \Delta$ and containing at least one change point η_r such that

$$\eta_{r-1} \leq s \leq \eta_r \leq \dots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 0$$

and that $\min\{s_0 - \eta_r, e_0 - \eta_{r+q}\} \geq \Delta/2$. Denote $\kappa_{\max}^{s,e} = \max\{\kappa_p : \min\{\eta_p - s_0, e_0 - \eta_p\} \geq \Delta/16\}$. Consider any generic $[s, e] \subset [s_0, e_0]$ such that $[s, e]$ contains at least one change point. Let $b \in \arg \max_{s < t < e} (\tilde{X}^{s,e}(t), \tilde{Y}^{s,e}(t))$. For some $c_1 > 0$, $\lambda > 0$ and $\delta > 0$, suppose that

$$(\tilde{X}^{s,e}(b), \tilde{Y}^{s,e}(b)) \geq c_1 (\kappa_{\max}^{s,e})^2 \Delta \tag{40}$$

$$\sup_{s < t < e} |(\tilde{X}^{s,e}(t), \tilde{Y}^{s,e}(t)) - \|\tilde{\mu}^{s,e}(t)\|_2^2| \leq \lambda \tag{41}$$

If there exists a sufficient small $c_2 > 0$ satisfying

$$c_2 < \min \left\{ \frac{c_3}{2C_R^2 + 2c_3}, \frac{1}{2 + 32C_R^2 / \min\{1/4, 1/2 - 2c_3\}} \right\}$$

with c_3 defined in Lemma 16, such that

$$\lambda \leq c_2 \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|_2^2 \tag{42}$$

then there exists a change point $\eta_k \in (s, e)$ such that

$$\min\{e - \eta_k, \eta_k - s\} > \Delta/4, \quad |\eta_k - b| \leq \frac{C_3 \Delta \lambda}{\|\tilde{\mu}^{s,e}(\eta_k)\|_2^2} \quad \text{and} \quad \|\tilde{\mu}^{s,e}(\eta_k)\|_2^2 \geq (1 - 2c_2) \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|_2^2,$$

where $C_3 = 2C_R^2 / \min\{1/4 - 1/2c_3\}$.

Proof. For any $t \in \{s+1, \dots, e-1\}$, denote $\tilde{Z}^{s,e}(t) = (\tilde{X}^{s,e}(t), \tilde{Y}^{s,e}(t))$. It follows from Proposition 1 that without loss of generality, we can assume $\|\tilde{\mu}^{s,e}(t)\|_2^2$ is locally decreasing at b . Observe that this implies there exists a change point $\eta_k \in [s, b]$, since otherwise $\|\tilde{\mu}^{s,e}(t)\|_2^2$ is increasing on $[s, b]$ as a consequence of Lemma 19. Therefore, we have

$$s \leq \eta_k \leq b \leq e.$$

Observe that

$$\tilde{Z}^{s,e}(b) \geq \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|_2^2 - \lambda \geq c_2^{-1} (1 - c_2) \lambda, \tag{43}$$

which follows from (41) and (42), and

$$\|\tilde{\mu}^{s,e}(b)\|_2^2 \geq \tilde{Z}^{s,e}(b) - \lambda \geq \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|_2^2 - 2\lambda \geq c_2^{-1} (1 - 2c_2) \lambda, \tag{44}$$

which follows from (43). We, consequently, have

$$\|\tilde{\mu}^{s,e}(b)\|^2 \geq \tilde{Z}^{s,e}(b) - \lambda \geq (1 - c_2(1 - c_2)^{-1})\tilde{Z}^{s,e}(b) > \tilde{Z}^{s,e}(b)/2 \geq (c_1/2)(\kappa_{\max}^{s,e})^2\Delta. \quad (45)$$

where the second inequality follows from (43) and the last inequality follows from (40).

Since $s \leq \eta_k \leq b \leq e$ and $\|\tilde{\mu}^{s,e}(t)\|_2^2$ is locally decreasing at b , by Proposition 1, $\|\tilde{\mu}^{s,e}(t)\|_2^2$ is decreasing within $[\eta_k, b]$. Therefore

$$\|\tilde{\mu}^{s,e}(\eta_k)\|^2 \geq \|\tilde{\mu}^{s,e}(b)\|^2. \quad (46)$$

Equation (46) combining with (44) gives

$$\|\tilde{\mu}^{s,e}(\eta_k)\|^2 \geq (1 - 2c_2) \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|^2.$$

Step 1. In this step, it is shown that $\min\{\eta_k - s, e - \eta_k\} \geq \min\{1, c_1\}\Delta/16$.

Suppose η_k is the only change point in (s, e) . It must hold that $\min\{\eta_k - s, e - \eta_k\} \geq \min\{1, c_1\}\Delta/16$, otherwise by Lemma 18,

$$\|\tilde{\mu}^{s,e}(\eta_k)\|^2 = \frac{(\eta_k - s)(e - \eta_k)}{e - s} \kappa_k^2 < \frac{c_1}{16} \kappa_k^2 \Delta \leq \frac{c_1}{2} (\kappa_{\max}^{s,e})^2 \Delta,$$

which contradicts (45).

Suppose $[s, e]$ contains at least two change points. For the sake of contradiction, suppose $\min\{\eta_k - s, e - \eta_k\} < \min\{1, c_1\}\Delta/16$. Reversing the time series if necessary, it suffices to consider

$$\eta_k - s < \min\{1, c_1\}\Delta/16. \quad (47)$$

Observe that (47) implies that η_k is the first change point in $[s, e]$. Therefore

$$\begin{aligned} \|\tilde{\mu}^{s,e}(\eta_k)\|^2 &\leq \frac{1}{8} \|\tilde{\mu}^{s,e}(\eta_{k+1})\|^2 + 4\kappa_r^2(\eta_k - s) \leq \frac{1}{8} \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|^2 + \frac{c_1}{4} \kappa_k^2 \Delta \\ &\leq \frac{1}{8} (1 - 2c_2)^{-1} \|\tilde{\mu}^{s,e}(b)\|^2 + \frac{1}{2} \|\tilde{\mu}^{s,e}(b)\|^2 < \|\tilde{\mu}^{s,e}(b)\|_2, \end{aligned}$$

where the first inequality follows from Lemma 20 and (47), the second inequality follows from (47), the third inequality follows from (44) and (45), and the fourth inequality follows from $c_2 < 3/8$. This contradicts (46).

Step 2. In order to apply Lemma 16, it suffices to check that (55) for $\tilde{\mu}^{s,e}(t)$. Observe that

$$\begin{aligned} \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|^2 - \|\tilde{\mu}^{s,e}(\eta_k)\|^2 &\leq 2c_2 \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|^2 \leq 2c_2(1 - 2c_2)^{-1} \|\tilde{\mu}^{s,e}(\eta_k)\|^2 \\ &\leq \frac{2c_2 C_R^2}{c_3(1 - 2c_2)} c_3 \|\tilde{\mu}^{s,e}(\eta_k)\|^2 \Delta^2 (e - s)^{-2} \leq c_3 \|\tilde{\mu}^{s,e}(\eta_k)\|^2 \Delta^2 (e - s)^{-2}, \end{aligned}$$

where c_3 is defined as in (55), the first and the second inequality follow from (44), the third inequality follows from $e - s \leq C_R \Delta$ and the last inequality hold for sufficiently small

$$0 < c_2 < \frac{c_3}{2C_R^2 + 2c_3}.$$

Let c be defined in Lemma 16. Since $e - s \leq C_R \Delta$,

$$\frac{2\lambda(e-s)^2}{c\Delta\|\tilde{\mu}^{s,e}(\eta_k)\|^2} \leq 2C_R^2 \frac{\lambda\Delta}{cc_2^{-1}(1-c_2)\lambda} < \Delta/16,$$

where the first inequality follows from (44) and the last inequality holds for sufficiently small c_2 satisfying

$$c_2 < \frac{1}{2 + 32C_R^2/c}.$$

By Lemma 16 if d is chosen such that

$$d - \eta_k = \frac{2\lambda(e-s)^2}{c\Delta\|\tilde{\mu}^{s,e}(\eta_k)\|^2} < \Delta/16, \quad (48)$$

and that

$$\|\tilde{\mu}^{s,e}(\eta_k)\|^2 - \|\tilde{\mu}^{s,e}(d)\|^2 > c\|\tilde{\mu}^{s,e}(\eta_k)\|^2 |d - \eta_k| \Delta (e-s)^{-2} \geq 2\lambda, \quad (49)$$

where the first inequality follows from Lemma 16 and the second inequality follows from (48).

For the sake of contradiction, suppose $b \geq d$. Then

$$\|\tilde{\mu}^{s,e}(b)\|^2 \leq \|\tilde{\mu}^{s,e}(d)\|^2 < \|\tilde{\mu}^{s,e}(\eta_k)\|^2 - 2\lambda \leq \max_{s < t < e} \|\tilde{\mu}^{s,e}(t)\|^2 - 2\lambda \leq \max_{s < t < e} \tilde{Z}(t) + \lambda - 2\lambda = \tilde{Z}(b) - \lambda,$$

where the first inequality follows Proposition 1, which ensures that $\|\tilde{\mu}^{s,e}(t)\|^2$ is decreasing on $[\eta_k, b]$ and $d \in [\eta_k, b]$, the second inequality follows from (49). This is a contradiction to (41). Thus $b \leq d$ and so

$$0 \leq b - \eta_k \leq d - \eta_k \leq \frac{2\lambda(e-s)^2}{c\Delta\|\tilde{\mu}^{s,e}(\eta_k)\|^2} \leq \frac{2C_R^2}{c} \frac{\Delta\lambda}{\|\tilde{\mu}^{s,e}(\eta_k)\|^2}$$

where the third inequality follows from $e - s \leq C_R \Delta$. □

C Technical details in Section 3

C.1 Matrix estimation

In this subsection, some results concerning matrix estimation is established.

Lemma 9. 1. *Let $\{A(t)\}_{t=1}^T$ be a collection of independent matrices with independent Bernoulli entries satisfying*

$$\sup_{1 \leq t \leq T} \|\mathbb{E}A(t)\|_\infty \leq \rho,$$

with $n\rho \geq \log(n)$. Let $\{w(t)\}_{t=1}^T \subset \mathbb{R}$ be a collection of scalar such that $\sum_{t=1}^T w(t)^2 = 1$ and $\sum_{t=1}^T w(t) = 0$. Then there exists an absolute constant $C > 32 \times 2^{1/4} e^2$ such that

$$\mathbb{P} \left(\left\| \sum_{t=1}^T w(t)A(t) - \mathbb{E} \left(\sum_{t=1}^T w(t)A(t) \right) \right\|_{\text{op}} \geq C\sqrt{n\rho} + \varepsilon \right) \leq \exp(-\varepsilon^2/2). \quad (50)$$

2. *If $\{A(t)\}_{t=1}^T$ are symmetric matrices, then (50) still holds.*

Proof. Observe that the conclusion in **2** is consequence of that in **1**, as if $A(t)$ is symmetric, then $A(t) = A'(t) + A''(t)$, where $A'(t)$ is the upper diagonal matrix of $A(t)$ including the diagonal and $A''(t)$ is the lower diagonal matrix of $A(t)$. Therefore the conclusion in **2** follows by applying the conclusion in **1** to $A'(t)$ and $A''(t)$. In the rest of this proof, we will only consider **1**.

Let $B(t) = A(t) - \mathbb{E}(A(t))$ and $\tilde{B} = \sum_{t=1}^T w(t)B(t)$. The function

$$H(B(1), \dots, B(T)) = \left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}} = \|\tilde{B}\|_{\text{op}}$$

is one-Lipschitz, therefore by Corollary 4 in [Samson \(2000\)](#), one has for any $\varepsilon > 0$,

$$\mathbb{P} \left\{ \left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}} \geq \mathbb{E} \left(\left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}} \right) + \varepsilon \right\} \leq \exp(-\varepsilon^2/2).$$

To complete the argument, it suffices to bound $\mathbb{E} \left(\left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}} \right)$. By Lemma 10 and for all $t \in \{1, \dots, T\}$, the entries of $w(t)B(t)$ are bounded on $[-\rho, 1]$, there exists a collection of random matrices $\{Z(t)\}_{t=1}^T \subset \mathbb{R}^{n \times n}$ such that $\mathbb{E}(Z|B) = B$, where $Z = (Z(1), \dots, Z(T))$ and $B = (B(1), \dots, B(T))$, and that $(Y_t)_{ij} = (1 - \rho)(Z(t))_{ij} + \rho$ are mutually independent Bernoulli random variables with parameter ρ . Denote $G(B) = \left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}}$. Then

$$\begin{aligned} \mathbb{E} \left(\left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}} \right) &= \mathbb{E}(G(B)) = \mathbb{E}(G(\mathbb{E}(Z|B))) \leq \mathbb{E}(\mathbb{E}(G(Z)|B)) = \mathbb{E} \left(\left\| \sum_{t=1}^T w(t)Z(t) \right\|_{\text{op}} \right) \\ &= \frac{1}{1 - \rho} \mathbb{E} \left(\left\| \sum_{t=1}^T w(t)Y(t) \right\|_{\text{op}} \right), \end{aligned}$$

where G being convex is used in the inequality and $\sum_{t=1}^T w(t) = 0$ is used in the last equality. Since the entries of $\sum_{t=1}^T w(t)Y(t)$ are independent and identically distributed, by Lemma 11,

$$\mathbb{E} \left(\left\| \sum_{t=1}^T w(t)B(t) \right\|_{\text{op}} \right) \leq C\sqrt{\rho n},$$

where $C > 32 \times 2^{1/4}e^2$. □

Lemma 10. *Let $X \in [-\rho, 1]$ be a centered Bernoulli random variable. Then there exists a random variable Y such that*

- $\mathbb{E}(Y|X) = X$, and
- $(1 - \rho)Y + \rho$ is a Bernoulli random variable with parameter ρ .

Proof. The proof is taken from the proof of Lemma 2 in [Tomozei and Massoulié \(2014\)](#), by letting

$$Y = 1 - \mathbb{1}\{X \leq (1 + \rho)U - \rho\},$$

where U is a Uniform $[0, 1]$ random variable independent with X . □

Lemma 11. Let $\{A(t)\}_{t=1}^T$ be a collection of independent adjacency matrices whose entries are independent Bernoulli random variables with parameter ρ satisfying with $n\rho \geq c_2 \log(n)$, $c_2 > 4$, and let $B_t = A_t - E(A_t)$. Suppose $\{w_t\}_{t=1}^T \subset \mathbb{R}$ be a collection of scalar such that $\sum_{t=1}^T w_t^2 = 1$. Then there exists an absolute constant $C > 32 \times 2^{1/4} e^2$ such that

$$\mathbb{E} \left(\left\| \sum_{t=1}^T w_t B_t \right\|_{\text{op}} \right) \leq C \sqrt{n\rho}.$$

Proof. Let $\tilde{B} = \sum_{t=1}^T w_t B(t)$. To bound, $\mathbb{E}(\|\tilde{B}\|_{\text{op}})$, since the entries of \tilde{B} are independent and identically distributed with $E(\tilde{B}) = 0$, by Corollary 2.2 in [Seginer \(2000\)](#), one has

$$\mathbb{E} \left(\|\tilde{B}\|_{\text{op}} \right) \leq C_1 \mathbb{E} \left(\max_{1 \leq i \leq n} \|\tilde{B}_{i*}\| \right),$$

where $C_1 = 16 \times 2^{1/4} e^2$. For any $i \in \{1, \dots, n\}$, since $\|\tilde{B}_{i*}\|$ is one-Lipschitz convex function, by Corollary 4 in [Samson \(2000\)](#), it holds that for any $\varepsilon > 0$,

$$\mathbb{P} \left(\|\tilde{B}_{i*}\| \geq \mathbb{E}\|\tilde{B}_{i*}\| + \varepsilon \right) \leq \exp(-\varepsilon^2/2).$$

Since

$$\mathbb{E}(\|\tilde{B}_{i*}\|)^2 \leq \mathbb{E}(\|\tilde{B}_{i*}\|^2) = \sum_{t=1}^T w_t^2 \mathbb{E}(\|(B(t))_{i*}\|^2) + \sum_{s \neq t} w_s w_t \mathbb{E}(B_s, B_t) = \sum_{t=1}^T w_t^2 n\rho(1 - \rho) \leq n\rho,$$

one has

$$\mathbb{P} \left(\|\tilde{B}_{i*}\| \geq \sqrt{n\rho} + \varepsilon \right) \leq \exp(-\varepsilon^2/2). \quad (51)$$

Using the above display, it follows that

$$\begin{aligned} \mathbb{E} \left(\max_{1 \leq i \leq n} \|\tilde{B}_{i*}\| \right) &= \int_0^\infty \mathbb{P} \left(\max_{1 \leq i \leq n} \|\tilde{B}_{i*}\|_2 \geq \varepsilon \right) d\varepsilon \leq \int_0^{2\sqrt{\rho n}} 1 d\varepsilon + \int_{2\sqrt{\rho n}}^\infty n \mathbb{P}(\|\tilde{B}_{1*}\| \geq \varepsilon) d\varepsilon \\ &= 2\sqrt{\rho n} + \int_{\sqrt{\rho n}}^\infty n \mathbb{P}(\|\tilde{B}_{1*}\| \geq \varepsilon + \sqrt{\rho n}) d\varepsilon \leq 2\sqrt{\rho n} + \int_{\sqrt{\rho n}}^\infty n \exp(-\varepsilon^2/2) d\varepsilon \\ &\leq 2\sqrt{\rho n} + \frac{1}{\sqrt{\rho n}} \int_{\sqrt{\rho n}}^\infty n \varepsilon \exp(-\varepsilon^2/2) d\varepsilon \leq 2\sqrt{\rho n} + n^{1-c_2/2} \frac{1}{\sqrt{c_2 \log(n)}} < C_2 \sqrt{\rho n}, \end{aligned}$$

where $C_2 > 2$, the first inequality follows from the observation that $\|\tilde{B}_{i*}\|$ are identically distributed, the second inequality follows from (51) and the last inequality follows from $\rho n \geq c_2 \log(n)$, $c_2 > 2$. \square

Lemmas 12 and 13 are from Lemma 1 in [Xu \(2017\)](#).

Lemma 12. Let $A, B \in \mathbb{R}^{n \times n}$ be two symmetric matrices with $\|A - B\|_{\text{op}} < \tau/(1 + \delta)$, $\tau > 0$. Then for a fixed $\delta < 1$, we have

$$\|\text{USVT}(A, \tau, \infty) - B\|_{\text{F}}^2 \leq 16 \min_{0 \leq r \leq n} \left\{ r\tau^2 + (1 + \delta)^2 \delta^{-2} \sum_{i=r+1}^n \lambda_i^2(B) \right\}.$$

Lemma 13. Let A and B be defined as in Lemma 12, and that $\|B\|_{\infty} \leq \tau'$, then

$$\|\text{USVT}(A, \tau, \tau') - B\|_{\text{F}}^2 \leq 16 \min_{0 \leq r \leq n} \left\{ r\tau^2 + (1 + \delta)^2 \delta^{-2} \sum_{i=r+1}^n \lambda_i^2(B) \right\}.$$

C.2 Proofs in Section 3.2

Lemma 14. *Suppose $A, \Lambda \in \mathbb{R}^{n \times n}$ are symmetric matrices with Bernoulli entries satisfying $\|\Lambda\|_\infty \leq \rho$ and $\|A - (\Lambda - \text{diag}(\Lambda))\|_{\text{op}} \leq (1 + \delta)\tau$. Then*

$$\|\text{USVT}(A, \tau, \infty) - \Lambda\|_{\text{F}}^2 \leq 16 \min_{0 \leq r \leq n} \left\{ r\tau^2 + 2(1 + \delta)^2 \delta^{-2} \sum_{i=r+1}^n \lambda_i \right\} + 32(1 + \delta)^2 \delta^{-2} \|\text{diag}(\Lambda)\|_{\text{F}}^2,$$

where $\{\lambda_i\}_{i=1}^n$ are the eigenvalues of Λ ordered in decreasing absolute values.

Proof. Let $\{\lambda'_i\}_{i=1}^n$ be the eigenvalues of $\Lambda - \text{diag}(\Lambda)$ ordered in absolute value, $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of Λ ordered in absolute value and $\{v_i\}_{i=1}^n$ be the eigenvectors of Λ . Observe that for any orthonormal basis $\{u_i\}_{i=1}^n$, and any $r = 1, \dots, n-1$,

$$\sum_{i=r+1}^n (\lambda'_i)^2 \leq \sum_{i=r+1}^n u_i^\top (\Lambda - \text{diag}(\Lambda))^2 u_i.$$

By Lemma 12, one has

$$\|\text{USVT}(A, \tau, \infty) - (\Lambda - \text{diag}(\Lambda))\|_{\text{F}}^2 \leq 16 \min_{0 \leq r \leq n} \left\{ r\tau^2 + (1 + \delta)^2 \delta^{-2} \sum_{i=r+1}^n (\lambda'_i)^2 \right\}.$$

For any $r = 1, \dots, n$,

$$\begin{aligned} \sum_{i=r+1}^n (\lambda'_i)^2 &\leq \sum_{i=r+1}^n v_i^\top (\Lambda - \text{diag}(\Lambda))^2 v_i - v_i^\top \Lambda^2 v_i + \sum_{i=r+1}^n \lambda_i^2 \\ &= \sum_{i=r+1}^n v_i^\top (-2\Lambda \text{diag}(\Lambda) + \text{diag}(\Lambda)^2) v_i + \sum_{i=r+1}^n \lambda_i^2 \leq \sum_{i=r+1}^n \|\Lambda v_i\|_2^2 + 2v_i^\top \text{diag}(\Lambda)^2 v_i + \sum_{i=r+1}^n \lambda_i^2 \\ &\leq 2 \sum_{i=r+1}^n \lambda_i^2 + 2\|\text{diag}(\Lambda)\|_{\text{F}}^2, \end{aligned}$$

which leads to the desired results. \square

D Properties of population CUSUM statistics

Recall that in Definition 1 we introduced a general version of CUSUM statistics, which can be applied to various types of data. In Sections D.1 and D.2, we apply Definition 1 to vectors and scalars respectively.

D.1 Vector version CUSUM

Assumption 9. *Let $\{V(t)\}_{t=1}^T \subset \mathbb{R}^p$. Assume there exists $\{\nu_m\}_{m=0}^M \subset \{1, \dots, T\}$ such that*

$$V(\nu_m + 1) = \dots = V(\nu_{m+1}) \quad \text{for all } m = 0, \dots, M-1,$$

and that $\inf_{m=0, \dots, M} \|V(\nu_m) - V(\nu_{m+1})\| = \inf_{m=0, \dots, M} \kappa_m \geq \kappa = \kappa_0 \rho \sqrt{p}$.

The results in this subsection are used in the proofs of the main theorems. Below, $\{V(t)\}_{t=1}^T$ corresponds to $\{\mu(t)\}_{t=1}^T$ as defined in Assumption 1, and $\kappa = \kappa_0 \rho \sqrt{p}$ (see Assumption 1). For brevity, we introduce new notation in this subsection such that it is self-contained within this subsection.

For $1 \leq s < t < e \leq T$, denote the CUSUM statistics

$$\tilde{V}^{s,e}(t) = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{r=s+1}^t V(r) - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{r=t+1}^e V(r). \quad (52)$$

For simplicity denote $\tilde{V}(t) = \tilde{V}^{0,T}(t)$. It is desired to show that this vector version CUSUM statistics have the same properties as the 1D CUSUM.

Remark 16. *The CUSUM statistics defined in (52) is translational invariant. In other words, let $W \in \mathbb{R}^p$ and $U(t) = V(t) + W$ for all t , then*

$$\tilde{V}(t) = \tilde{U}(t).$$

Consequently it can be assumed that $\sum_{t=1}^T V(t) = 0$, and

$$\tilde{V}(t) = \left(\sum_{r=1}^t V(r) - \frac{t}{T} \sum_{r=1}^T V(r) \right) / \sqrt{\frac{t(T-t)}{T}} = \left(\sum_{r=1}^t V(r) \right) / \sqrt{\frac{t(T-t)}{T}}. \quad (53)$$

Proposition 1. *The quantity $\|\tilde{V}(t)\|^2$ is maximized at the change points. For $t \in [\nu_{m-1}, \nu_m]$, $\|\tilde{V}(t)\|^2$ is either monotone or decreases and then increases.*

Proof. Let $t \in (\nu_{m-1}, \nu_m)$. By Equation (2.7) of Lemma 2.2 in Venkatraman (1992), for every $j = 1, \dots, p$, $\tilde{V}_j(t)$ can be continuously extended to the function

$$f_j(x) = \frac{a_j - b_j x}{\sqrt{x(1-x)}},$$

where $x = t/T$, a_j and b_j are defined similarly as in Lemma 2.2 in Venkatraman (1992). Thus it suffices to show that for $x \in (c, d)$ where $0 \leq c \leq d \leq 1$, the function

$$f(x) = \sum_{j=1}^p \frac{(a_j - b_j x)^2}{x(1-x)}$$

is maximized at either c or d .

Let

$$f'(x) = \sum_{j=1}^n \frac{-(2a_j x - b_j x - a_j)(b_j x - a_j)}{(x-1)^2 x^2} = \frac{g(x)}{(x-1)^2 x^2}.$$

The desired result follows if $f'(x)$ is either nonpositive, or nonnegative or that there exists $x_0 \in (0, 1)$ such that

$$f'(x) \begin{cases} \leq 0 & \text{when } x \leq x_0 \\ \geq 0 & \text{when } x \geq x_0 \end{cases} \quad (54)$$

Since $(x-1)^2 x \geq 0$ for all $x \in (0, 1)$. Observe that g is quadratic and that $g(0) = -\sum_{i=1}^n a_i^2 \leq 0$ and $g(1) = (b_i x - a_i)^2 \geq 0$. Therefore $g(x)$ can have at most one root in (c, d) . If $g(x)$ has no root in (c, d) , then $g(x)$ is either positive or negative. If $g(x)$ has a root $x_0 \in (c, d)$, then (54) holds. \square

Lemma 15. *Suppose there exists a change point $\nu \in (0, T)$ such that any other change point ν' within $(0, T)$ satisfies $\min\{|\nu' - \nu|\} \geq \Delta$. Then*

$$\max_{0 \leq t \leq T} \|\tilde{V}(t)\|^2 \geq \frac{\|V(\nu) - V(\nu + 1)\|^2 \Delta^2}{48T}.$$

Proof. Denote $\kappa = \|V(\nu) - V(\nu + 1)\|$.

Step 1. Let

$$\begin{aligned} I_1 &= \left\{ i : \left| \sum_{r=1}^{\nu-\Delta} V_i(r) \right| \geq \Delta |V_i(\nu) - V_i(\nu + 1)|/4 \right\}, \\ I_2 &= \left\{ i : \left| \sum_{r=1}^{\nu} V_i(r) \right| \geq \Delta |V_i(\nu) - V_i(\nu + 1)|/4 \right\}, \\ I_3 &= \left\{ i : \left| \sum_{r=1}^{\nu+\Delta} V_i(r) \right| \geq \Delta |V_i(\nu) - V_i(\nu + 1)|/4 \right\}, \end{aligned}$$

Then by Lemma 21, $I_1 \cup I_2 \cup I_3 = \{1, \dots, p\}$. We have

$$\sum_{l=1}^3 \left\{ \sum_{i \in I_l} (V_i(\nu) - V_i(\nu + 1))^2 \right\} \geq \sum_{i=1}^p (V_i(\nu) - V_i(\nu + 1))^2 = \kappa^2,$$

which implies that

$$\max_{l=1,2,3} \left\{ \sum_{i \in I_l} (V_i(\nu) - V_i(\nu + 1))^2 \right\} \geq \kappa^2/3.$$

Without loss of generality, suppose $\sum_{i \in I_1} (V_i(\nu) - V_i(\nu + 1))^2 \geq \kappa^2/3$. Then

$$\begin{aligned} \max_{1 \leq t \leq T} \|\tilde{V}(t)\|^2 &\geq \|\tilde{V}(\nu - \Delta)\|^2 = \frac{T}{(\nu - \Delta)(T - (\nu - \Delta))} \left\| \sum_{r=1}^{\nu-\Delta} V(r) \right\|^2 \\ &\geq \frac{1}{T} \sum_{i \in I_1} \left(\sum_{r=1}^{\nu-\Delta} V_i(r) \right)^2 \geq \frac{1}{T} \sum_{i \in I_1} (\Delta |V_i(\nu) - V_i(\nu + 1)|/4)^2 \\ &\geq \frac{\Delta^2}{48T} \kappa^2, \end{aligned}$$

where the first equality follows from (53) and the second last inequality follows from the definition of I_1 . \square

Lemma 16. *Let $[s, e] \subset [1, T]$ be any generic interval containing a change point ν satisfying*

$$\min\{\nu - s, e - \nu\} \geq c_1 \Delta.$$

If

$$\|\tilde{V}^{s,e}(\nu)\|^2 \geq \kappa^2 \Delta^2 (e - s)^{-1},$$

and there exists sufficient small $c_3 > 0$ such that

$$\max_{1 \leq t \leq T} \|\tilde{V}^{s,e}(t)\|^2 - \|\tilde{V}^{s,e}(\nu)\|^2 \leq c_3 \|\tilde{V}^{s,e}(\nu)\|^2 \Delta^2 (e-s)^{-2}, \quad (55)$$

then there exists an absolute constant $c, c_1 > 0$ such that $d \in [s, e]$ satisfying $|d - \nu| \leq c_1 \Delta / 16$, and

$$\|\tilde{V}^{s,e}(\nu)\|^2 - \|\tilde{V}^{s,e}(d)\|^2 > c \|\tilde{V}^{s,e}(\nu)\|^2 |\nu - d| \Delta (e-s)^{-2},$$

where $c = \min\{c_1, 1/2 - 2c_3\}$.

Proof. Denote $\tilde{V}^{s,e}(t) = \tilde{V}(t)$ and $l = d - \nu$. It suffices consider the case of $l \geq 0$, as the case of $l \leq 0$ follows by reversing the time series. Let $\nu' > \nu$ be the next change point. Then either $\nu' = e$ which means that ν is the last change point, or $\nu' < T$ which indicates that ν is not the last change point.

Case 1. Suppose $\nu' = T$. Let $i = \nu - s$ and $h = e - \nu$. For any $u \in \{1, \dots, p\}$, by Case 1 in Lemma 2.6 of Venkatraman (1992), it holds that

$$\tilde{V}_u(\nu) = \frac{a_u \sqrt{i+h}}{\sqrt{ih}}, \quad \tilde{V}_u(\nu+l) = \frac{h-l}{h} \frac{a_u \sqrt{i+h}}{\sqrt{(i+l)(h-l)}}.$$

Thus

$$\tilde{V}_u(\nu)^2 - \tilde{V}_u(\nu+l)^2 = l \frac{a_u^2(i+h)}{ih} \frac{h+i}{h(i+l)} = \frac{l(h+i)}{h(i+l)} \tilde{V}_u(\nu)^2.$$

So

$$\|\tilde{V}(\nu)\|^2 - \|\tilde{V}(\nu+l)\|^2 = \frac{l(h+i)}{h(i+l)} \|\tilde{V}(\nu)\|^2 \geq \frac{l(e-s)}{(e-s)^2} \|\tilde{V}(\nu)\|^2 \geq \frac{c_1 l \Delta}{(e-s)^2} \|\tilde{V}(\nu)\|^2.$$

Case 2. Suppose $\nu' < e$. Let $i = \nu - s$, $h = \Delta/2$ and $j = e - \nu - h$. Let $l \leq h/2$. For any $u \in \{1, \dots, p\}$, by Case 2 in Lemma 2.6 of Venkatraman (1992),

$$\tilde{V}_u(\nu) = \frac{a_u \sqrt{i+h}}{\sqrt{ih}}, \quad \tilde{V}_u(\nu+h) = \frac{(a_u + h\theta) \sqrt{i+j+h}}{\sqrt{(i+h)j}} \quad \text{and} \quad \tilde{V}_u(\nu+l) = \frac{(a_u + l\theta) \sqrt{i+j+h}}{\sqrt{(i+l)(j+h-l)}},$$

where θ is the solution of

$$\tilde{V}_u^2(\nu+h) - \tilde{V}_u^2(\nu) = \frac{(a_u + h\theta)^2(i+j+h)}{(i+h)j} - \frac{a_u^2(i+h)}{ih}.$$

Denote $B = \|\tilde{V}(\nu+h)\|^2 - \|\tilde{V}(\nu)\|^2$ and $B_u = \tilde{V}_u(\nu+h)^2 - \tilde{V}_u(\nu)^2$. Thus by (55),

$$B \leq c_3 \|\tilde{V}^{s,e}(\nu)\|_2^2 \Delta^2 (e-s)^{-2}. \quad (56)$$

Then by Lemma 17,

$$\begin{aligned} \|\tilde{V}(\nu)\|^2 - \|\tilde{V}(\nu+l)\|^2 &= \sum_{u=1}^p \left\{ \tilde{V}_u(\nu)^2 - \tilde{V}_u(\nu+l)^2 \right\} \\ &\geq \sum_{u=1}^p \left\{ \frac{\tilde{V}_u(\nu)^2 (hl - l^2)}{(i+l)(j+h-l)} - B_u \frac{l(i+h)j}{h(i+l)(j+h-l)} \right\} = \frac{\|\tilde{V}(\nu)\|_2^2 l(h-l)}{(i+l)(j+h-l)} - B \frac{l(i+h)j}{h(i+l)(j+h-l)} \\ &\geq \frac{\|\tilde{V}(\nu)\|_2^2 l \Delta}{2(e-s)^2} - 2B \frac{l}{\Delta} \geq (1/2 - 2c_3) \frac{\|\tilde{V}(\nu)\|_2^2 l \Delta}{(e-s)^2}, \end{aligned}$$

where the last inequality follows from (56). \square

Lemma 17. *Denote*

$$\Theta_\nu = \frac{a\sqrt{i+j+h}}{\sqrt{i(j+h)}}, \quad \Theta_{\nu+h} = \frac{(a+h\theta)\sqrt{i+j+h}}{\sqrt{(i+h)j}} \quad \text{and} \quad \Theta_{\nu+l} = \frac{(a+l\theta)\sqrt{i+j+h}}{\sqrt{(i+l)(j+h-l)}}.$$

Then

$$\Theta_\nu^2 - \Theta_{\nu+l}^2 \geq \frac{\Theta_\nu^2(hl - l^2)}{(i+l)(j+h-l)} - (\Theta_{\nu+h}^2 - \Theta_\nu^2) \frac{l(i+h)j}{h(i+l)(j+h-l)}.$$

Proof. Observe that

$$\begin{aligned} \Theta_\nu^2 - \Theta_{\nu+l}^2 &= \frac{a^2(i+j+h)}{i(j+h)} - \frac{(a+l\theta)^2(i+j+h)}{(i+l)(j+h-l)} \\ &= \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)} ((i+l)(j+h-l) - i(j+h)) \\ &\quad - \frac{(2l\theta a + l^2\theta^2)(i+j+h)}{(i+l)(j+h-l)} \\ &= \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)} (-il + lj + lh - l^2) - (2l\theta a + l^2\theta^2) \frac{(i+j+h)}{(i+l)(j+h-l)}. \end{aligned}$$

To bound the term $2l\theta a + l^2\theta^2$, let $b = \Theta_{\nu+h}^2 - \Theta_\nu^2$. Then

$$b = \frac{(a+h\theta)^2(i+j+h)}{(i+h)j} - \frac{a^2(i+j+h)}{i(j+h)}.$$

Therefore

$$\frac{bj(i+h)(j+h)}{i+j+h} = (a^2 + 2h\theta a + h^2\theta^2)i(j+h) - a^2(i+h)j,$$

which gives

$$2h\theta a + h^2\theta^2 = \frac{bj(i+h)}{i+j+h} + \frac{a^2(j-i)h}{i(j+h)}.$$

Therefore

$$\begin{aligned} 2l\theta a + l^2\theta^2 &\leq 2l\theta a + lh\theta^2 \\ &= \frac{l}{h}(2h\theta a + h^2\theta^2), \\ &= \frac{l}{h} \left(\frac{bj(i+h)}{i+j+h} + \frac{a^2(j-i)h}{i(j+h)} \right) \end{aligned}$$

which implies that

$$\begin{aligned}
\Theta_\nu^2 - \Theta_{\nu+l}^2 &= \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)}(-il+lj+lh-l^2) - (2l\theta a + l^2\theta^2) \frac{(i+j+h)}{(i+l)(j+h-l)} \\
&\leq \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)}(-il+lj+lh-l^2) \\
&\quad - \frac{l}{h} \left(\frac{bj(i+h)}{i+j+h} + \frac{a^2(j-i)h}{i(j+h)} \right) \frac{(i+j+h)}{(i+l)(j+h-l)} \\
&= \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)}(-il+lj+lh-l^2) \\
&\quad - \frac{lbj(i+h)}{h(i+l)(j+h-l)} - \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)}(j-i)l \\
&= \frac{a^2(i+j+h)}{i(j+h)(i+l)(j+h-l)}(lh-l^2) - b \frac{l(i+h)j}{h(i+l)(j+h-l)}
\end{aligned}$$

□

Lemma 18. *Suppose $[s, e]$ contains one change point η_k , then*

$$\|\tilde{V}^{s,e}(t)\|^2 = \begin{cases} \frac{t-s}{(e-s)(e-t)}(e-\eta_k)^2\|V(\eta) - V(\eta+1)\|^2, & t \leq \eta_k, \\ \frac{e-t}{(e-s)(t-s)}(\eta_k-s)^2\|V(\eta) - V(\eta+1)\|^2, & t \geq \eta_k. \end{cases}$$

Proof. This is a straightforward result from the definitions. □

Lemma 19. *Let η_1 be the first change point in $\{1, \dots, T\}$. Then for any $1 \leq t \leq \eta_1$,*

$$\|\tilde{V}^{1,T}(t)\|^2 = \frac{t(T-\eta_1)}{\eta_1(T-t)}\|\tilde{V}^{1,T}(\eta_1)\|^2.$$

Proof. This is a direct consequence of Lemma 22. □

Lemma 20. *Let $[s, e]$ contain two or more change points such that*

$$\eta_{r-1} \leq s \leq \eta_r \leq \dots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 1.$$

If $\eta_r - s \leq c\Delta$ for some $c \leq 1/4$ and $\eta_{r+1} - \eta_r \geq \Delta$, then

$$\|\tilde{V}^{s,e}(\eta_r)\|^2 \leq 2c\|\tilde{V}^{s,e}(\eta_{r+1})\|^2 + 4\kappa_r^2(\eta_r - s)$$

Proof. This follows a similar calculation as in Lemma 23. □

D.2 1D CUSUM

Assumption 10. *Let $\{f(t)\}_{t=1}^T \subset \mathbb{R}$. Assume there exists $\{\nu_m\}_{m=0}^M \subset \{1, \dots, T\}$ such that*

$$f(\nu_m + 1) = \dots = f(\nu_{m+1}) \quad \text{for all } 0 \leq m \leq M-1$$

and that $|f(\nu_m) - f(\nu_{m+1})| = \kappa_m \geq \kappa$.

For the same reasons as we described after Assumption 9, in this subsection we use a self-contained notation system, and one can interpret $\kappa = \kappa_0 n \rho$ as we used in Assumption 1.

Lemma 21. *Suppose ν_m is a change point of $\{f(t)\}_{t=1}^T$ such that $\min_{m' \neq m} \{\nu_m - \nu_{m'}\} \geq \Delta$. Then*

$$\max \left\{ \left| \sum_{r=1}^{\nu_m - \Delta} f(r) \right|, \left| \sum_{r=1}^{\nu_m} f(r) \right|, \left| \sum_{r=1}^{\nu_m + \Delta} f(r) \right| \right\} \geq \Delta |f(\nu_m) - f(\nu_m + 1)|/4. \quad (57)$$

Proof. For simplicity denote $\nu_m = \nu$. Observe that

$$\max\{|f(\nu)|, |f(\nu + 1)|\} \geq |f(\nu) - f(\nu + 1)|/2.$$

Thus

$$\max \left\{ \left| \sum_{r=\nu-\Delta}^{\nu} f(r) \right|, \left| \sum_{r=\nu+1}^{\nu+\Delta} f(r) \right| \right\} \geq \Delta |f(\nu) - f(\nu + 1)|/2. \quad (58)$$

Since

$$\left| \sum_{r=\nu-\Delta}^{\nu} f(r) \right| \leq \left| \sum_{r=1}^{\nu-\Delta} f(r) \right| + \left| \sum_{r=1}^{\nu} f(r) \right|, \quad \left| \sum_{r=\nu+1}^{\nu+\Delta} f(r) \right| \leq \left| \sum_{r=1}^{\nu} f(r) \right| + \left| \sum_{r=1}^{\nu+\Delta} f(r) \right|, \quad (59)$$

(58) and (59) directly imply (57). \square

Lemma 22. *Let η_1 be the first change point in $\{1, \dots, T\}$. Then for any $1 \leq t \leq \eta_1$,*

$$\tilde{f}_t^{1,T} = \sqrt{\frac{t(T - \eta_1)}{\eta_1(T - t)}} \tilde{f}_{\eta_1}^{1,T}.$$

Proof. Without loss of generality assume $\sum_{t=1}^T f_t = 0$. Thus $\eta_1 f_1 = \sum_{t=1}^{\eta_1} f_t = -\sum_{t=\eta_1+1}^T f_t$. As a result, for any $1 \leq t \leq \eta_1$,

$$\begin{aligned} \tilde{f}_t^{1,T} &= \sqrt{\frac{T-t}{Tt}} \sum_{i=1}^t f_i - \sqrt{\frac{t}{T(T-t)}} \sum_{i=t+1}^T f_i \\ &= \sqrt{\frac{T-t}{Tt}} t f_1 - \sqrt{\frac{t}{T(T-t)}} \left((\eta_1 - t) f_1 + \sum_{i=\eta_1+1}^T f_i \right) \\ &= \sqrt{\frac{T-t}{Tt}} t f_1 - \sqrt{\frac{t}{T(T-t)}} \{(\eta_1 - t) f_1 - \eta_1 f_1\} \\ &= \frac{(T-t)\sqrt{t} + t\sqrt{t}}{\sqrt{T(T-t)}} f_1 = \sqrt{\frac{Tt}{T-t}} f_1. \end{aligned}$$

\square

Remark 17. *If there exists $b \in [1, \eta_1]$ such that $\tilde{f}_b^{1,T} > 0$, then by Lemma 22, $\tilde{f}_{\eta_1}^{1,T} > 0$. Since for $t \in [1, \eta_1]$, $\sqrt{\frac{t(T-\eta_1)}{\eta_1(T-t)}}$ is an increasing function of t , this also implies $\tilde{f}_t^{1,T} > 0$ is increasing within $[1, \eta_1]$.*

Lemma 23. Let $[s, e]$ contain two or more change points such that

$$\eta_{r-1} \leq s \leq \eta_r \leq \dots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 1.$$

If $\eta_r - s \leq c_1^2 \Delta$ for some $c_1 \leq 1/4$ and $\eta_{r+1} - \eta_r \geq \Delta$, then

$$|\tilde{f}_{\eta_r}^{s,e}| \leq c_1 |\tilde{f}_{\eta_{r+1}}^{s,e}| + 2\kappa_r \sqrt{\eta_r - s}.$$

Proof. Consider the sequence $\{g_t\}_{t=s+1}^e$ be such that

$$g_t = \begin{cases} f_{\eta_{r+1}}, & \text{if } s+1 \leq t \leq \eta_r, \\ f_t, & \text{if } \eta_r+1 \leq t \leq e. \end{cases}$$

For any $t \geq \eta_r + 1$,

$$\begin{aligned} & \tilde{f}_t^{s,e} - \tilde{g}_t^{s,e} \\ &= \sqrt{\frac{e-t}{(e-s)(t-s)}} \left(\sum_{i=s+1}^{\eta_r} f_{\eta_r} + \sum_{i=\eta_r+1}^t f_{\eta_{r+1}} - \sum_{i=s+1}^{\eta_r} g_{\eta_r} - \sum_{i=\eta_r+1}^t g_{\eta_{r+1}} \right) \\ & \quad - \sqrt{\frac{t-s}{(e-s)(e-t)}} \left(\sum_{i=t+1}^e f_t - \sum_{i=t+1}^e g_t \right) \\ &= \sqrt{\frac{e-t}{(e-s)(t-s)}} (\eta_r - s)(f_{\eta_{r+1}} - f_{\eta_r}) \leq \sqrt{\eta_r - s} \kappa_r. \end{aligned}$$

Thus

$$\begin{aligned} |\tilde{f}_{\eta_r}^{s,e}| &\leq |\tilde{g}_{\eta_r}^{s,e}| + \sqrt{\eta_r - s} \kappa_r \\ &\leq \sqrt{\frac{(\eta_r - s)(e - \eta_{r+1})}{(\eta_{r+1} - s)(e - \eta_r)}} |\tilde{g}_{\eta_{r+1}}^{s,e}| + \sqrt{\eta_r - s} \kappa_r \\ &\leq \sqrt{\frac{c_1^2 \Delta}{\Delta}} |\tilde{g}_{\eta_{r+1}}^{s,e}| + \sqrt{\eta_r - s} \kappa_r \\ &\leq c_1 |\tilde{f}_{\eta_{r+1}}^{s,e}| + 2\sqrt{\eta_r - s} \kappa_r, \end{aligned}$$

where the first inequality follows from Lemma 22 and the observation that the first change point of g_t in $[s, e]$ is η_{r+1} . \square

E Additional lemmas

Lemma 24. Suppose $x > 0$ and that $x^2 + bx - c \geq 0$ where $b, c > 0$ and that

$$b \leq \sqrt{c}/4.$$

Then $x \geq 7\sqrt{c}/8$.

Proof. We have either $x \geq \frac{-b+\sqrt{b^2+4c}}{2}$ or $x \leq \frac{-b-\sqrt{b^2+4c}}{2}$. Since $x, b, c > 0$ and $b \leq \sqrt{c}/4$, we have

$$x \geq \frac{-b + \sqrt{b^2 + 4c}}{2} \geq 7\sqrt{c}/8.$$

□

Let $\{\alpha_m\}_{m=1}^M, \{\beta_m\}_{m=1}^M$ be two sequences independently selected at random from $\{1, \dots, T\}$, and

$$\mathcal{M} = \bigcap_{k=1}^K \{\alpha_m \in \mathcal{S}_k, \beta_m \in \mathcal{E}_k, \text{ for some } m \in \{1, \dots, M\}\}, \quad (60)$$

where $\mathcal{S}_k = [\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\mathcal{E}_k = [\eta_k + \Delta/2, \eta_k + 3\Delta/4]$, $k = 1, \dots, K$. In the lemma below, we give a lower bound on the probability of \mathcal{M} .

Lemma 25. *For the event \mathcal{M} defined in (60), we have*

$$\mathbb{P}(\mathcal{M}) \geq 1 - \exp \left\{ \log \left(\frac{T}{\Delta} \right) - \frac{M\Delta^2}{16T^2} \right\}.$$

Proof. Since the number of change points are bounded by T/Δ ,

$$\mathbb{P}(\mathcal{M}^c) \leq \sum_{k=1}^K \prod_{m=1}^M \{1 - \mathbb{P}(\alpha_m \in \mathcal{S}_k, \beta_m \in \mathcal{E}_k)\} \leq K(1 - \Delta^2/(16T^2))^M \leq (T/\Delta)(1 - \Delta^2/(16T^2))^M.$$

□