Message-passing sequential detection of multiple change points in networks

XuanLong Nguyen[1], Arash Ali Amini[1] and Ram Rajagopal[2]
1 Department of Statistics, University of Michigan
2 Department of Civil and Environmental Engineering, Stanford University

***Abstract*—We propose a probabilistic formulation that enables sequential detection of multiple change points in a network setting. We present a class of sequential detection rules for functionals of change points, and prove their asymptotic optimality properties in terms of expected detection delay time. Drawing from graphical model formalism, the sequential detection rules can be implemented by a computationally efficient message-passing protocol which may scale up linearly in network size and in waiting time. The effectiveness of our exact and approximate inference algorithms are demonstrated by simulations.**

I. Introduction

Classical sequential detection is the problem of detecting changes in the distribution of data collected sequentially over time [1]. In a decentralized network setting, the decentralized sequential detection problem concerns with data sequences aggregaged over the network, while sequential detection rules are constrained to the network structure (see, e.g., [2], [3], [4]). The focus was still on a *single* change point variable taking values in (discrete) time. In this paper, our interests lie in sequential detection in a network setting, where multiple change point variables may be simultaneously present.

As an example, quickest detection of traffic jams concerns with multiple potential hotspots (i.e., change points) spatially located across a highway network. A simplistic approach is to treat each change point variables independently, so that the sequential analysis of individual change points can be applied separately. However, it has been shown that accounting for the statistical dependence among the change point variables can provide significant improvement in reducing both false alarm probability and detection delay time [5].

This paper proposes a general probablistic formulation for the multiple change point problem in a network setting, adopting the perspective of probabilistic graphical models for multivariate data [6]. We consider estimating functionals of multiple change points defined globally and locally across the network. The probablistic formulation enables the borrowing of statistical strengh from one network site (associated with a change point variable) to another. We propose a class of sequential detection rules, which can be implemented in a message-passing and distributed fashion across the network. The computation of the proposed sequential rules scales up linearly in both network size and in waiting time, while an approximate version scales up constantly in waiting time. The proposed detection rules are shown to be asymptotically optimal in a Bayesian setting. Interestingly, the expected detection delay can be expressed in terms of Kullback-Leibler divergences defined along edges of the network structure. We provide simulations that demonstrate both statistical and computational efficiency of our approach.

Related Work. The rich statistical literature on sequential analysis tends to focus almost entirely on the inference of a single change point variable [1]. There are recent formulations for sequential diagnosis of a single change point, which may be associated with multiple causes [7], or multiple sequences [8]. Another approach taken in [9] considers a change propagating in a Markov fashion across an array of sensors. These are interesting directions but the focus is still on detecting the onset of a single event. Graphical models have been considered for distributed learning and decentralized detection before, but not in the sequential setting [10], [11]. This paper follows the line of work of [5], [12], but our formulation based on graphical models is more general, and we impose less severe constraints on the amount of information that can be exchanged across network sites.

II. Graphical model for multiple change points

In this section, we shall formulate the multiple change point detection problem, where the change point variables and observed data are linked using a graphical model. Consider a sensor network with d sensors, each of which is associated with a random variable $\lambda_j \in \mathbb{N}$, for $j \in [d] := \{1, 2, \ldots, d\}$, representing a *change point*, the time at which a sensor fails to function properly. We are interested in detecting these change points as accurately and as early as possible, using the data that are associated with (e.g., observed by) the sensors. Taking a Bayesian approach, each λ_j is independently endowed with a prior distribution $\pi_j(\cdot)$.

A central ingredient in our formalism is the notion of a *statistical graph*, denoted as $G = (V, E)$, which specifies the probabilistic linkage between the change point variables and observed data collected in the network. The vertex set of the graph, $V = [d]$ represents the indices of the change point variables λ_j. The edge set E represents pairings of change point variables, $E = \{e = \{s_1, s_2\} \mid s_1, s_2 \in V\}$. With each vertex and each edge, we associate a sequence of *observation* variables,

$$\mathbf{X}_j = (X_j^1, X_j^2, \ldots), \quad j \in V, \tag{1}$$

$$\mathbf{X}_e = (X_e^1, X_e^2, \ldots), \quad e \in E, \tag{2}$$

where the superscript denotes the time index. The $\mathbf{X}_j$ models the private information of node j, while $\mathbf{X}_e$ models the shared information of nodes connected by e. We will use the notation $\mathbf{X}_j^n = (X_j^1, \ldots, X_j^n)$ and similarly for $\mathbf{X}_e^n$; notice the distinction between X_j^n, the observation at time n, versus bold $\mathbf{X}_j^n$, the observations up to time n, both at node j. The aggregate of all the observations in the network is denoted as $\mathbf{X}_* = (\mathbf{X}_j, j \in V, \mathbf{X}_e, e \in E)$. Similarly, $\mathbf{X}_*^n$ represents all the observations up to time n. We will also use $\lambda_* = (\lambda_j, j \in V)$.

The joint distribution of λ_* and $\mathbf{X}_*^n$ is given by a graphical model,

$$P(\lambda_*, \mathbf{X}_*^n) = \prod_{j \in V} \pi_j(\lambda_j) \prod_{j \in V} P(\mathbf{X}_j^n | \lambda_j) \prod_{e \in E} P(\mathbf{X}_e^n | \lambda_{s_1}, \lambda_{s_2}). \tag{3}$$

Given $\lambda_j = k$, we assume $X_j^1, \ldots, X_j^{k-1}$ to be i.i.d. with density g_j and $X_j^k, X_j^{k+1}, \ldots$ to be i.i.d. with density f_j. Given $(\lambda_{s_1}, \lambda_{s_2})$, we assume that the distribution of $\mathbf{X}_e^n$ only depends on $\lambda_e := \lambda_{s_1} \wedge \lambda_{s_2}$, the minimum of the two change points; hence we often write $P(\mathbf{X}_e^n | \lambda_e)$ instead of $P(\mathbf{X}_e^n | \lambda_{s_1}, \lambda_{s_2})$. Given $\lambda_e = k$, $X_e^1, \ldots, X_e^{k-1}$ are i.i.d. with density g_e and $X_e^k, X_e^{k+1}, \ldots$ are i.i.d. with density f_e. All the densities are assumed to be with respect to some underlying measure μ. These specifications can be summarized as,

$$P(\mathbf{X}_j^n | \lambda_j) = \prod_{t=1}^{k-1} g_j(X_j^t) \prod_{t=k}^{n} f_j(X_j^t) \tag{4}$$

and similarly for $P(\mathbf{X}_e^n | \lambda_e)$. We will assume the prior on λ_j to be geometric with parameter $\rho_j \in (0,1)$, i.e. $\pi_j(k) := (1-\rho_j)^{k-1}\rho_j$, for $k \in \mathbb{N}$. Note that these change point variables are dependent a posteriori, despite being independent a priori.

A. Sequential rules and optimality

Although our primary interest is in sequential estimation of the change points $\lambda_* = (\lambda_j)$, we are in general interested in the following functionals,

$$\phi := \phi(\lambda_*) := \lambda_{\mathcal{S}} := \min_{j \in \mathcal{S}} \lambda_j. \tag{5}$$

for some subset $\mathcal{S} \subset [d]$. Examples include a single change point $\mathcal{S} = \{j\}$, the earliest among a pair $\mathcal{S} = \{i, j\}$ and the earliest in the entire network $\mathcal{S} = [d]$. Let $\mathcal{F}_n = \sigma(\mathbf{X}_*^n)$ be the σ-algebra induced by the sequence $\mathbf{X}_*^n$. A sequential detection rule for ϕ is formally a stopping time τ with respect to filtration $(\mathcal{F}_n)_{n \geq 0}$. To emphasize the subset $\mathcal{S}$, we will use $\tau_{\mathcal{S}}$ to denote a rule when the functional $\phi = \lambda_{\mathcal{S}}$. For example τ_1 is a detection rule for λ_1 and τ_{12} is a rule for $\lambda_{12} = \lambda_1 \wedge \lambda_2$.

In choosing τ, there is a trade-off between the false alarm probability $\mathbb{P}(\tau \leq \phi)$ and the detection delay $\mathbb{E}(\tau - \phi)_+$. Here, we adopt the Neyman-Pearson setting to consider all stopping rules for ϕ, having false alarm at most α,

$$\Delta_\phi(\alpha) := \{\tau : \ \mathbb{P}(\tau \leq \phi) \leq \alpha\}, \tag{6}$$

and pick a rule in Δ_ϕ that has minimum detection delay. It is worth mentioning that there are non-Bayesian optimality criteria for the single change point problem, e.g. [13], and it would be an interesting direction to study our multiple change point model in such settings.

B. Communication graph and message passing (MP)

Another ingredient of our formalism is the notion of a *communication graph* representing constraints under which the data can be transmitted across network to compute a particular stopping rule, say τ_j. In general, such a rule depends on all the aggregated data $\mathbf{X}_*^n$. We are primarily interested in those rules that can be implemented in a distributed fashion by passing messages from one sensor only to its neighbors in the communication graph. Although, conceptually, the statistical graph and communication graphs play two distinct roles, they usually coincide in practice and this will be assumed throughout this paper. See Fig. 1 for an illustration.

III. Asymptotically optimal MP rules

We suspect that it is not feasible to derive strictly optimal sequential stopping rules in closed from (say by stochastic dynamic programming) for the multiple change point problem introduced earlier. More crucially, even if such rules are obtained, they are not computationally tractable for large networks, due to the exponential complexity of the state-space. In this section, we shall present a class of detection rules that scale linearly in the size of the network, d, and can be implemented in a distributed fashion by message passing.

Consider the following posterior probabilities

$$\gamma_{\mathcal{S}}^n(k) := \mathbb{P}(\lambda_{\mathcal{S}} = k \mid \mathbf{X}_*^n), \tag{7}$$

$$\gamma_{\mathcal{S}}^n[n] := \mathbb{P}(\lambda_{\mathcal{S}} \leq n \mid \mathbf{X}_*^n) = \sum_{k=1}^{n} \gamma_{\mathcal{S}}^n(k). \tag{8}$$

We propose to stop at the first time $\gamma_{\mathcal{S}}^n[n]$ goes above a threshold,

$$\tau_{\mathcal{S}} = \inf\{n \in \mathbb{N} : \ \gamma_{\mathcal{S}}^n[n] \geq 1 - \alpha\} \tag{9}$$

where α is the maximum tolerable false alarm. It is easily verified that these rules have a false alarm at most α.

Lemma 1. *For $\phi = \lambda_{\mathcal{S}}$, the rule $\tau_{\mathcal{S}} \in \Delta_\phi(\alpha)$.*

More interestingly, we will show that $\tau_{\mathcal{S}}$ is asymptotically optimal for detecting $\lambda_{\mathcal{S}}$. To do so, let us extend the edge set to $\widetilde{E} := E \cup \{\{j\} : j \in V\}$. This allows us to treat the private data associated with node j, i.e. $\mathbf{X}_j$, as (shared) data associated with a self-loop in the graph $(V, \widetilde{E})$. For any $e \in \widetilde{E}$, let $I_e := \int f_e \log \frac{f_e}{g_e} d\mu$ be the KL divergence between f_e and g_e. For $\phi = \lambda_{\mathcal{S}}$, let

$$I_\phi := \sum_{e \subset \mathcal{S}} I_e \tag{10}$$

where the sum runs over all $e \in \widetilde{E}$ which are subsets of $\mathcal{S}$. For example, for a chain graph on $\{1,2,3\}$ with node 2 in the middle, $\widetilde{E} = \{\{1,2\}, \{2,3\}, \{1\}, \{2\}, \{3\}\}$ and we have $I_{\lambda_{12}} := I_1 + I_2 + I_{12}$ while $I_{\lambda_{13}} := I_1 + I_3$. (Here, we abuse notation to write I_{12} instead of $I_{\{1,2\}}$ and so on.)

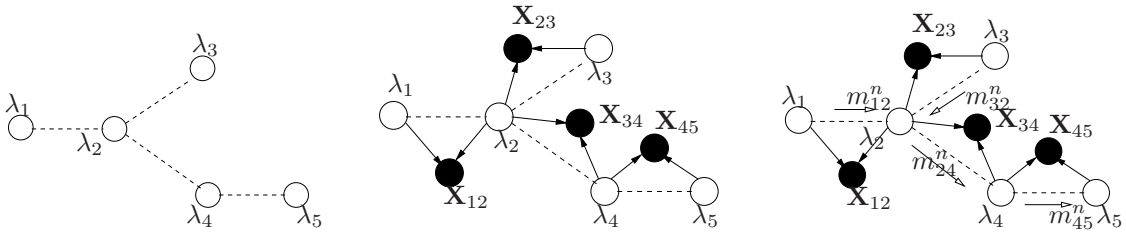

Fig. 1. Left panel illustrates a statistical graph, which induces a graphical model in the middle panel. Right panel illustrates statistical messages passed at time n along some edges in a communication graph (which coincides with statistical graph in this case).

Recall the geometric prior on λ_j (with parameter ρ_j) and the definition of $\phi = \lambda_{\mathcal{S}}$ as the minimum of $\lambda_j, j \in \mathcal{S}$. Then, ϕ is geometrically distributed a priori with parameter $1 - e^{-q_\phi} := 1 - \prod_{j\in\mathcal{S}}(1-\rho_j)$. We can now state our main result on asymptotic optimality.

Theorem 1. *Assume* $|\log \frac{f_e}{g_e}| \le M$ *for all* $e \in \widetilde{E}$. *Then,* $\tau_{\mathcal{S}}$ *is asymptotically optimal for* $\phi = \lambda_{\mathcal{S}}$*; more specifically, as* $\alpha \to 0$,

$$\begin{aligned}\mathbb{E}\big[\tau_{\mathcal{S}} - \phi \mid \tau_{\mathcal{S}} \ge \phi\big] &= \frac{|\log\alpha|}{q_\phi + I_\phi}(1+o(1)) \\ &= \inf_{\widetilde{\tau}\in\Delta_\phi(\alpha)} \mathbb{E}\big[\widetilde{\tau} - \phi \mid \widetilde{\tau} \ge \phi\big].\end{aligned}$$

Remark 1. A notable feature of this result is the decomposition (10) of information along the edges of the graph. For example, in the case of a paired delay $\phi = \lambda_{12}$, for which the information $I_\phi = I_1 + I_2 + I_{12}1_{\{\{1,2\}\in E\}}$ increases (hence the asymptotic delay decreases) if there is an edge between nodes 1 and 2. This has no counterpart in the classical theory where one looks at change points independently.

Remark 2. Another feature of the result is observed for a single delay, say $\phi = \lambda_1$, where one has $I_\phi = I_1$ regardless of whether there is an edge between nodes 1 and 2. Thus, the asymptotic delay for the threshold rule which bases its decision on the posterior probability of λ_1 given all the data in the network ($\mathbf{X}_*^n$) is the same as the one which bases its decision on the posterior given only private data of node 1 ($\mathbf{X}_1^n$). Although this rather counter-intuitive result holds asymptotically, the simulations show that even for moderately low values of α, having access to extra information in $\mathbf{X}_{12}^n$ does indeed improve performance as one expects. (cf. Section VI).

IV. Exact message passing algorithm

It is relatively simple to adapt the well-established belief propagation algorithm, also known as sum-product, to the graphical model (3). The algorithm produces exact values of the posterior $\gamma_{\mathcal{S}}^n$, as defined in (7), in the cases where G is a polytree (and provides a reasonable estimate otherwise.) In this section, we provide the details for $\mathcal{S} = \{j\}$ or $\mathcal{S} = \{i,j\} \in E$.

One issue in adapting the algorithm is the possible infinite support of $\gamma_{\mathcal{S}}^n$. Thanks to a "constancy" property of the likelihood, it is possible to lump all the states after n when computing $\gamma_{\mathcal{S}}^n[n]$.

Lemma 2. *Let* $\{i_1, i_2, \dots, i_r\} \subset [d]$ *be a distinct collection of indices. The function*

$$(k_1, k_2, \dots, k_r) \mapsto P(\mathbf{X}_*^n | \lambda_{i_1} = k_1, \lambda_{i_2} = k_2, \dots, \lambda_{i_r} = k_r)$$

is constant over $\{n+1, n+2, \dots\}^r$.

The algorithm is invoked at each time step n, by passing messages between nodes according to the following protocol: a node sends a message to one of its neighbors (in G) when and only when it has received messages from all its other neighbors. Message passing continues until any node can be linked to any other node by a chain of messages, assuming a connected graph. For a tree, this is usually achieved by designating a node as root and passing messages from the root to the leaves and then backwards.

The message that node j sends to its neighbor i, at time n, is denoted as $m_{ji}^n = [m_{ji}^n(1), \dots, m_{ji}^n(n+1)] \in \mathbb{R}^{n+1}$ and computed as

$$m_{ji}^n(k) = \sum_{k'=1}^{n+1} \Big\{ \widetilde{\pi}_j(k') P(\mathbf{X}_j^n | k') P(\mathbf{X}_{ij}^n | k \wedge k') \prod_{r\in\partial j\setminus\{i\}} m_{rj}^n(k') \Big\}$$

for $k \in [n+1]$, where $\widetilde{\pi}_j(k) := \pi_j(k)$ for $k \in [n]$ and $\widetilde{\pi}_j(n+1) := \pi_j[n]^c = \sum_{k=n+1}^\infty \pi_j(k)$, and ∂j is the neighborhood set of j. Once the message passing ends, γ_j^n and γ_{ij}^n are readily available. We have

$$\gamma_j^n(k) \propto \widetilde{\pi}_j(k) P(\mathbf{X}_j^n | k) \prod_{r\in\partial j} m_{rj}^n(k), \quad k \in [n]. \tag{11}$$

It also holds for $k = n+1$ if the LHS is interpreted as $\gamma_j^n[n]^c$. Similarly, when $\{i,j\} \in E$, $P(\lambda_i = k_1, \lambda_j = k_2 | \mathbf{X}_*^n)$ is proportional to

$$\begin{aligned}&\widetilde{\pi}_i(k_1)\widetilde{\pi}_j(k_2) P(\mathbf{X}_i^n|k_1) P(\mathbf{X}_j^n|k_2) P(\mathbf{X}_{ij}^n | k_1 \wedge k_2) \\ &\qquad \times \prod_{r\in\partial i\setminus\{j\}} m_{ri}^n(k_1) \prod_{r\in\partial j\setminus\{i\}} m_{rj}^n(k_2)\end{aligned}$$

for $(k_1, k_2) \in [n]^2$, from which γ_{ij}^n can be computed.

Lemma 3. *When* G *is a tree, the message passing algorithm produces correct values of* γ_j^n *and* γ_{ij}^n *at time step* n*, with computational complexity* $O((|V| + |E|)n)$.

V. Fast approximation algorithm

We now turn to an approximate message passing algorithm which, at each time step, has computational complexity $O(|V| + |E|)$. Let us define binary variables

$$Z_j^n = 1\{\lambda_j \le n\}, \quad Z_*^n = (Z_1^n, \dots, Z_d^n). \tag{12}$$

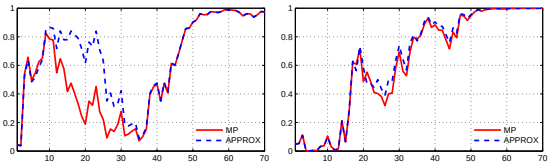

Fig. 2. Examples of posterior paths, $n \mapsto \gamma^n[n]$, obtained by exact (MP) and approximate (APPROX) message passing.

The idea is to compute $P(Z_*^n|\mathbf{X}_*^n) = P(Z_*^n|X_*^n, \mathbf{X}_*^{n-1})$ recursively based on $P(Z_*^{n-1}|\mathbf{X}_*^{n-1})$. The former is proportional (in Z_*^n) to $P(Z_*^n, X_*^n|\mathbf{X}_*^{n-1})$ and we have

$$P(Z_*^n, X_*^n|\mathbf{X}_*^{n-1}) = \prod_{j\in V} P(X_j^n|Z_j^n) \prod_{\{i,j\}\in E} P(X_{ij}^n|Z_i^n, Z_j^n)\, P(Z_*^n|\mathbf{X}_*^{n-1}). \tag{13}$$

Let $u_e(z;\xi) := [g_e(\xi)]^{1-z}[f_e(\xi])^z$ for $e \in \widetilde{E}$, $z \in \{0,1\}$. Then, $P(X_j^n|Z_j^n) = u_j(Z_j^n; X_j^n)$, and $P(X_{ij}^n|Z_i^n, Z_j^n) = u_{ij}(Z_i^n \vee Z_j^n; X_{ij}^n)$. It remains to express $P(Z_*^n|\mathbf{X}_*^{n-1})$ in terms of $P(Z_*^{n-1}|\mathbf{X}_*^{n-1})$. This is possible at a cost of $O(2^{|V|})$, but we omit the details for brevity. To obtain a fast algorithm (i.e., $O(\text{poly}(|V|))$), we instead approximate

$$P(Z_*^n|\mathbf{X}_*^{n-1}) \approx \prod_{j\in V} P(Z_j^n|\mathbf{X}_*^{n-1}) = \prod_{j\in V} \nu(Z_j^n; \gamma_j^{n-1}[n]), \tag{14}$$

where $\nu(z;\beta) := \beta^z(1-\beta)^{1-z}$. By constancy Lemma 2, Bayes rule and algebra, we get the recursion

$$\gamma_j^{n-1}[n] = \frac{\pi_j(n)}{\pi_j[n-1]^c} + \frac{\pi_j[n]^c}{\pi_j[n-1]^c}\gamma_j^{n-1}[n-1].$$

Thus, at time n, the RHS of (14) is known based on values computed at time $n-1$ (with initial value $\gamma_j^0[0] = 0, j \in V$). Inserting this RHS into (13) in place of $P(Z_*^n|\mathbf{X}_*^{n-1})$, we obtain a graphical model in variables Z_*^n (instead of λ_*) which has the same form as (3) with $\nu(Z_j^n; \gamma_j^{n-1}[n])$ playing the role of the prior $\pi(\lambda_j)$.

Hence, we can apply a message passing algorithm similar to that described in Section IV, to marginalize this approximate version of $P(Z_*^n, X_*^n|\mathbf{X}_*^{n-1})$, providing approximate values of $\gamma_j^n[n] = P(Z_j^n = 1|\mathbf{X}_*^n)$ and $\gamma_{ij}^n[n]$. The message update equations are similar to those of Section IV and are omitted. The difference is that the messages are now binary and do not grow in size with n. Fig. 2 shows examples of posterior tracking by approximate algorithm. Theoretical analysis of the algorithm will appear in a longer version of this paper.

VI. Simulations

We present simulation results as depicted in Fig. 3. The setting is that of graphical model (3) on $d = 4$ nodes, where the statistical graph is a star with node 2 in the middle. Conditioned on λ_*, all the data sequences, $\mathbf{X}_*$, are assumed Gaussian of variance 1, with pre-change mean 1 and post-change mean zero. All priors are geometric with parameters $\rho_j = 0.1$. Fig. 3 shows plots of expected delay over $|\log\alpha|$, against $|\log\alpha|$, for

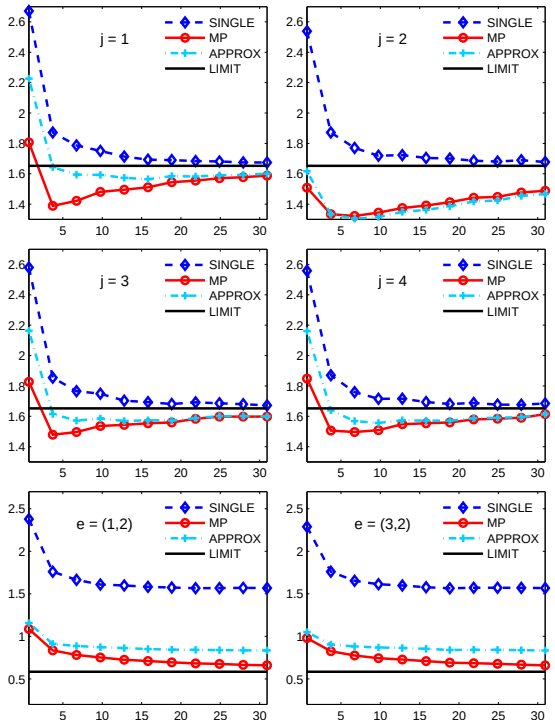

Fig. 3. Plots of the slope $\frac{1}{-\log\alpha}\mathbb{E}[\tau_S - \phi|\tau_S \geq \phi]$ against $-\log\alpha$ for message-passing algorithm (MP), approximate algorithm (APPROX) and SINGLE algorithm which disregards shared information. The graph is the star graph of 4 nodes with node 2 in the center. Estimates of both single and paired change points (λ_j and λ_{ij}) are shown together with theoretical limit of Theorem 1. (The case $e = (4,2)$ is omitted to conserve space; it looks very similar to $e = (1,2), (3,2)$). False alarm tolerance α ranges in $[0.5, 10^{-13}]$.

three methods: the message-passing algorithm of Section IV (MP), approximate algorithm of Section V (APPROX) and the method which bases its inference on posteriors calculated based only on each node's private information (SINGLE). This latter method estimates a single change point λ_j by $\widehat{\tau}_j := \inf\{n : P(\lambda_j \leq n|\mathbf{X}_j^n) \geq 1-\alpha\}$ and a paired $\lambda_{ij} = \lambda_i \wedge \lambda_j$ by $\widehat{\tau}_i \wedge \widehat{\tau}_j$. Also shown in the figure is the limiting value of the normalized expected delay as predicted by Theorem 1. All plots are generated by Monte Carlo simulation over 5000 realizations.

In estimating single change points, MP, which takes shared information into account, has a clear advantage over SINGLE, for high to relatively low false alarm values (even, say, around $\alpha \approx e^{-5}$); though, both methods seem to converge to the same slope in the $\alpha \to 0$ limit, as suggested by Theorem 1. (The particular value is $(-\log 0.9 + 0.5)^{-1} = 1.6519$.) Also note that the advantage of MP over SINGLE is more emphasized for node 2, as expected by its access to shared information from all the three nodes. We also note that APPROX does a reasonable job at approximating MP, with delays between those of SINGLE and MP, getting closer to MP as $\alpha \to 0$.

For paired change points, the advantage of MP and APPROX over SINGLE is more emphasized. It is also interesting to note that while MP seems to converge to the expected theoretical limit $(-2\log 0.9 + 3\cdot 0.5)^{-1} = 0.5845$, SINGLE

seems to converge to a higher slope (with a reasonable guess being 1.6519 as in the case of single change points).

In regard to false alarm probability, nonzero values were only observed for the first few values of α considered here, and those were either below or very close to the specified tolerance.

Appendix A
Proof sketch of Theorem 1

We provide a proof sketch for the case $d = 2$ here (the full proof is quite technical, and can be found in [14]). Fix some $\phi = \tau_{\mathrm{S}}$ and consider the likelihood ratio

$$D_\phi^k(\mathbf{X}_*^n) := \frac{P(\mathbf{X}_*^n \mid \phi = k)}{P(\mathbf{X}_*^n \mid \phi = \infty)}.$$

Let $\mathbb{P}_\phi^k = \mathbb{P}(\cdot \mid \phi = k)$. Our asymptotic analysis hinges on the asymptotic behavior of $\frac{1}{n}\log D_\phi^k(\mathbf{X}_*^n)$, as $n \to \infty$, under probability measure $\mathbb{P}_\phi^k$. In particular, building on the results of [15], it is straightforward to derive the following sufficient condition. Let $\mathbb{P}_{\lambda_1,\lambda_2}^{m_1,m_2}$ be the probability measure conditioned on $\lambda_1 = m_1$ and $\lambda_2 = m_2$. Also, let $\pi_\phi^k(m_1, m_2) := \mathbb{P}_\phi^k(\lambda_1 = m_1, \lambda_2 = m_2)$. Suppose that for all (m_1, m_2) with positive probability under $\pi_\phi^k(m_1, m_2)$, we can show the "concentration inequality"

$$\mathbb{P}_{\lambda_1,\lambda_2}^{m_1,m_2}\Big(\Big|\frac{1}{n}\log D_\phi^k(\mathbf{X}_*^n) - I_\phi\Big| > \varepsilon\Big) \le q(n)\exp(-c_1 n\varepsilon^2)$$

for all $n \ge \frac{1}{\varepsilon}p(m_1, m_2, k)$ for polynomials $p(\cdot)$ and $q(\cdot)$. Then, if both $\pi_\phi^k(\cdot,\cdot)$ and $\mathbb{P}(\phi = \cdot)$ have finite polynomial moments, which is the case for our geometric priors, conclusions of Theorem 1 hold for ϕ. We say that $|n^{-1}\log D_\phi^k(\mathbf{X}_*^n) - I_\phi| \le \varepsilon$ holds with high probability, abbreviated w.h.p.

Next, we define $R_p^n(e) := R_p^n(\mathbf{X}_e) := \prod_{t=p}^n \frac{f_e(\mathbf{X}_e)}{g_e(\mathbf{X}_e)}$ if $e \in \widetilde{E}$ and $p \le n$, and $R_p^n(\mathbf{X}_e) = 1$ otherwise. Similarly let I_e be defined as in (10) if $e \in \widetilde{E}$ and $I_e = 0$ otherwise. We note that

$$D_\phi^k(\mathbf{X}_*^n) = \frac{\sum_{k_1,k_2}\pi_\phi^k(k_1,k_2)\, R_{k_1}^n\{1\}\, R_{k_1}^n\{2\}\, R_{k_1\wedge k_2}^n\{12\}}{\sum_{k_1,k_2}\pi_\phi^\infty(k_1,k_2)\, R_{k_1}^n\{1\}\, R_{k_1}^n\{2\}\, R_{k_1\wedge k_2}^n\{12\}},$$

with some abuse of notation. In addition, for a collection $\mathcal{E} = \{e_1, \ldots, e_J\}$, define

$$S_m^{q,n}(\mathcal{E}) = \sum_{p=m}^{q} Ae^{-\beta p}\prod_{e\,\in\,\mathcal{E}} R_p^n(e)$$

for some $A, \beta > 0$, and let $I_{\mathcal{E}} := \sum_{e\in\mathcal{E}} I_e$.

We will say that $a_n \overset{\varepsilon}{\asymp} b_n$ if $|a_n - b_n| \le \varepsilon$ for $n \ge \frac{c_0}{\varepsilon}$. This relation is transitive and stable under addition, subtraction and taking maximum. Furthermore,

$$n^{-1}\log(a_n + b_n) \overset{\varepsilon}{\asymp} \max\{n^{-1}\log a_n,\, n^{-1}\log b_n\}. \quad (15)$$

At the heart of the proof are two concentration inequalities:

$$\frac{1}{n}\log R_m^n(e) \overset{\varepsilon}{\asymp} I_e, \quad \frac{1}{n}\log S_m^{n,n}(\mathcal{E}) \overset{\varepsilon}{\asymp} I_{\mathcal{E}}, \quad (16)$$

where the first holds conditioned on $\lambda_e \le m$ w.h.p. and the second conditioned on $\lambda_e \le m$ for all $e \in \mathcal{E}$, w.h.p.

Let us focus on the case $\phi = \lambda_{12}$. For $k < \infty$, $\pi_\phi^k(k_1, k_2) \propto \overline{\rho}_1^{k_1}\overline{\rho}_2^{k_2} 1_{\{k_1\wedge k_2 = k\}}$ where $\overline{\rho}_j := 1 - \rho_j$. Let

$$\mathcal{T}_1 := R_k^n\{1\}[S_{k+1}^{n,n}\{2\} + Q_{2,n}],$$

where $Q_{2,n} := \sum_{k_2>n}\pi_\phi^k(k, k_2) \propto \overline{\rho}_2^n$ (symmetrically for $\mathcal{T}_2$ with roles of 1 and 2 reversed) and $\mathcal{T}_0 := \pi_\phi^k(k,k)R_k^n\{1\}R_k^n\{2\}$. Then, $D_\phi^k(\mathbf{X}_*^n) = R_k^n\{12\}\sum_{i=0}^2 \mathcal{T}_i$. Now, condition on $\lambda_1 = k$ and $\lambda_2 = r \ge k$ so that $\lambda_{12} = k$. (The other case with roles of 1 and 2 reversed follows by symmetry.) By (15),

$$n^{-1}\log D_\phi^k(\mathbf{X}_*^n) \overset{\varepsilon}{\asymp} n^{-1}\log R_k^n\{12\} + \max_{i=0,1,2} n^{-1}\log \mathcal{T}_i$$

Consider $\mathcal{T}_1$ and note that $n^{-1}\log\mathcal{T}_1 \overset{\varepsilon}{\asymp} n^{-1}\log R_k^n\{1\} + \max\{\frac{1}{n}\log S_{k+1}^{n,n}\{2\}, \log\overline{\rho}_2\}$ where the first term is $\overset{\varepsilon}{\asymp} I_1$ w.h.p. by (16). Similarly, $n^{-1}\log S_{k+1}^{n,n}\{2\}$ equals

$$n^{-1}\log\big(S_{k+1}^{r,r}\{2\}R_r^n\{2\} + S_{r+1}^{n,n}\{2\}\big) \overset{\varepsilon}{\asymp} I_2$$

and since $\log\overline{\rho}_1 < 0$, we have $n^{-1}\log\mathcal{T}_1 \overset{\varepsilon}{\asymp} I_1 + I_2$. Other terms are dealt with similarly, and we get, w.h.p.

$$\frac{1}{n}\log D_\phi^k(\mathbf{X}_*^n) \overset{\varepsilon}{\asymp} I_{12} + \max\{I_1 + I_2, I_2 + I_1, I_1 + I_2\}.$$

The case $\phi = \lambda_1$ follows along similar lines.

References

[1] T. L. Lai, "Sequential analysis: Some classical problems and new challenges (with discussion)," *Statist. Sinica*, vol. 11, pp. 303–408, 2001.

[2] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decentralized sequential detection with a fusion center performing the sequential test," *IEEE Trans. Info. Theory*, vol. 39, no. 2, pp. 433–442, 1993.

[3] Y. Mei, "Asymptotic optimality theory for decentralized sequential hypothesis testing in sensor networks," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2072–2089, 2008.

[4] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "On optimal quantization rules in some problems in sequential decentralized detection," *IEEE Transactions on Information Theory*, vol. 54(7), pp. 3285–3295, 2008.

[5] R. Rajagopal, X. Nguyen, S. Ergen, and P. Varaiya, "Distributed online simultaneous fault detection for multiple sensors," in *Proc. of 7th Int'l Conf. on Info. Proc. in Sensor Networks (IPSN)*, April 2008.

[6] M. I. Jordan, "Graphical models," *Statistical Science*, vol. 19, pp. 140–155, 2004.

[7] S. Dayanik, C. Goulding, and H. V. Poor, "Bayesian sequential change diagnosis," *Mathematics of Operations Research*, vol. 33, no. 2, pp. 475–496, 2008.

[8] Y. Xie and D. Siegmund, "Sequential multi-sensor change-point detection," in *Joint Statistical Meeting*, 2011.

[9] V. Raghavan and V. V. Veeravalli, "Quickest change detection of a markov process across a sensor array," *IEEE Transactions on Information Theory*, vol. 56(4), pp. 1961–1981, 2010.

[10] M. Cetin, L. Chen, J. W. Fisher III, A. Ihler, R. Moses, M. Wainwright, and A. Willsky, "Distributed fusion in sensor networks: A graphical models perspective," *IEEE Signal Processing Magazine*, vol. July, pp. 42–55, 2006.

[11] O. P. Kreidl and A. Willsky, "Inference with minimum communication: a decision-theoretic variational approach," in *NIPS*, 2007.

[12] R. Rajagopal, X. Nguyen, S. Ergen, and P. Varaiya, "Simultaneous sequential detection of multiple interacting faults," *http://arxiv.org/abs/1012.1258*, 2010.

[13] A. G. Tartakovsky and M. Pollak, "Nearly minimax changepoint detection procedures," in *ISIT*, 2011.

[14] A. A. Amini and X. Nguyen, "Sequential detection of multiple change points in networks: A graphical model based approach," Department of Statistics, University of Michigan, Tech. Rep., 2012.

[15] A. G. Tartakovsky and V. V. Veeravalli, "General asymptotic bayesian theory of quickest change detection," *Theory Probab. Appl.*, vol. 49, no. 3, pp. 458–497, 2005.