Submitted to the Annals of Statistics
arXiv: arXiv:0000

OPTIMAL CHANGE POINT DETECTION AND LOCALIZATION IN SPARSE DYNAMIC NETWORKS

By Daren Wang*, Yi Yu† and Alessandro Rinaldo‡

*University of Chicago *, University of Warwick † and Carnegie Mellon University ‡*

We study the problem of change point localization in dynamic networks models. We assume that we observe a sequence of independent adjacency matrices of the same size, each corresponding to a realization of an unknown inhomogeneous Bernoulli model. The underlying distribution of the adjacency matrices are piecewise constant, and may change over a subset of the time points, called change points. We are concerned with recovering the unknown number and positions of the change points. In our model setting we allow 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 a region of impossibility 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 algorithm for network change point localization, called Network Binary Segmentation, that 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. Under appropriate conditions, Local Refinement guarantees a minimax optimal rate for network change point localization while remaining computationally feasible.

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, 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 collection of results of computational, methodological and theoretical nature exist.

In contrast to the basic premise of the static network modeling framework, many modern 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 1.3 below for some literature review).

Keywords and phrases: Change point detection; Low-rank networks; Stochastic block model; Minimax optimality.
MSC 2010 subject classifications: Primary 62M10; secondary 91B84

In this article we are concerned with a discrete time network dynamic setting in which the set of nodes is fixed but the edge probabilities are time-varying. We assume that we observe a sequence of T independent and possibly sparse networks of constant size whose distributions may change at $K < T$ unknown time points, or change points. We impose minimal restrictions on the number and locations of the possible 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. Our goal is to detect whether any such change has taken place, and to accurately estimate the time of the corresponding change point. Importantly, we are not interested in estimating the underlying data-generating distributions. As our analysis will reveal, although we only consider a fairly straightforward form of network dynamics, the associated inference problem is rather subtle and far from trivial. Furthermore, if one is interested in the underlying distributions, then static network estimation methods can be applied to the sample means of the adjacency matrices between two consecutive change point estimators.

1.1. *Problem setup.* 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 $\{1, \ldots, n\}$ 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 $\mathbb{E}(A_{ij}) = \Theta_{ij}$.

Definition 1 covers a wide range of models for undirected networks, including the Erdős–Rényi random graph (Erdős and Rényi, 1959), the stochastic block model (Holland et al., 1983), the degree corrected block model (Karrer and Newman, 2011) and the 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, discussions on the possible relaxations on the independence and Bernoulli assumptions can be found in Section 5.

We further assume that the probability distributions of the networks change only over an unknown subset of the time points, called change points. We formalize our setting below.

ASSUMPTION 1 (Change point dynamic network model). *Let $\{A(t)\}_{t=1}^T$ be a sequence of $n \times n$ adjacency matrices of independent inhomogeneous Bernoulli networks with means $\{\Theta(t)\}_{t=1}^T$ satisfying the following properties.*

1. *The sparsity parameter*

$$\rho := \max_{t=1,\ldots,T} \|\Theta(t)\|_\infty \tag{1}$$

is such that

$$\rho n \geq \log(n), \tag{2}$$

where $\|\cdot\|_\infty$ denotes the entrywise maximum norm of a matrix.

2. *There exists a sequence* $(\eta_0, \ldots, \eta_{K+1})$ *of time points, called change points, such that* $1 = \eta_0 < \eta_1 < \ldots < \eta_K \leq T < \eta_{K+1} = T + 1$ *and, for* $t = 2, \ldots, T$,

$$\Theta(t) \neq \Theta(t-1) \quad \text{if and only if} \quad t \in \{\eta_1, \ldots, \eta_K\}.$$

We let

$$\Delta := \min_{k=1,\ldots,K+1} \{\eta_k - \eta_{k-1}\} \leq T$$

be the minimal spacing between two consecutive change points and set

$$\kappa_0 := \frac{\min_{k=1,\ldots,K} \|\Theta(\eta_k) - \Theta(\eta_k - 1)\|_{\mathrm{F}}}{n\rho} \in (0, 1], \tag{3}$$

to be the normalized magnitude of the smallest changes in the data generating distribution, where $\|\cdot\|_{\mathrm{F}}$ *denotes the Frobenius norm.*

A few comments on our modeling assumptions are in order. First, we rely on the Frobenius norm of the difference between two consecutive expected adjacency matrices at a change point to quantify the magnitude of the corresponding distributional change. This is a fairly general metric, able to capture both "dense" changes caused by small variations spread across many edge probabilities as well by "sparse" changes due to large difference only along few coordinates. Next, the quantity $\kappa_0 \in (0, 1]$ appearing in (3) measures the size of the smallest distributional change in the model in a manner that is independent of the choice of the other parameters. Indeed, the terms $\|\Theta(\eta_k) - \Theta(\eta_k - 1)\|_{\mathrm{F}}$'s depend on both the sparsity parameter ρ and the size of the networks n. To avoid such confounding, and using the fact that $\max_k \|\Theta(\eta_k) - \Theta(\eta_k - 1)\|_{\mathrm{F}} \leq n\rho$, setting κ_0 as in (3) yields a scale-free parameter in $(0, 1]$ that is independent of both ρ and n.

The model described above is defined by the parameters Δ, κ_0, n and ρ. We adopt a high-dimensional framework whereby T grows unbounded and all the defining parameters are allowed to change as a function of T. The number of change points K also may change with T, but since $K \leq \frac{T}{\Delta}$ by definition, we will capture any dependence on K only through Δ. We refer to any relationship among all the model parameters $(\Delta, \kappa_0, n, \rho)$ and T that holds as $T \to \infty$ as a **scaling**. For ease of readability we will not make the dependence on T explicit in our notation.

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 $(A(1), \ldots, A(T))$ of adjacency matrices satisfying the above assumptions. More precisely, for a given scaling of the model parameters, we aim to construct an estimator of $(\eta_1, \ldots, \eta_K)$ of the form

$$(A(1), \ldots, A(T)) \mapsto (\hat{\eta}_1, \ldots, \hat{\eta}_{\hat{K}}) \subset (2, \ldots, T) \tag{4}$$

and with $\hat{\eta}_1 < \hat{\eta}_2 < \ldots < \hat{\eta}_{\hat{K}}$ satisfying the following notion of localization consistency.

Definition 2 (Consistent localization). *A change point estimator of the form* (4) *is consistent if, with probability tending to* 1 *as* $T \to \infty$,

$$\widehat{K} = K \quad \text{and} \quad \max_{k=1,\ldots,K} |\hat{\eta}_k - \eta_k| \leq \epsilon, \tag{5}$$

where $\epsilon = \epsilon(T, \Delta, \kappa_0, \rho, n)$ *is such that*

$$\frac{\epsilon}{\Delta} \to 0. \tag{6}$$

The term ϵ *is called the* **localization error** *of the estimator and the sequence* $\left\{\frac{\epsilon}{\Delta}\right\}$ *the* **localization rate**.

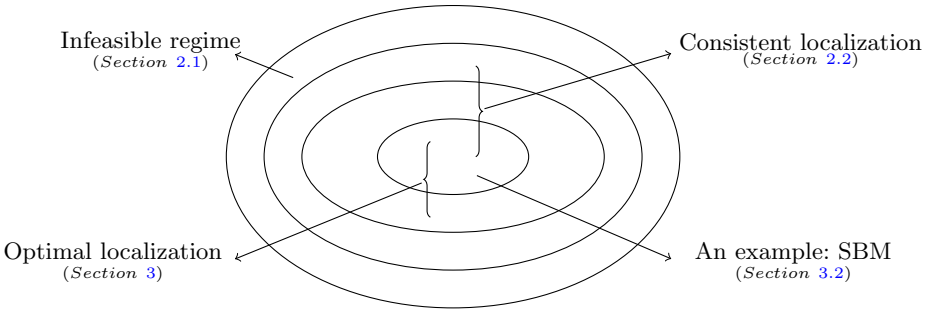

FIG 1. *Reading guide.*

Thus, we will deem a change point estimator consistent if, with high probability as the number of time points grows, its localization error is a vanishing fraction of the minimal distance between consecutive change points. The limiting probability (in T) of the event in (5) and the value of the localization error ϵ depend on the choice of the scaling. For instance, it is intuitively clear that scalings in which all parameters decrease with T will lead to a sequence of change point problems of increasing difficulty.

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 efficient 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 the quantity

$$\sqrt{\rho}\kappa_0, \tag{7}$$

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 standard deviation of the entries of A.

1.2. *List of contributions.* The main theoretical contribution of the paper is the identification of three regions inside the space of model parameters corresponding to different types of scaling or regimes: (i) an impossibility regimes, where no change point localization algorithm is guaranteed to be consistent (see Section 2.1); (ii) a feasibility regime, described in Assumption 2), for which we demonstrate the existence of a polynomial-time, consistent change point estimator (see Section 2.2); and (iii) a subset of (ii), described in Assumption 3, for which we further show that change point localization can be achieved at a nearly minimax optimal rate, again using a polynomial-time algorithm (see Section 3). The partition of scaling regimes, represented pictorially in Figure 1, is relatively sharp, in the sense that regimes (i) and (ii) are only off by any diverging factor in T.

To be specific, 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

$$\rho\kappa_0^2 \lesssim \frac{\log^2(T)}{n\Delta}, \tag{8}$$

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 $\xi > 0$[1],

$$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta n}, \tag{9}$$

we demonstrate a computationally-efficient procedure, called Network Binary Segmentation (NBS) (see Algorithm 1 below) that is provably consistent. The procedure combines sample splitting 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 is of independent interest.

The NBS is consistent under nearly the weakest possible conditions, since it leads to a vanishing localization rate under the scaling (9) which, save for a $\log^{2\xi}(T)$ term, matches the phase transition boundary in (8). Remarkably, no structural assumptions on the distributions of the networks themselves are used. Indeed, in deriving the bound (8), we construct a worst-case class of distributions consisting of dynamic networks satisfying stochastic block models. This reveals that, under the scaling in which the 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. The dynamic version optimality is shown in Pensky (2016).

- 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 $\Theta(\eta_k) - \Theta(\eta_k - 1)$, $k = 1, \ldots, K$, be a matrix of rank at most $r \leq n$, an additional parameter that 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 stronger scaling

$$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta}\frac{r}{n}, \tag{10}$$

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

$$\epsilon \lesssim \frac{\log^2(T)}{\kappa_0^2 n^2 \rho}. \tag{11}$$

The proposed procedure takes as input the estimates of the change point locations from any reasonable (not necessarily consistent nor optimal) estimator, including the NBS, and further improves their accuracy to deliver the above localization rate. At its core, the LR algorithm

[1] In fact, ξ is allowed to be zero if n diverges with T. More generally, in that case, we may replace the term $\log^{\xi}(T)$ with any other quantity one diverging in T.

relies on exactly K (this, we recall, being the 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 (11), is in fact nearly minimax rate-optimal, aside for the $\log^2(T)$ term. Interestingly, the expression of the rate (11) 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). More discussions on the gap between the scalings (9) and (10), and on the comparisons with Wang et al. (2017) are provided later in the paper.

- 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 common 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, the diagonal entries of the expected adjacency matrix are set to be zero, which in general would prevent the low-rank assumption. In fact, this complication is often ignored in the existing literature. In this case, we show that with a very mild additional assumption, we are still able to recover the nearly optimal localization rate (11). In our analysis we borrow tools and ideas from several areas, including change point detection, network analysis and graphon estimation.

The rest of this paper is organized as follows. Section 1.3 summarize some of the related literature. In Section 2, we first identify the scalings for which consistent localization is impossible and then present the NBS change point estimator, which we show to be consistent under almost any scaling outside this impossibility regime. In Section 3 we develop the more sophisticated algorithm LR, which we then show to be almost minimax rate-optimal under an additional low-rank assumption. We further demonstrate in Section 3.2 how our procedure is applicable to the dynamic stochastic block model. Section 4 presents few illustrative simulations that verify the effectiveness of our procedures. Finally, we conclude with more discussions including potential future work directions in Section 5. The proofs of our results are presented in the the appendix and supplementary material.

1.3. *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 a 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 (2019) 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. Bhattacharjee et al. (2018) dealt with a change point detection problem in a one-change-point stochastic block model sequences and focused on recovering underlying models, which resulted in a cost of sub-optimal change point 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. Zhao et al. (2019) provided a two-step algorithm, which first estimates the networks and then uses a moving window to detect change points. The results thereof are consistent yet optimal. Another consistent yet optimal result on network change point detection problems is derived in Chapter 5 in Mukherjee (2018).

1.4. *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 ith row and jth column of A. Let $\kappa_i(A)$ be the ith eigenvalue of A with ordering $|\kappa_1(A)| \geq |\kappa_2(A)| \geq \ldots \geq |\kappa_n(A)|$, 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 ith 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 $f(n) \gtrsim g(n)$.

We now recall the definition of cumulative sum (CUSUM) statistic (Page, 1954).

DEFINITION 3 (CUSUM statistics). *For a collection of any type of data $\{X(t)\}_{t=1}^T$, any pair of time points $(s,e) \subset \{0,\ldots,T\}$ with $s < e-1$, and any time point $t = s+1,\ldots,e-1$, let the CUSUM statistics be*

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

Since the CUSUM statistic is linear in its arguments, we have that, for any $0 \leq s < t < e \leq T$,

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

2. Consistent localization. In this section we study 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 a vanishing localization rate from the ones for which no estimator is consistent. To be precise, when we say that consistent localization is impossible for a given scaling, we mean it in a minimax sense that there exists *some* change point model satisfying Assumption 1 for which no estimator of the change points is consistent.

2.1. *The impossibility regime.* Below we establish an information-theoretic lower bound, which demonstrates that, if

$$\rho\kappa_0^2 \lesssim \frac{\log^2(T)}{n\Delta}, \tag{12}$$

then no consistent estimator of the change points exists. 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 and therefore low rank) does not necessarily lead to a consistent estimator under the scaling in (12). The details are given in Section S.1.

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

$$\mathcal{P} = \left\{ P^T_{\kappa_0,\Delta,n,\rho} : \Delta = \min\left\{ \left\lfloor \frac{4\zeta \log(T)}{n\rho\kappa_0^2} \right\rfloor, \lfloor T/4 \rfloor \right\}, \rho \le 1/2, \kappa_0 \le 1, 0 < \zeta < 1/2 \right\}.$$

Then there exists a $T(\zeta)$, which depends on ζ, such that for all $T \ge T(\zeta)$,

$$\inf_{\widehat{\eta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P(H(\widehat{\eta}, \eta(P))) \ge \Delta/2,$$

where the infimum is over all estimators $\widehat{\eta} = \{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ of the change point locations, $\eta(P)$ is the set of the change points of $P \in \mathcal{P}$ and $H(\cdot,\cdot)$ denotes the Hausdorff distance.

2.2. *Network Binary Segmentation.* In our next result, we show that parameter scalings of the form given in (12) 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 the following signal-to-noise condition is met.

ASSUMPTION 2. *For a constant $C_\alpha > 0$ and any $\xi > 0$, we have that*

$$\kappa_0\sqrt{\rho} \ge C_\alpha \sqrt{\frac{1}{n\Delta}} \log^{1+\xi}(T). \tag{13}$$

Recalling (12), our results cover all parameter scalings, aside from a $\log^\xi(T)$ term, where $\xi > 0$ can be arbitrarily small. When the size of the networks n diverges with T, arguably a very natural asymptotic regime, one can take ξ in Assumption 2 to be zero. In fact, in this case the signal-to-noise ratio condition (13) can be weakened to be of the form $\kappa_0\sqrt{\rho} \ge C_\alpha \sqrt{\frac{1}{n\Delta}} \log(T) e_T$, for any sequence of positive numbers $\{e_T\}_{T=1,2,\ldots}$ diverging to infinity arbitrarily slowly.

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

- Assume a non-sparse setting (i.e. $\rho \asymp 1$). If the minimal spacing Δ is of order $\log^{2+2\xi}(T)$, then Assumption 2 demands that $n\kappa_0 \succeq n^{1/2}$. 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 chosen to be $\log(n)/n$ as in (2), 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

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

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

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

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

INPUT: Two independent samples $\{A(t)\}_{t=1}^T, \{B(t)\}_{t=1}^T \in \mathbb{R}^{n\times n}$, τ_1.
  for $m = 1,\ldots,M$ **do**
    $[s'_m, e'_m] \leftarrow [s,e] \cap [\alpha_m, \beta_m]$
    $(s_m, e_m) \leftarrow [s'_m + 64^{-1}(e'_m - s'_m), e'_m - 64^{-1}(e'_m - s'_m)]$
    if $e_m - s_m \geq 1$ **then**
      $b_m \leftarrow \arg\max_{t=s_m+1,\ldots,e_m-1}(\widetilde{A}^{s_m,e_m}(t), \widetilde{B}^{s_m,e_m}(t))$
      $a_m \leftarrow (\widetilde{A}^{s_m,e_m}(b_m), \widetilde{B}^{s_m,e_m}(b_m))$
    else
      $a_m \leftarrow -1$
    end if
  end for
  $m^* \leftarrow \arg\max_{m=1,\ldots,M} a_m$
  if $a_{m^*} > \tau_1$ **then**
    add b_{m^*} to the set of estimated change points
    NBS$((s, b_{m*}), \{(\alpha_m,\beta_m)\}_{m=1}^M, \tau_1)$
    NBS$((b_{m*}+1, e), \{(\alpha_m,\beta_m)\}_{m=1}^M, \tau_1)$
  end if
OUTPUT: The set of estimated change points.

The NBS is a novel algorithm that builds on the traditional machinery developed for the univariate mean change point detection problem. The cornerstones of the NBS are the CUSUM statistics $\widetilde{A}^{s_m,e_m}(t)$ and $\widetilde{B}^{s_m,e_m}(t)$ (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. Vostrikova, 1981; Fryzlewicz, 2014; Wang and Samworth, 2018), the 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 $X^2 = X$. As a result, $\|\widetilde{A}^{s_m,e_m}(t)\|_{\mathrm{F}}^2$ cannot serve as a good estimator of $\|\widetilde{\Theta}^{s_m,e_m}(t)\|_{\mathrm{F}}^2$. In practice, these two independent samples can be acquired by splitting the data into, say, odd and even time points. In addition, every random interval (s'_m, e'_m) provided to the algorithm is shrunk by a constant fraction of its original length. This is done in order to avoid false positives around newly-found change points, a correction usually performed in WBS-style algorithm: see, e.g., the parameter δ used in Algorithm 3 in Wang et al. (2017) and the parameter β used in Algorithm 4 Wang and Samworth (2018). Note that in our paper, however, the amount of shrinking does not depend on unknown quantities.

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 3) –

and does not rely on any network or graphon estimation procedures, which are computationally costly. Though the NBS is not estimating 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. In our next result we show that the NBS yields in fact a consistent estimator the change points.

THEOREM 1. *Assume the model described in Assumption 1 and the condition of Assumption 2. There exist absolute positive constants $C_R > 3/2$, C_β, $c_2 \in (0,1)$, c, c_T and C_1 such that, letting $\{(\alpha_m,\ \beta_m)\}_{m=1}^M \subset (0,T)$ be a collection of random intervals whose end points are drawn independently and uniformly from $\{1,\dots,T\}$ and such that*

$$\max_{m=1,\dots,M}(\beta_m - \alpha_m) \leq C_R\Delta, \tag{14}$$

and

$$C_\beta \rho n \log^{3/2}(T) < \tau < c_2\kappa_0^2 n^2\rho^2\Delta \tag{15}$$

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

$$\mathbb{P}\Big\{\widehat{K} = K;\quad \max_{k=1,\dots,K}|\eta_k - \hat{\eta}_k| \leq \epsilon\Big\} \geq 1 - \exp\left(\log\left(\frac{T}{\Delta}\right) - M\frac{\Delta}{4C_RT}\right) - (6T^{3-c_T} + 2T^{3-c}), \tag{16}$$

where

$$\epsilon = C_1\log(T)\left(\frac{\sqrt{\Delta}}{\kappa_0 n\rho} + \frac{\sqrt{\log(T)}}{\kappa_0^2 n\rho}\right). \tag{17}$$

The constants in the theorem statement and their hierarchy of dependencies can be explicitly tracked in the proof; in particular, we require that the signal-to-noise ratio constant C_α in Assumption 2 to be sufficiently large. See the remark at the beginning of the proof of Theorem 1 in Appendix A.

To see how Theorem 1 implies that the NBS is consistent according to Definition 2, 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 Assumptions 2 and 1, respectively, into the bound (17) on the localization error to get that

$$\begin{aligned}\frac{\epsilon}{\Delta} &= C_1\log(T)\left(\frac{\sqrt{\Delta}}{\kappa_0 n\rho} + \frac{\sqrt{\log(T)}}{\kappa_0^2 n\rho}\right)\frac{1}{\Delta} \\ &\leq C_1\left(\frac{1}{C_\alpha\sqrt{\log(n)}\log^{\xi}(T)} + \frac{1}{C_\alpha^2\log^{1/2+2\xi}(T)}\right) \to 0,\end{aligned} \tag{18}$$

as $T \to \infty$ (with all the remaining parameters also possibly changing in accordance to any scaling compatible with Assumption 2). The last expression also shows that, if n diverges as T grows unbounded, the parameter ξ can be taken to be 0 in Assumption 2 and consistent localization is still guaranteed. More interestingly, (18) continues to hold also when $n \asymp 1$, so that consistent

localization is possible even when the number of nodes remains bounded. Of course, this is in striking contrast with the problem of consistent estimation of the edge probabilities – or, more generally, of an underlying graphon – which requires $n \to \infty$.

We remark that, while Theorem 1 shows that the NBS algorithm is consistent, we make no claim as to whether the localization rate is optimal. In the next section we will propose a two-step algorithm for change point localization that is not only consistent but nearly minimax rate-optimal under more favorable scalings on the parameters than the ones considered in Theorem 1.

We conclude this section with few technical remarks on the assumptions of Theorem 1. In order for the NBS algorithm to be consistent, the threshold parameter τ needs to belong to an appropriate range: see (15). Such choice essentially guarantees that τ is both large enough to avoid false positives and small enough to never miss any true change points, both events occurring with high probability. Next, the condition in (14) requires that each of the random intervals fed to the NBS algorithm is not too large, compared to the minimal spacing parameter Δ. Without assuming (14), and using the trivial bound $C_R \leq T/\Delta$, it can be shown that the NBS will achieve a larger localization error of

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

under the scaling

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

which is stronger than the one in Assumption 2. Assumption (14) about the length of the random time intervals used as input to the algorithms is of somewhat technical nature, but it appears necessary to yield the localization error in (17). Indeed, this condition, or analogous ones requiring some knowledge of Δ, are commonly assumed in the literature for change point localization to derive theoretical guarantees for WBS-style methods: see, e.g., Fryzlewicz (2014), Wang and Samworth (2018), Wang et al. (2018b), Baranowski et al. (2019), Anastasiou and Fryzlewicz (2019) and Eichinger et al. (2018). Finally, the parameter M, the number of random intervals used by the procedure, affects the results through the probability lower bound in (16). In order to guarantee that the probability tends to 1, one needs that

$$M \gtrsim \frac{T}{\Delta} \log\left(\frac{T}{\Delta} \right).$$

3. Optimal localization. In the previous section we saw how the NBS algorithm 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 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 procedure that first applies the NBS and then refines the resulting estimators of the locations of the change points, will achieve a minimax optimal localization rate. The additional step beyond the NBS is named local refinement (LR) and is detailed in Algorithm 3.

The LR algorithm takes as input two identically distributed sequences of networks fulfilling Assumption 1 (obtained for instance by sample splitting), along with a sequence $\{\nu_k\}_{k=1}^K$ of initial change point estimates that are sufficiently close to the locations of the true change points, in way made precise in (20) below. In particular, this preliminary estimates may be computed on the same data. The procedure then inspects all the triplets of consecutive change point estimators one at a time (with the time points 1 and $T+1$ as two dummy change points, for notational consistency).

Algorithm 2 Universal Singular Value Thresholding. 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 ith eigen-pair of A, with $|\kappa_1(A)| \geq \cdots |\kappa_n(A)|$
$A' \leftarrow \sum_{i:|\kappa_i(A)|\geq\tau_2} \kappa_i(A) v_i v_i^\top$
USVT$(A, \tau_2, \tau_3) \leftarrow (A''_{ij})$ with

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

OUTPUT: USVT(A, τ_2, τ_3).

Algorithm 3 Local Refinement

INPUT: $\{A(t)\}_{t=1}^T, \{B(t)\}_{t=1}^T \in \mathbb{R}^{n\times n}$, τ_2, τ_3, $\{\nu_k\}_{k=1}^K \subset \{1, \ldots, T-1\}$, $\nu_0 = 1$, $\nu_{K+1} = T+1$.
for $k = 1, \ldots, K$ **do**
$[s, e] \leftarrow [2^{-1}(\nu_{k-1} + \nu_k),\ 2^{-1}(\nu_k + \nu_{k+1})]$
$\widetilde{\Delta}_k \leftarrow \sqrt{\frac{(e-\nu_k)(\nu_k - s)}{e-s}}$
$\widehat{\Theta}_k \leftarrow \text{USVT}(\widetilde{B}^{s,e}(\nu_k), \tau_2, \tau_3\widetilde{\Delta}_k)$
$b_k \leftarrow \text{argmax}_{s\leq t\leq e}(\widetilde{A}^{s,e}(t), \widetilde{\Theta}_k)$
end for
OUTPUT: $\{b_k\}_{k=1}^K$.

For each such triplet, the 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. From a computational standpoint, Algorithm 3 is parallelizable in the sense that we can deal with each $k \in \{1, \ldots, K\}$ separately.

The signal-to-noise ratio conditions under which the LR improves upon the NBS are stronger than 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, 2018). We formalize them next.

Assumption 3. *Let $\{\Theta(t)\}_{t=1}^T$ be defined as in Assumption 1. For some $0 < r \leq n$,*

$$\max_{k=1,\ldots,K} \text{rank}\left(\Theta(\eta_k) - \Theta(\eta_k - 1)\right) \leq r.$$

Furthermore, for a constant $C_\alpha > 0$ and any $\xi > 0$,

$$\kappa_0\sqrt{\rho} \geq C_\alpha \frac{\log^{1+\xi}(T)}{\sqrt{\Delta}}\sqrt{\frac{r}{n}}. \tag{19}$$

The fixed quantity $\xi > 0$ in in the previous assumtion is required only for the case of $r \asymp n \asymp 1$ and can be set to zero in all other scenarios. The parameter r controlling the maximal rank of the difference of consecutive expected adjacency matrices is, like all the other parameters, also allowed to change with T. The first condition in Assumption 3 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 is of low rank. Using the fact that, for any matrices $A, B \in \mathbb{R}^{n\times n}$ of rank r_1 and r_2 respectively, it holds that

$$\text{rank}(A - B) = \min\{r_1 + r_2,\ n\},$$

we see that Assumption 3 indirectly constraints the ranks of $\{\Theta(t)\}_{t=1}^{T}$. In particular, if $\Theta(\eta_k)$ and $\Theta(\eta_{k-1})$ are the expected adjacency matrices of stochastic block models with M_1 and M_2 communities respectively, then $\operatorname{rank}(\Theta(\eta_k) - \Theta(\eta_{k-1})) \leq \min\{M_1 + M_2, n\}$.

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 (19) match the one in Assumption 2.
- 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 (though, of course, necessarily, $r \leq n$.)

3.1. *Upper and lower bounds on the localization error.* The next theorem derives improved localization rates for the LR procedure under and is the main result of this section.

THEOREM 2. *Assume the model described in Assumption 1 and the conditions of Assumption 3. There exist absolute positive constants C, C_ϵ, C_2 and C_3 such that if $\{\nu_k\}_{k=1}^{K} \subset (2, \ldots, T)$ is an increasing sequence satisfying*

$$\max_{k=1,\ldots,K} |\nu_k - \eta_k| < \Delta/6, \tag{20}$$

then the collection of the estimated change points $\mathcal{B} = \{\widehat{\eta}_k\}_{k=1}^{K}$ returned by the LR procedure with input parameters $(0, T)$, $\{\nu_k\}_{k=1}^{K}$,

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

is such that

$$\mathbb{P}\Big\{\max_{k=1,\ldots,K} |\eta_k - \widehat{\eta}_k| \leq \epsilon\Big\} \geq 1 - 2T^{3-3C_\varepsilon/4} - 4T^{3-3C_3^2/8},$$

where

$$\epsilon = C_2 \frac{\log^2(T)}{\kappa_0^2 n^2 \rho}. \tag{21}$$

The proof of Theorem 2 is given in Appendix A. The values and dependence among the constants can be tracked throughout and, just like with Theorem 1, demand that the constant C_α in the signal-to-noise ratio condition (19) is chosen large enough.

It is immediate to see that Theorem 2 offers stronger consistency guarantees than Theorem 1. Indeed, using Assumption 3 along with the assumption that $\rho \geq \frac{\log(n)}{n}$, we see that the localization rate implied by (21) is

$$\frac{\epsilon}{\Delta} \leq \frac{1}{C_\alpha^2 \log^{2\xi}(T) r \log n} \to 0, \tag{22}$$

as $T \to \infty$. This upper bound on the localization error is of smaller order than the one in (18) afforded by Theorem 1. Furthermore, as remarked above, change point consistency is still guaranteed even as $n \asymp 1$. On the other hand, if n is diverging in T, we may set $\xi = 0$ in Assumption 3.

To gain a further appreciation for the type of improvement Theorem 2 delivers over Theorem 1, assume that $r \asymp n$. Then, according to Theorem 1, in order for the NSB procedure to yield the

same localization error as in Theorem 2 it appears necessary to strengthen the signal-to-noise ratio requirement to be

$$\kappa_0\sqrt{n\rho\Delta} \gtrsim \sqrt{n}\log^{1+\xi}(T)$$

instead of just $\kappa_0\sqrt{n\rho\Delta} \gtrsim \log^{1+\xi}(T)$.

In addition to Assumption 3, Theorem 2 further requires that the sequence $\{\nu_k\}_{k=1}^K$ of preliminary estimates used as an input to the procedure to be within a constant fraction of Δ from the true change points; see (20). Notice that this assumption may be satisfied even if the ratio $\max_{k=1,\dots,K}|\nu_k - \eta_k|$ is not a vanishing fraction of Δ, thus failing to fulfill Definition 2. Of course, the change point estimators obtained using the NBS algorithm satisfy (20) with high probability and for all large enough T, as demonstrated above in Theorem 1, and therefore can be used as inputs to the LR algorithm.

Finally, the choices of threshold parameters τ_2 and τ_3 stem from the analysis of the USVT procedure for network estimation in Xu (2018). In particular, the parameter τ_2 serves as a cutoff for the upper bound of the operator norm difference between the sample and population version of certain matrices of interest.

In the second result of the section we prove that the localization rate demonstrated in Theorem 2 is nearly minimax optimal, save for a term poly-logarithmic in T.

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 $P^T_{\kappa_0,\Delta,n,\rho}$ denote the corresponding joint distribution. Consider the class of distributions*

$$\mathcal{Q} = \left\{P^T_{\kappa_0,\Delta,n,\rho} : \kappa_0 \le 1/2,\ \rho \le 1/2\right\}.$$

Then,

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

The family of distributions $\mathcal{Q}$ allows for a wide range of changes. Indeed, the constraints that $\kappa_0 \le 1/2$ is fairly general and, in particular, include the challenging scenario where all edge probabilities change at the change points. The constant $1/2$ is arbitrary and can be replaced by any constant between 0 and 1.

3.2. *Sparse stochastic block model.* In Theorem 2 we show that, for network models with rank constraints, combining the NBS and the LR algorithms yields nearly optimal localization under the low rank assumption and the scaling described in Assumption 3. Low rank network models include a wide range of common network models, e.g. the Erdős–Rényi random graph model (Erdős and Rényi, 1959), stochastic block models (e.g. Holland et al., 1983) and random dot product models (Young and Scheinerman, 2007). However, in these models, it is often also assumed that no self-loops are allowed, i.e. the diagonal entries of the adjacency matrices are always 0. As a result, the low rank assumption no longer holds. In this section we show that, for the case of stochastic block models, this issue can be overcome and that the guarantees of Theorem 2 hold also in this case. For completeness, we include the definition of a sparse stochastic block model and some of its properties.

DEFINITION 4 (Sparse Stochastic Block Model). *A network is from a sparse stochastic block model with size n, sparsity parameter ρ, membership matrix $Z \in \{0,1\}^{n\times s}$ and connectivity matrix*

$Q \in [0,1]^{r\times r}$, *if the expected adjacency matrix satisfies*

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

Each of the rows of the membership matrix Z contains only one non-zero entry; moreover, Z is a column full rank matrix, i.e. $\operatorname{rank}(Z) = r$. *In particular,* $\operatorname{rank}(ZQZ^\top) \le r$, *with identity holding when Q is a full rank matrix.*

In order to accommodate for the lack of self-loops we rely on a new set of conditions, described next.

Assumption 4. *Let $\{A(t)\}_{t=1}^T \in \mathbb{R}^{n\times n}$ be a sequence of independent adjacency matrices satisfying the dynamic network model of Assumption 1. Assume that, for all $k = 1,\dots,K$,*

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

where $\Lambda(k) = Z_k Q_k Z_k^\top$, Z_k is a membership matrix such that $\operatorname{rank}(Z_k) \le r$, *and Q_k is a connectivity matrix. Furthermore, for a constant $C_\alpha > 0$ and any $\xi > 0$,*

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

Assumption 4 differs from Assumption 3 only in the how it constraints the difference of the expected adjacency matrices. Indeed, under Assumption 4, $\Theta(\eta_k) - \Theta(\eta_k - 1)$ is typically not a low rank matrix, and therefore Assumption 3 would not hold. Aside from this, the signal-to-noise condition is identical in the two sets of assumptions.

Now, unlike in the problem of recovering the community assignment in a stochastic block model, where zeroing out the diagonal entries of the low rank matrix corresponding to the expected adjacency matrix is essentially inconsequential, in the localization problem this is not the case. To see this, observe that if the time interval $(s+1,\dots,e)$ contains one change point η_k, then for $t \in (s+1,\dots,e-1)$,

$$\widetilde{\Theta}^{s,e}(t) = \begin{cases} \sqrt{\frac{t-s}{(e-s)(e-t)}}(e-\eta_k)(\Lambda(k) - \operatorname{diag}(\Lambda(k))), & \text{if } t \le \eta_k, \\ \sqrt{\frac{e-t}{(e-s)(t-s)}}(\eta_k - s)(\Lambda(k) - \operatorname{diag}(\Lambda(k))), & \text{if } t \ge \eta_k. \end{cases}$$

In particular, at $t = \eta_k$,

$$\left\| \sqrt{\frac{(t-\eta_k)(e-\eta_k)}{(e-s)}} \operatorname{diag}(\Lambda(k)) \right\|_F \lesssim \rho\sqrt{n}\sqrt{\min\{e-\eta_k, \eta_k - s\}},$$

a quantity that depends on the spacing between change points and may potentially be quite large. In order to handle such issue we make the following assumption.

Assumption 5. *For each $k = 0,\dots,K$, set*

$$\Theta(\eta_k) = \Gamma(k) - \operatorname{diag}(\Gamma(k)),$$

where $\Gamma(k) = Z'_k Q'_k Z_k'^\top$, Z'_k is a membership matrix and Q'_k is a connectivity matrix. For an absolute constant $C_\Gamma > 0$, it holds that

$$\|\Gamma(k)\|_{\mathrm{F}} \ge C_\Gamma \|\operatorname{diag}(\Gamma(k))\|_{\mathrm{F}}.$$

Since $\|\Gamma(k)\|_{\mathrm{F}}$ is of order no larger than ρn and $\|\mathrm{diag}(\Gamma(k))\|_{\mathrm{F}}$ is of order no larger than $\rho\sqrt{n}$, overall Assumption 5 is a mild condition. Of course, if $\Gamma(k)$ is a diagonally-dominant matrix, then it is unclear how to estimate $\Gamma(k)$ 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 4 and Assumption 5, then the same conclusion still holds.*

The proof of Theorem 3 can be found in Appendix A. The main difference between this proof and the proof of Theorem 2 is the treatment on the diagonal entries under Assumption 5.

4. Illustrative Simulations. In this section we will present the results of various illustrative simulations intended to corroborate the theory developed in the paper and to demonstrate the type of improvements the LR delivers over the NBS. As for this, we will use well-tuned tuning parameters, which will be reported.

We point out that we could not find a methodology for the problem of multiple change point localization in network models with which to directly compare the NBS and LR. We have looked into existing methods for multiple change point localization that have been proposed for change point localization in settings different than dynamic network models, such as the ones put forward in Keshavarz et al. (2018), Cho (2015), Cho and Fryzlewicz (2015) and Wang and Samworth (2018), among others. However, none of these procedures could be successfully deployed in the simulation settings described below. For this reason, we do not report the results of these comparison.

We consider the following three simulation settings. All settings have equally-spaced change points, therefore the total number of time points $T = (K+1)\Delta$.

Setting (i). We set $\Delta = 60, 80, 120, 200$, $K = 2$, $n = 150$ and $\rho = 0.02$. Each network is generated from a balanced 3-community stochastic block model. At the change points, the connectivity matrices are

$$Q_1 = \rho \begin{pmatrix} 0.6 & 1 & 0.6 \\ 1 & 0.6 & 0.5 \\ 0.6 & 0.5 & 0.6 \end{pmatrix}, \quad Q_2 = \rho \begin{pmatrix} 0.6 & 0.5 & 0.6 \\ 0.5 & 0.6 & 1 \\ 0.6 & 1 & 0.6 \end{pmatrix} \quad \text{and} \quad Q_3 = Q_1,$$

respectively.

Setting (ii). We set $\Delta = 60, 80, 120, 200$, $K = 2$, $n = 150$, $\rho = 0.015$ and the connectivity matrix be

$$Q = \rho \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{pmatrix}.$$

Each network is generated from a balanced 3-community stochastic block model. At the change points, membership are reshuffled randomly.

Setting (iii). We set $\Delta = 80$, $K = 2$, $n = 150, 180, 210, 240$, $\rho = 0.01$. Each network is generated from a balanced 3-community stochastic block model. At the change points, the connectivity matrices are

$$Q_1 = \rho \begin{pmatrix} 0.9 & 0.8 & 0.3 \\ 0.8 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 \end{pmatrix}, \quad Q_2 = \rho \begin{pmatrix} 0.3 & 0.3 & 0.7 \\ 0.3 & 0.6 & 0.3 \\ 0.7 & 0.3 & 0.3 \end{pmatrix} \quad \text{and} \quad Q_3 = \rho \begin{pmatrix} 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.6 \\ 0.3 & 0.6 & 0.1 \end{pmatrix},$$

respectively.

For each of the above settings we simulated a dynamic network realization and applied both the NBS and LR 200 times. In fact, we have applied a simplified version of the NBS algorithm based on the BS procedure (see, e.g. Vostrikova, 1981) instead of WBS. Since the number of change points is small, it can be shown that the guarantees of Theorem 3 hold true even for this simpler, computationally less demanding algorithm[2].

To evaluate the performance of the algorithms, for each simulation we recorded

- $d(\widehat{S}, S)/T$, the Hausdorff distance between the set of change point estimators and the set of the true change points, normalized by T,
- $|\widehat{K} - K|$, the absolute difference between the numbers of the change point estimators and the true change points,
- and Prop, the proportion of simulations (out of 200) for which $\widehat{K} = K$.

Table 1 presents the results in the form of mean(standard error). The columns labeled by sub. $d(\widehat{S}, S)/T$ displays the results only for the simulations in which $\widehat{K} = K$. All the numerical analysis were conducted on machines with CPU Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz.

Since the LR is a local refinement to the NBS, the columns corresponding to the LR algorithm report, by construction, the same $\widehat{K}$ and therefore the same correct proportion. Due to (20), which requires the LR to be deployed only as a refinement of an estimator that returns the correct number of change points, in order to show the improvement afforded by the LR we only considered simulations in which the NBS outputs the correct number of change points.

	$d(\widehat{S}, S)/T$		$\lvert\widehat{K} - K\rvert$	Prop	sub. $d(\widehat{S}, S)/T$	
	NBS	LR			NBS	LR
			Setting (i)			
$T = 180$	0.164(0.010)	0.130(0.011)	0.955(0.062)	0.400	0.043(0.005)	0.008(0.004)
$T = 240$	0.113(0.009)	0.078(0.009)	0.820(0.063)	0.485	0.023(0.002)	0.000(0.000)
$T = 360$	0.049(0.006)	0.027(0.006)	0.450(0.051)	0.675	0.010(0.001)	0.000(0.000)
$T = 600$	0.019(0.003)	0.003(0.001)	0.265(0.036)	0.770	0.004(0.000)	0.000(0.000)
			Setting (ii)			
$T = 180$	0.033(0.003)	0.004(0.002)	0.195(0.033)	0.830	0.021(0.002)	0.000(0.000)
$T = 240$	0.013(0.002)	0.001(0.000)	0.070(0.018)	0.930	0.009(0.001)	0.000(0.000)
$T = 360$	0.006(0.001)	0.001(0.000)	0.070(0.018)	0.930	0.003(0.000)	0.000(0.000)
$T = 600$	0.002(0.000)	0.000(0.000)	0.055(0.016)	0.945	0.001(0.000)	0.000(0.000)
			Setting (iii)			
$n = 150$	0.115(0.010)	0.095(0.010)	0.415(0.038)	0.610	0.029(0.004)	0.014(0.005)
$n = 180$	0.027(0.003)	0.008(0.003)	0.250(0.034)	0.775	0.012(0.001)	0.000(0.000)
$n = 210$	0.013(0.002)	0.000(0.000)	0.165(0.027)	0.840	0.004(0.001)	0.000(0.000)
$n = 240$	0.013(0.002)	0.000(0.000)	0.165(0.026)	0.835	0.002(0.000)	0.000(0.000)

TABLE 1

Simulation results for both the NBS and LR.

As for tuning parameters, recall that we have one tuning parameter τ_1 for the NBS, and two tuning parameters, τ_2 and τ_3, for the LR. The choices of tuning parameters in these three different settings are given in Table 2, where Inf is equivalent to no entrywise truncation in the USTV step, and M is the number of communities in the stochastic block model. In selecting the tuning

[2]In general, however, when the number of change points increases with T, BS is sub-optimal compared to WBS. Note that the default choices in R packages based on Cho (2015), Cho and Fryzlewicz (2015) and Wang and Samworth (2018) are all based on BS instead of WBS.

parameters we have used the true number of communities M; of course, in practice, this quantity needs to be estimated from the data (e.g. Chen and Lei, 2018). Finally we estimate ρ using $\hat{\rho}$, defined to be the 95% quantile of

$$\left\{ T^{-1} \sum_{t=1}^{T} A_{ij}(t),\ 1 \le i, j \le n \right\}.$$

Setting	τ_1	τ_2	τ_3
(i)	$n\hat{\rho}\log^2(T)/21$	$Mn\hat{\rho}$	$\hat{\rho}$
(ii)	$n\hat{\rho}\log^2(T)/20$	$Mn\hat{\rho}$	Inf
(iii)	$3n\hat{\rho}/4$	$Mn\hat{\rho}$	Inf

TABLE 2

Tuning parameter choices.

It can be seen from Table 1 that, with these choices of the tuning parameters, the performance of both the NBS and LR improves as T, n and ρ increase. In addition, the LR significantly outperforms the NBS.

For all the settings described above, we have also conducted additional simulations with an omnibus default choice for the tuning parameter which does not require knowledge of M : $\tau_1 = n\hat{\rho}\log^2(T)/20$. The results are shown in Table 3. Due to the default choice of the tuning parameter, it is not easy to show how the performance changes with different model parameters. Therefore we only collect the NBS results to demonstrate that we can achieve good performances in terms of $|\widehat{K} - K|$, $d(\widehat{S}, S)/T$ and Prop, with easily chosen tuning parameter.

	$d(\widehat{S}, S)/T$	$\lvert\widehat{K} - K\rvert$	Prop	Time(second/repetition)
		Setting (i)		
$T = 180$	0.166(0.010)	1.025(0.062)	0.360	1.607(0.030)
$T = 240$	0.121(0.010)	0.760(0.061)	0.520	3.104(0.055)
$T = 360$	0.042(0.006)	0.285(0.044)	0.805	7.126(0.060)
$T = 600$	0.011(0.002)	0.125(0.023)	0.875	20.837(0.149)
		Setting (ii)		
$T = 180$	0.332(0.000)	0.970(0.012)	0.030	1.061(0009)
$T = 240$	0.444(0.000)	0.955(0.015)	0.045	2.032(0.028)
$T = 360$	0.667(0.000)	0.985(0.009)	0.015	3.950(0.023)
$T = 600$	1.111(0.000)	1.000(0.000)	0.000	10.994(0.022)
		Setting (iii)		
$n = 150$	0.154(0.013)	0.415(0.038)	0.610	1.861(0.004)
$n = 180$	0.050(0.006)	0.255(0.035)	0.770	2.683(0.010)
$n = 210$	0.015(0.002)	0.195(0.032)	0.825	4.936(0.021)
$n = 240$	0.009(0.001)	0.210(0.033)	0.815	8.785(0.039)

TABLE 3

Simulation results for the NBS with a default tuning parameter.

It can be seen in Table 3 that with this default choice of tuning parameter, the NBS is still producing good results.

5. Discussion. We have studied the change point localization problem in sparse dynamic network settings. We have proposed two computationally-efficient algorithms based on CUSUM statistics: Network Binary Segmentation (NBS) and Local Refinement (LR). The NBS is able to

localize multiple change points consistently under virtually all parameter scalings for which this task is feasible. The LR guarantees sharper localization errors under slightly stronger scalings and is nearly minimax rate-optimal under those scalings. Our results are applicable to a wide class of dynamic network models and, in particular to the ones assuming a sequence of time-varying stochastic block models.

While we are able to demonstrate a nearly optimal localization procedure only under a certain low rank assumption (see Assumption 3), it remains an open problem to design a computationally efficient algorithm that is provably optimal across all scalings for which consistent localization is possible, described in Assumption 2.

The assumptions used in this paper can be possibly generalized in a few directions. If one wishes to relax the independence across time and/or within networks, or replace the Bernoulli assumption with other distributional assumptions (e.g. sub-Gaussian), then it will be necessary to change in the proofs of the concentration inequalities and the corresponding large probability events. This in turn may lead to different scaling requirements for consistency and optimality, as well as possibly different localization error bounds.

It is worth noting that, assuming a stochastic block model at each time point, replacing the USVT algorithm used in the LR procedure with an NP-hard graphon-based algorithm (see, e.g. Pensky, 2016; Gao et al., 2015) will produce the nearly optimal rate (11) under the scaling

$$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta}\frac{\left(1+r^2/n\right)}{n}, \tag{23}$$

which is weaker than the scaling we assume for our polynomial time algorithms (NBS and LR), namely (10). Equations (9), (10) and (23) reveal that

(i) in the very sparse regime, i.e. $r \lesssim \sqrt{n}$, there is no gap between the scaling (23) required by NP-hard algorithms and the scaling (9);
(ii) in the moderately sparse regime, i.e. $\sqrt{n} \lesssim r \lesssim n$, then there is a gap between statistical and computational limits;
(iii) in the very dense regime, i.e. $r \asymp n$, (10) and (23) are the same, which means NP-hard algorithms are not gaining over polynomial methods.

These observations is consistent with similar phenomena observed in other statistical problems, see e.g. Zhang et al. (2012), Loh and Wainwright (2013), to name but a few.

To summarize, we have the following Table 4.

Rate	Scaling	Algorithm
$\epsilon/T = o(1)$	$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta}\frac{1}{n}$	Poly
$\epsilon/T = \epsilon_{\text{opt}}\frac{\log^2(T)}{T}$	$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta}\frac{r}{n}$	Poly
	$\rho\kappa_0^2 \gtrsim \frac{\log^{2+2\xi}(T)}{\Delta}\frac{\left(1+r^2/n\right)}{n}$	NP

TABLE 4
Summary of our rates results.

Acknowledgments. We would like to thank an anonymous reviewer and the associate editor for constructive comments that led to improvements in the presentation of the paper.

SUPPLEMENTARY MATERIAL

Supplement: Optimal Change point detection and localization in sparse dynamic networks

(doi: COMPLETED BY THE TYPESETTER; .pdf). We moved the appendices containing many of the technical proofs and detailed discussions to the supplementary document (Wang et al., 2018a).

References.

Anastasiou, A. and Fryzlewicz, P. (2019). Detecting multiple generalized change-points by isolating single ones. *arXiv preprint arXiv:1901.10852*.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286** 509–512.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **81** 649–672.

Bhattacharjee, M., Banerjee, M. and Michailidis, G. (2018). Change point estimation in a dynamic stochastic block model. *arXiv preprint arXiv:1812.03090*.

Bhattacharyya, S. and Chatterjee, S. (2017). Spectral clustering for dynamic stochastic block model. Tech. rep., Working Paper.

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.

Chen, K. and Lei, J. (2018). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, **113** 241–251.

Cho, H. (2015). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics* in press.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77** 475–507.

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 Statistcs)*, **66** 607–627.

Eichinger, B., Kirch, C. et al. (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli*, **24** 526–564.

Erdős, P. and Rényi, A. (1959). On random graphs, I. *Publicationes Mathematicae (Debrecen)*, **6** 290–297.

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 ®in Machine Learning* 129–233.

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.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* 016107.

Keshavarz, H., Michailidis, G. and Atchade, Y. (2018). Sequential change-point detection in high-dimensional gaussian graphical models. *arXiv preprint arXiv:1806.07870*.

Kolaczyk, E. D. (2017). *Topics at the Frontier of Statistics and Network Analysis:(re) visiting the Foundations*. Cambridge University Press.

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*.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*. 476–484.

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.

Mukherjee, S. S. (2018). *On Some Inference Problems for Networks*. Ph.D. thesis.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41** 100–115.

Pensky, M. (2016). Dynamic network models and graphon estimation. *arXiv preprint arXiv:1607.00673*.
Pensky, M. and Zhang, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, **13** 678–709.
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.
Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems*. Ph.D. thesis.
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, D., Yu, Y. and Rinaldo, A. (2018a). Supplement to "optimal change point detection and localization in sparse dynamic networks".
Wang, D., Yu, Y. and Rinaldo, A. (2018b). Univariate mean change point detection: Penalization, cusum and optimality. *arXiv preprint arXiv:1810.09498*.
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. (2018). High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 57–83.
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. (2018). Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*. 5429–5438.
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.
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.
Zhang, Y., Wainwright, M. J. and Duchi, J. C. (2012). Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems*. 1502–1510.
Zhao, Z., Chen, L. and Lin, L. (2019). Change-point detection in dynamic networks via graphon estimation. *arXiv preprint arXiv:1908.01823*.

APPENDIX A: PROOFS OF THEOREMS 1, 2 AND 3

For simplicity, we set

$$\|\Theta(\eta_k) - \Theta(\eta_k - 1)\|_{\mathrm{F}} = \kappa_k > 0, \text{ for any } k = 1, \ldots, K.$$

Proof of Theorem 1. The value of the constants in statememt of the theorem can be tracked in the proof. The hierarchy can be abstracted as follows: first, c and c_T are chosen such that (16) tends to 0 as $T \to \infty$; then, C_β can be chosen depending on c and c_T; the constant c_2 therefore depends on C_β and C_α; finally, a sufficiently large $C_1 > 0$ is chosen and depends on all the aforementioned constants and C_R. In particular, increasing C_α would decrease the lower bound of C_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 (25) in Section S.5 holds. By Lemma S.24,

the probability of the complementary event is smaller than

$$\exp\left\{\log\left(\frac{T}{\Delta}\right) - \frac{M\Delta}{4C_R T}\right\},$$

which vanishes provided that

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

For $0 \leq s < t < e \leq T$, we consider the event

$$\mathcal{A}(s,t,e) = \left\{\left|(\widetilde{A}^{s,e}(t), \widetilde{B}^{s,e}(t)) - \|\widetilde{\Theta}^{s,e}(t)\|_{\mathrm{F}}^2\right| \leq C_\beta \log(T)\left(\|\widetilde{\Theta}^{s,e}(t)\|_{\mathrm{F}} + \log^{1/2}(T)\rho n\right)\right\}. \tag{24}$$

Due to Lemma S.6, 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 on 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 (2018) 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, on 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

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

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 (17).

Observe that

$$\epsilon = C_1 \log(T)\left(\frac{\sqrt{\Delta}}{\kappa_0 n\rho} + \frac{\log^{1/2}(T)}{\kappa_0^2 n\rho}\right) \leq C_1\left(\frac{\Delta}{C_\alpha \log^{1/2}(n)\log^{\xi}(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)}\right), \tag{25}$$

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 $\mathrm{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau)$ (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 (25) and (15),

$$\epsilon \leq 64^{-1}\Delta \leq 64^{-1}(e'_m - s'_m),$$

where the second inequality is fulfilled by choosing a sufficiently large C_α. Therefore, the interval

$$[s_m, e_m] = [s'_m + 64^{-1}(e'_m - s'_m), e'_m - 64^{-1}(e'_m - s'_m)],$$

contains no change points. To see this, notice that, on the event $\mathcal{A}$, $\widetilde{\Theta}^{s_m, e_m}(t) = 0$ for all $t \in (s_m, e_m)$, as there is no change point in $[s_m, e_m]$. Furthermore, by Lemma S.6, there exists a large enough constant $C_\beta > 0$ such that

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

Thus, with the input parameter τ satisfying

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

we conclude that $\mathrm{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau)$ 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 $\mathrm{NBS}((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau)$. On 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 $[s_m, e_m] = [s'_m + 64^{-1}(e'_m - s'_m), e'_m - 64^{-1}(e'_m - s'_m)]$, then one has that

$$\eta_p - \Delta 3/4 \leq s_m \leq \eta_p - \Delta/8 \quad \text{and} \quad \eta_p + \Delta/8 \leq e_m \leq \eta_p + \Delta 3/4. \tag{26}$$

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

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

It then follows from Lemma S.17 that

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

where the last inequality stems from (26). Thus, due to Assumption 1 part *1.* and Assumption 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

$$\kappa_p\sqrt{\Delta}/4 \geq \kappa_0 n\rho\sqrt{\Delta}/4 \geq C_\alpha/4\sqrt{n\rho}\log^{1+\xi}(T) \geq C_\alpha/4\log^{1/2}(n)\log^{1+\xi}(T) > 2C_\beta\log(T), \tag{27}$$

provided that, for $n, T \geq 2$,

$$C_\beta < \min\left\{8^{-1}C_\alpha\log^{\xi}(T)\log^{1/2}(n),\, C_\alpha^2/16\log^{1/2+2\xi}(T)\right\}. \tag{28}$$

We remark that as for the hierarchy of all the absolute constants involved, (28) is a constraint on C_α. Thus, with a large enough C_α, there exists an absolute constant $c_2 > 0$, such that

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

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

$$a_{m*} = (\widetilde{A}^{s_{m*},e_{m*}}(b_{m*}), \widetilde{B}^{s_{m*},e_{m*}}(b_{m*})) \geq c_2(\kappa_{\max}^{s,e})^2\Delta, \tag{29}$$

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

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

The 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 Algorithm 1.

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

$$\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4 \tag{30}$$

and that

$$|b_{m*} - \eta_p| \leq \epsilon. \tag{31}$$

In this step we will prove that (30) holds. Denote

$$[s_{m*}, e_{m*}] = [s'_{m*} + 64^{-1}(e'_{m*} - s'_{m*}), e_{m*} - 64^{-1}(e'_{m*} - s'_{m*})].$$

Suppose for the sake of contradiction that

$$\max_{s_{m*}<t<e_{m*}} \|\widetilde{\Theta}^{s_{m*},e_{m*}}(t)\|_{\mathrm{F}}^2 < c_2(\kappa_{\max}^{s,e})^2\Delta/2. \tag{32}$$

Then

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

where the first inequality is due to the definition of the event $\mathcal{A}$, the second inequality follows from (32) and the third inequality from Assumption 2, for an appropriately large C_α. This contradicts (29). Therefore

$$\max_{s_{m*}<t<e_{m*}} \|\widetilde{\Theta}^{s_{m*},e_{m*}}(t)\|_{\mathrm{F}}^2 \geq c_2(\kappa^{s,e}_{\max})^2\Delta/2. \tag{33}$$

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 S.17,

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

which contradicts (33).

Therefore, since $e_{m*} - s_{m*} \geq c_2\Delta/2$,the bound (25) implies that

$$\epsilon \leq C_1 \left(\frac{\Delta}{C_\alpha \log^{1/2}(n) \log^{\xi}(T)} + \frac{\Delta}{C_\alpha^2 \log^{1/2+2\xi}(T)} \right) \leq 64^{-1}(e'_{m*} - s'_{m*}), \tag{34}$$

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 (29), $[s_{m*}, e_{m*}]$ contains at least one undetected change point.

Step 4. In the final step of the proof we will show that (31) occurs. To that end, we will apply Lemma S.7. Let

$$\lambda = \max_{s_{m*}<t<e_{m*}} \left| (\widetilde{A}^{s_{m*},e_{m*}}(t), \widetilde{B}^{s_{m*},e_{m*}}(t)) - \|\widetilde{\Theta}^{s_{m*},e_{m*}}(t)\|_{\mathrm{F}}^2 \right|. \tag{35}$$

Observe that (33) and (27) imply that

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

and

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

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

$$\lambda \leq C_\beta \log(T) \left(\log^{1/2}(T)\rho n + \max_{s_{m*}<t<e_{m*}} \|\widetilde{\Theta}^{s_{m*},e_{m*}}(t)\|_{\mathrm{F}} \right) \leq c_3 \max_{s_{m*}<t<e_{m*}} \|\widetilde{\Theta}^{s_{m*},e_{m*}}(t)\|_{\mathrm{F}}^2. \tag{36}$$

Since (5) follows from (29), (6) follows from (35), and (7) follows from (36), all the conditions in Lemma S.7 hold. Lemma S.7 implies that there exists an undetected change point η_p within $[s, e]$ such that

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

and this combining with (33) provides that

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

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. □

Proof of Theorem 2. The dependence among the constants involved in Theorem 2 is as follows. Firstly, C and C_ε are chosen to guarantee that $2T^{3-3C_\varepsilon/4} \to 0$. Secondly, C_3 is chosen such that $4T^{3-3C_3^2/8} \to 0$. In particular, we may take $C > 64 \times 2^{1/4} e^2$, $C_\varepsilon > 12$ and $C_3 > 2\sqrt{2}$. Finally, the leading constant $C_2 > 0$ in the error bound depends on all the aforementioned constants and the signal-to-noise ratio constant C_α in Assumption 3, which should be chosen to be sufficiently large.

For convenience, we have broken down the proof in five steps, each of which is applied to every $k \in \{1, \ldots, K\}$. 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 (20), $\eta_k \in [\nu_{k-1}, \nu_{k+1}]$ and

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

Similar calculations show also that

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

Therefore, it holds that

$$1/2 \min\{\nu_k - \nu_{k-1}, \nu_{k+1} - \nu_k\} \geq \Delta/6.$$

As a result, the interval

$$[s, e] = [\nu_{k-1} + 1/2(\nu_k - \nu_{k-1}), \nu_{k+1} - 1/2(\nu_{k+1} - \nu_k)]$$

contains only one change point η_k. We have that

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

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 S.17,

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

Next, we set

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

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

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

we obtain that

$$\begin{aligned} \|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}^2 &= \frac{\nu_k - s}{(e-s)(e-\nu_k)}(e-\eta_k)^2\|\Lambda(k)\|_{\mathrm{F}}^2 = \widetilde{\Delta}_k^2\left(\frac{e-\eta_k}{e-\nu_k}\right)^2\kappa_k^2 \\ &= \widetilde{\Delta}_k^2\left(1 - \frac{\eta_k - \nu_k}{e-\nu_k}\right)^2\kappa_k^2 \geq \frac{\Delta}{6}\left(1 - \frac{\Delta/6}{\Delta/3}\right)^2\kappa_k^2 \geq \Delta\kappa_k^2/24. \end{aligned} \tag{37}$$

Step 3. We next apply Lemma S.8 by letting $\varepsilon = C_\varepsilon \log(T)$, with $C_\varepsilon > 12$. Define the event

$$\mathcal{A} = \left\{ \sup_{0 \leq s < t < e \leq T} \|\widetilde{A}^{s,e}(t) - \widetilde{\Theta}^{s,e}(t)\|_{\mathrm{op}} \leq C\sqrt{n\rho} + C_\varepsilon \log(T) \right\},$$

where $C > 64 \times 2^{1/4}e^2$. Due to Lemma S.8, we have $\mathbb{P}(\mathcal{A}) \geq 1 - 2T^{3-C_\varepsilon/4}$.

We then apply Lemma S.11. Set $\tau_2 = (3/4)(C\sqrt{n\rho} + C_\varepsilon \log(T))$, and define

$$\mathcal{B} = \left\{ \sup_{0 \leq s < t < e \leq T} \|\operatorname{USVT}(\widetilde{A}^{s,e}(t), \tau_2, \infty) - \widetilde{\Theta}^{s,e}(t)\|_{\mathrm{F}} \leq 3\sqrt{r}\big(C\sqrt{n\rho} + C_\varepsilon \log(T)\big) \right\}.$$

In order to apply Lemma S.11, let $A = \widetilde{A}^{s,e}(t)$, $B = \widetilde{\Theta}^{s,e}(t)$ and $\tau = \tau_2$. We then have $\mathbb{P}(\mathcal{B}) \geq 1 - 2T^{3-C_\varepsilon/4}$.

Let

$$\widehat{A}^{s,e}(\nu_k) = \operatorname{USVT}(\widetilde{A}^{s,e}(\nu_k), \tau_2, \tau_3\widetilde{\Delta}_k). \tag{38}$$

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

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

On the event $\mathcal{B}$,

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

By the triangle inequality and Assumption 3, we have that

$$\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}} \geq \|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}} - 3\sqrt{r}\big(C\sqrt{n\rho} + C_\varepsilon \log(T)\big) \geq c_1'\sqrt{\Delta}\kappa_k, \tag{39}$$

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,

$$\begin{aligned} 2\left(\frac{\widetilde{\Theta}^{s,e}(\nu_k)}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}}, \frac{\widehat{A}^{s,e}(\nu_k)}{\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}}\right) &= 2 - \left\| \frac{\widetilde{\Theta}^{s,e}(\nu_k)}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}} - \frac{\widehat{A}^{s,e}(\nu_k)}{\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}} \right\|_{\mathrm{F}}^2 \\ \geq 2 - 4\left(\frac{\|\widetilde{\Theta}^{s,e}(\nu_k) - \widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}}{\max\left\{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}, \|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}\right\}}\right)^2 &\geq 2 - \frac{9r\big(C\sqrt{n\rho} + C_\varepsilon \log(T)\big)^2}{(c_1')^2 \kappa_k^2 \Delta} \geq 1, \end{aligned}$$

where the second inequality follows from the definition of the event $\mathcal{B}$ and from (37), while the last inequality follows from Assumption 3 with a sufficiently large C_α. Therefore,

$$(\widetilde{\Theta}^{s,e}(\nu_k), \widehat{A}^{s,e}(\nu_k)/\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}) \geq \|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}/2 \geq (4\sqrt{6})^{-1}\sqrt{\Delta}\kappa_k, \tag{40}$$

where in the last inequality we have used again (37).

Step 4. Since $\{B(t)\}_{t=1}^T$ is independent of $\{A(t)\}_{t=}^T$, the distribution of $\{B(t)\}_{t=1}^T$ does not change on the event $\mathcal{B}$. Observe that, from (38),

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

In combination with (39), 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 S.4, we obtain that, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{e-s}}\sum_{t=s+1}^{e}\left(\Theta(t) - B(t),\, \widehat{A}^{s,e}(\nu_k)/\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}\right)\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-3/2\varepsilon^2}{3\rho + \varepsilon\rho/(c_1'\kappa_k\sqrt{\Delta})}\right).$$

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

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{e-s}}\sum_{t=s}^{e}\left(\Theta(t) - B(t),\, \widehat{A}^{s,e}(\nu_k)/\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}\right)\right| \geq C\sqrt{\rho}\log(T)\right) \leq 2T^{-3C^2/8}. \tag{41}$$

Similar arguments also show that

$$\mathbb{P}\left(\left|\left(\widetilde{\Theta}^{s,e}(t) - \widetilde{B}^{s,e}(t),\, \widehat{A}^{s,e}(\nu_k)/\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}\right)\right| \geq C\sqrt{\rho}\log(T)\right) \leq 2T^{-3C^2/8}. \tag{42}$$

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

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

is a piecewise constant function with only one change point, namely η_k. Due to (40), it holds that

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

and, by (41) and (42),

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

and

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

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 $b_k=\arg\max_{s<t<e}|\widetilde{x}^{s,e}(t)|$ is an undetected change point such that, for a large enough constant $C_2>0$,

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

□

Proof of Theorem 3. 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 hold.

Let $[s,e]$ be defined as that in the proof of Theorem 2. We apply Lemma S.8 by letting $\varepsilon=C_\varepsilon\log(T)$, with $C_\varepsilon>12$. Define the event

$$\mathcal{A}'=\left\{\sup_{0\le s<t<e\le T}\|\widetilde{A}^{s,e}(t)-\widetilde{\Theta}^{s,e}(t)\|_{\mathrm{op}}\le C\sqrt{n\rho}+C_\varepsilon\log(T)\right\},$$

where $C>64\times 2^{1/4}e^2$. Due to Lemma S.8, we have $\mathbb{P}(\mathcal{A}')\ge 1-2T^{3-C_\varepsilon/4}$.

For $t\in\{1,\dots,T\}$, define $\Gamma(t)$ to be the block structure matrix satisfying

$$\Gamma(t)-\mathrm{diag}(\Gamma(t))=\Theta(t);$$

in addition, for any $s<t<e$, define

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

By Lemma S.13, on the event $\mathcal{A}'$, it holds that

$$\begin{aligned}\mathcal{B}'=\Bigg\{&\sup_{0\le s<t<e\le T}\|\,\mathrm{USVT}(\widetilde{A}^{s,e}(t),\tau_2,\infty)-\widetilde{\Gamma}^{s,e}(t)\|_{\mathrm{F}}^2\\ \le&9r\big(C\sqrt{n\rho}+C_\varepsilon\log(T)\big)^2+512\|\mathrm{diag}(\widetilde{\Gamma}^{s,e}(\nu_k))\|_{\mathrm{F}}^2\Bigg\}.\end{aligned}$$

Let

$$\widehat{A}^{s,e}(\nu_k)=\mathrm{USVT}(\widetilde{A}^{s,e}(\nu_k),\tau_2,\widetilde{\Delta}_k\tau_3).$$

Observe that since $\nu_k\le\eta_k$ and $\|\widetilde{\Lambda}^{s,e}(\nu_k)\|_\infty\le\widetilde{\Delta}_k\tau_3$, on the event $\mathcal{B}'$ it holds that

$$\begin{aligned}&\|\widehat{A}^{s,e}(\nu_k)-\widetilde{\Gamma}^{s,e}(\nu_k)\|_{\mathrm{F}}\le\|\,\mathrm{USVT}(\widetilde{A}^{s,e}(\nu_k),\tau_2,\infty)-\widetilde{\Gamma}^{s,e}(\nu_k)\|_{\mathrm{F}}\\ \le&3\sqrt{r}\big(C\sqrt{n\rho}+C_\varepsilon\log(T)\big)+16\sqrt{2}\|\mathrm{diag}(\widetilde{\Gamma}^{s,e}(\nu_k))\|_{\mathrm{F}}.\end{aligned}$$

Since $[s,e]$ contains only one change point η_k, by Assumption 5 and Lemma S.17,

$$\begin{aligned}\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}} &\geq \|\widetilde{\Gamma}^{s,e}(\nu_k)\|_{\mathrm{F}} - 3\sqrt{r}\big(C\sqrt{n\rho}+C_\varepsilon\log(T)\big) - 16\sqrt{2}\|\mathrm{diag}(\widetilde{\Gamma}^{s,e}(\nu_k))\|_{\mathrm{F}}\\ &\geq (1-16\sqrt{2}/C_\Gamma)\|\widetilde{\Gamma}^{s,e}(\nu_k)\|_{\mathrm{F}} - 3\sqrt{r}\big(C\sqrt{n\rho}+C_\varepsilon\log(T)\big)\\ &\geq \frac{1-16\sqrt{2}/C_\Gamma}{1+C_\Gamma}\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}} - 3\sqrt{r}\big(C\sqrt{n\rho}+C_\varepsilon\log(T)\big) \geq c_1'\sqrt{\Delta}\kappa_k, \qquad (43)\end{aligned}$$

with $c_1'>0$ by choosing proper constants. Equation (43) follows from the fact that

$$\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}} \leq \|\widetilde{\Gamma}^{s,e}(\nu_k)\|_{\mathrm{F}} + \|\mathrm{diag}(\widetilde{\Gamma}^{s,e}(\nu_k))\|_{\mathrm{F}} \leq (1+C_\Gamma)\|\widetilde{\Gamma}^{s,e}(\nu_k)\|_{\mathrm{F}}.$$

As a consequence,

$$\begin{aligned}&2\left(\frac{\widetilde{\Theta}^{s,e}(\nu_k)}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}}, \frac{\widehat{A}^{s,e}(\nu_k)}{\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}}\right) = 2 - \left\|\frac{\widetilde{\Theta}^{s,e}(\nu_k)}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}} - \frac{\widehat{A}^{s,e}(\nu_k)}{\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}}\right\|_{\mathrm{F}}^2\\ =&2 - \frac{\|\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}\Theta^{s,e}(\nu_k) - \|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}^2}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}^2\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}^2}\\ \geq&2 - \frac{\|\widetilde{\Theta}^{s,e}(\nu_k) - \widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}^2}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}^2} - \frac{\left|\|\widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}^2 - \|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}^2\right|}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}^2}\\ \geq&2 - 2\frac{\|\widetilde{\Theta}^{s,e}(\nu_k) - \widehat{A}^{s,e}(\nu_k)\|_{\mathrm{F}}^2}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}^2}\\ \geq&2 - 2\left(\frac{9r\big(C\sqrt{n\rho}+C_\varepsilon\log(T)\big)^2}{(c_1')^2\kappa_k^2\Delta} + \frac{513\|\mathrm{diag}(\widetilde{\Lambda}^{s,e}(\nu_k))\|_{\mathrm{F}}}{\|\widetilde{\Theta}^{s,e}(\nu_k)\|_{\mathrm{F}}}\right) \geq 1,\end{aligned}$$

where the second inequality follows from (37) and the event $\mathcal{B}'$, and the last inequality follows from Assumption 4 and (43). Therefore

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

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

Department of Statistics
University of Chicago
Chicago, IL 60637
E-mail: darenw@uchicago.edu

Department of Statistics
University of Warwick
Coventry, Cv4 7AL
United Kingdom
E-mail: yi.yu.2@warwick.ac.uk

Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA,15213
E-mail: arinaldo@cmu.edu