Active Sampling for the Quickest Detection of Markov Networks

Ali Tajer, *Senior Member, IEEE*, Javad Heydari, *Member, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*— Consider n random variables forming a Markov random field (MRF). The true model of the MRF is unknown, and it is assumed to belong to a binary set. The objective is to sequentially sample the random variables (one-at-a-time) such that the true MRF model can be detected with the fewest number of samples, while in parallel, the decision reliability is controlled. The core element of an optimal decision process is a rule for selecting and sampling the random variables over time. Such a process, at every time instant and adaptively to the collected data, selects the random variable that is expected to be most informative about the model, rendering an overall minimized number of samples required for reaching a reliable decision. The existing studies on detecting MRF structures generally sample the entire network at the same time and focus on designing optimal detection rules without regard to the data-acquisition process. This paper characterizes the sampling process for general MRFs, which is shown to be optimal in the asymptote of large n. The critical insight in designing the sampling process is devising an information measure that captures the decisions' inherent statistical dependence over time. Furthermore, when the MRFs can be modeled by acyclic probabilistic graphical models, the sampling rule is shown to take a computationally simple form. Performance analysis for the general case is provided, and the results are interpreted in several special cases: Gaussian MRFs, non-asymptotic regimes, Chernoff's rule for controlled (active) sensing, and the problem of cluster detection.**

***Index Terms*— Active sampling, controlled sensing, correlation detection, Markov network, quickest detection.**

I. Introduction

A. Overview

DRIVEN by advances in information sensing and acquisition, many application domains have evolved towards interconnected networks of information sources in which large-scale and complex data is constantly generated and processed for various inferential and decision-making purposes. Induced by their physical couplings, such information sources generate data streams that often bear strong statistical dependence structures. Probabilistic graphical models, in general, and Markov random fields (MRFs), in particular, provide effective analytical frameworks for encoding the statistical relationship among the datasets generated by different agents in a network [1]–[4].

Forming inferential decisions in an MRF strongly hinges on determining the dependence structure embedded in the MRF. There are two distinct aspects to determining an MRF structure: selecting (estimating) versus differentiating (detecting) the models. In **model selection** (structure learning), the objective is to sample the random variables that form an MRF, and select (estimate) the edge set of the graphical model associated with the MRF (a representative list of the existing approaches includes [5]–[18]). While graph structure learning is NP-hard in its general form [19], it becomes feasible under proper restrictions on the structure of the graph, e.g., limiting the graph to the classes of sparsely-connected graphs, edge-bounded graphs, and degree-bounded graphs. There is rich literature investigating the algorithmic and information-theoretic aspects of structure learning, especially for Gaussian and Ising graphical models. The existing studies can be distinguished based on the sampling mechanisms that they adopt. Broadly, there exists two distinct approaches to sampling: (i) pre-specific sampling, in which sampling is agnostic to the data and follows pre-specified rules [5]–[12], and (ii) active sampling, in which the sampling decisions are data-driven and they are updated dynamically as the data is collected [13]–[18]. In active sampling methods, sampling and model selection processes are inherently coupled, and their emphasis is on co-designing these two processes. In contrast, when the sampling mechanism is pre-specified, the sampling and model selection processes are decoupled, and the emphasis is placed on forming reliable decisions given a set of samples.

In contrast to model selection, in **model detection**, the unknown model of an MRF is assumed to belong to a finite set of known models, and the objective is to sample the random variables in order to identify the true model. MRF model detection, in its simplest form, is used for deciding whether a given set of random variables are independent, which is referred to as testing against independence. More generally, dependence model detection is the process of deciding in favor of one dependence model against a group of alternative ones (a representative list of relevant literature

Manuscript received July 15, 2019; revised July 20, 2021; accepted July 25, 2021. Date of publication October 28, 2021; date of current version March 17, 2022. This work was supported in part by the U.S. National Science Foundation under Grant ECCS-1455228, Grant CAREER ECCS-1554482, Grant ECCS-1933107, Grant DMS-1737976, and Grant CCF-1908308. An earlier version of this paper was presented in part at the 53rd Annual Allerton Conference on Communication, Control, and Computing in 2015, and in part at the 2019 IEEE International Symposium on Information Theory. *(Corresponding author: Ali Tajer.)*

Ali Tajer is with the Electrical, Computer, and System Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: tajer@ecse.rpi.edu).

Javad Heydari is with the Advanced AI Lab, LG Electronics, Santa Clara, CA 95054 USA (e-mail: javad.heydari@lge.com).

H. Vincent Poor is with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: poor@princeton.edu).

Communicated by A. Singh, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3124166.

Digital Object Identifier 10.1109/TIT.2021.3124166

includes [20]–[28]). The existing studies on MRF model detection adopt pre-specified sampling mechanisms and focus on forming detection rules. Furthermore, existing studies primarily investigate Gaussian MRFs.

In this paper, we investigate *active* sampling for model detection in general MRFs. The objectives are (i) characterizing the fundamentally minimum number of samples required for forming decisions with target reliability, and (ii) characterizing the attendant sampling and detection rules. Characterizing an optimal active sampling algorithm that can detect the model of an MRF with the minimal number of samples is especially imperative as an MRF's size or dimension grows, in which case sampling incurs substantial communication, sensing, and decision delay costs. An active sampling process in an MRF is specified by the aggregate number of samples to be collected as well as the order in which they are collected. When the order is pre-specified, determining the optimal sampling strategy reduces to minimizing the (average) number of samples. This can be effectively facilitated via sequential hypothesis testing, which is well-investigated. In sequential hypothesis testing, the samples are collected sequentially according to a *pre-specified* order, and the sampling strategy dynamically decides whether to take more samples or to terminate the process and form a decision [29]–[32]. However, incorporating dynamic decisions about the order of sampling introduces a new dimension to decision-making, which is investigated less. Forming such dynamic decisions that pertain to data acquisition naturally arises in a broad range of applications such as sensor management [33], inspection, and classification [34], medical diagnosis [35], cognitive science [36], generalized binary search [37], and channel coding with feedback [38], to name a few.

The contributions of this paper can be summarized as follows. First, we design an active sampling algorithm that guides the data collection over an MRF and specifies the final decision rules. Next, we analyze the performance of the proposed algorithm in terms of decision reliability, error exponents, and average decision delay. For delay analysis, we quantify the expected number of samples required to make a reliable decision in both non-asymptotic and asymptotic regimes. The non-asymptotic analysis provides a tool to determine the feasibility of the problem. In contrast, the asymptotic analysis highlights the optimality of the proposed algorithm in the asymptote of small error rates. Then, we provide some special cases and examples to further illustrate the quantities defined throughout the paper and showcase the effectiveness of the proposed algorithm. Finally, some experiments quantify the gap between our algorithm and the state-of-the-art approaches. We remark that some of the results in this paper have also appeared in [39] and [40].

B. Related Literature

1) Hypothesis Testing of Precision Matrices: For Gaussian MRFs with identical means, determining the covariance matrix (or its inverse, precision matrix) is equivalent to detecting the true model of the data. Testing simple *global* null hypothesis structures is investigated in [41]–[43]. Specifically, the focus of these studies is on the case where the number of samples and the data dimension are comparable and they focus on testing a single global model against the identity covariance matrix. The Gaussian setting is investigated in [41], the sub-Gaussian setting in [42], and the more general setting with moment conditions in [43]. Testing a composite global null hypothesis with only one sample is investigated in [44]–[47]. Testing for equality of two covariance matrices is investigated in [48]–[51], while [46], [47], and [52] study a composite global null hypothesis with a diagonal covariance matrix. A more recent review of these settings is available in [53].

In the context of testing precision matrices, testing for two different correlation matrices investigated in [55] and [56], and testing for two different precision matrices studied in [56], are most relevant to the problem in this paper. The key distinction of these studies with the setting considered in this paper is that they consider the fixed sample-size setting and use only the covariance matrices. In this paper, on the other hand, we focus on the sequential setting, and instead of considering only the covariance matrices, we leverage the entire distribution, which is necessary for the non-Gaussian MRFs.

2) Estimation of Covariance Matrices: In another related direction, model selection is performed via estimating the covariance matrix (or its inverse) of the data [5], [9], [10], [12], [28]. In [28], the temporal correlation of the data is treated as a nuisance parameter. By making a Gaussian assumption, a testing procedure is proposed to identify all the non-zero elements of the precision matrix with guaranteed performance. Estimating sparse covariance matrices via adaptive thresholding is considered in [9]. By adapting the threshold to the variability of individual entries in a data-driven setting, it is shown that, compared to the commonly used universal thresholding estimators, these estimators achieve the optimal rate of convergence over a large class of sparse covariance matrices under the spectral norm and enjoy excellent performance both theoretically and numerically. In [5], [10], and [12] estimating sparse precision matrices is considered, and ℓ_1 minimization is used to solve the problem by using the estimated covariance matrix of the given data. All the studies above consider the problem in the fixed sample-size setting, and their application to model selection is limited to Gaussian distributions.

3) Controlled (Active) Sensing for Detection: One directly applicable approach to treat coupled sampling and decision-making process is *controlled sensing*, originally developed by Chernoff for binary composite hypothesis testing through incorporating a controlled information gathering process that dynamically decides about taking one of a finite number of possible actions at each time [57]. Under the assumption of uniformly distinguishable hypotheses and having independent control actions, Chernoff's rule decides in favor of the action with the best *immediate* return according to proper information measures. It achieves optimal performance in the asymptote of a diminishing rate of erroneous decisions. Chernoff's rule, specifically, at each time, identifies the most likely true hypothesis based on the collected data and takes the action that reinforces that decision.

Extensions of Chernoff's rule to various settings are studied in [58]–[61]. Specifically, the studies in [58] and [59] investigate an extension to accommodate an infinite number of available actions and an infinite number of hypotheses, and [60] and [61] provide alternative rules that are empirically shown to outperform Chernoff's rule in the non-asymptotic regimes. Recent advances in controlled sensing that are relevant to the scope of this paper include [62]–[65]. In [62], Chernoff's rule is modified to relax the assumption that the hypotheses should be uniformly distinguishable in the multi-hypothesis setting. In this modified rule, a randomization policy is introduced into the selection rule such that at certain time instants, it ignores Chernoff's rule and randomly selects one action according to a uniform distribution. This rule is shown to admit the same asymptotic performance as Chernoff's rule. The results are extended to the setting in which the available data belongs to a discrete alphabet and follows a stationary Markov model in [63]. An application of Chernoff's rule to anomaly detection in a dataset is investigated in [64], where it is shown that when facing a finite number of sequences consisting of an anomalous one, Chernoff's rule is asymptotically optimal even without assuming that the hypotheses are distinguishable or exerting randomized actions. The study in [65] imposes a cost on switching among different actions and offers a modification of Chernoff's rule, which randomly decides between repeating the previous action, and a new action based on Chernoff's rule. It achieves the same asymptotic optimality as Chernoff's rule. Similarly, Chernoff's rule is also applied to sparse signal recovery [66], sequential estimation [67], and classification problems [68] and [69].

Besides Chernoff's rule and its variations, there exist alternative strategies admitting certain optimality guarantees. In pioneering studies, [70] and [71] offer a strategy that initially takes a number of sampels according to a pre-designated rule in order to identify the true hypothesis, after which it selects the action that maximizes the information under the identified hypothesis. The study in [72] proposes a heuristic strategy and characterizes the deviation of its average delay from the optimal rule. Other studies have investigated the Bayesian setting [73]–[77]. The study in [73] considers a sequential multi-hypothesis testing problem with multiple control actions for which the optimal strategy is the solution to dynamic programming that is computationally intractable. Hence, it designs two heuristic policies and investigates their non-asymptotic and asymptotic performances. For the same problem, performance bounds and the gains of sequential sampling and optimal data-adaptive selection rules are analyzed in the asymptote of the high cost of erroneous decisions [74]. The study in [75] restricts the samples to be generated by the exponential family distributions and shows that the dimension of the sufficient statistic space is less than both the number of parameters governing the exponential family and the number of hypotheses. Hence, an exactly optimal policy is characterized by only moderate computational complexity. Other heuristic approaches for anomaly detection are also investigated in [76] and [77], which select the action with the minimum immediate effect on the total Bayesian cost and are shown to achieve the same optimality guarantees suggested by Chernoff [57].

Despite their discrepancies in settings and approaches, all the studies above on controlled sensing assume that the available actions are *independent* or follow a first-order stationary Markov process. This is in contrast to the setting of this paper, in which the correlation structure in the generated data under one hypothesis or both induces co-dependence among the control actions. In this paper, we devise a sequential sampling strategy for detecting MRFs, in which the correlation model plays a significant role in forming the sampling decisions. Specifically, the devised selection rule, unlike Chernoff's rule, incorporates the correlation structure into decision-making via accounting for the impact of each action on the future ones and selecting the one with the largest *expected* information under the most likely true hypothesis. The associated optimality guarantees are established, and the specific results for the special case of Gaussian distributions are characterized. The gains of the proposed selection rule are also delineated analytically and numerically.

4) Active Learning for Model Selection: Unlike for model detection, active sampling for model selection (structure learning) is investigated in more depth [13]–[18]. In [13], a model selection problem in a supervised setting is considered, in which active learning is applied in order to identify the set of training examples that should be used to minimize the integrated variance of the model. The studies in [17] and [18] propose active learning algorithms for selecting the structures of MRFs. Their main distinction from our work is that they are concerned with a structure learning problem, while in this paper, the true model is selected from a finite set of candidate models. In [14]–[16], active learning over Bayesian networks is considered. In [14], it is assumed that the graphical model underlying the Bayesian network is known, and the objective is estimating network parameters. The studies in [15], [16], and [78] are concerned with learning the connectivity structure, the parameters, and the direction of the causal relationship among the nodes.

II. Data Model and Problem Formulation

A. Notation

Throughout the paper $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space on which all the probability measures are defined. In this space, consider n random variables $\mathcal{X} \triangleq \{X_1, \ldots, X_n\}$ forming an MRF with respect to an undirected graph $\mathcal{G}(V, E)$ with nodes $V \triangleq \{1, \ldots, n\}$ and the edge set $E \subseteq V \times V$. For any given set $A \subseteq V$, we define $X_A \triangleq \{X_i \ : \ i \in A\}$. Random variables $\mathcal{X}$ satisfy the global Markov property, that is, any two disjoint subsets of random variables are conditionally independent given a separating set, i.e.,

$$X_A \perp\!\!\!\perp X_B \mid X_C, \tag{1}$$

where C separates disjoint sets A and B such that every path between a node in A and a node in B passes through at least one node in C. One immediate result of the global Markov

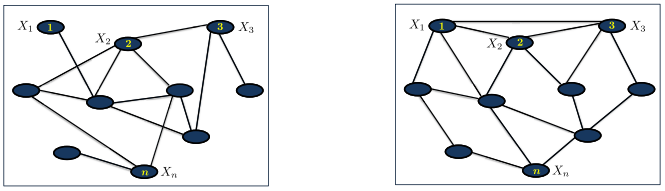

Fig. 1. Data model with two different correlation structures.

property is the pair-wise Markov property, i.e.,

$$\forall\ (i,j) \notin E \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}. \tag{2}$$

The model of the underlying $\mathcal{X}$ is unknown, and it is assumed to obey one of the two possible known models. Detecting the MRF model can be formalized as the solution to the binary hypothesis test:

$$\begin{aligned} \mathsf{H}_0 \ &: \ (X_1, \ldots, X_n) \sim \mathbb{P}_0, \\ \mathsf{H}_1 \ &: \ (X_1, \ldots, X_n) \sim \mathbb{P}_1, \end{aligned} \tag{3}$$

where $\mathbb{P}_0$ and $\mathbb{P}_1$ denote the two known and completely distinct probability measures governing the two models. We denote the undirected dependency graphs associated with joint measures $\mathbb{P}_0$ and $\mathbb{P}_1$ by $\mathcal{G}_0(V, E_0)$ and $\mathcal{G}_1(V, E_1)$, respectively. Figure 1 depicts the graphs associated with the precision matrices of a dichotomous Gaussian MRF model, in which the edges encode the conditional dependency structures. For convenience in notations, we assume that the distributions of the random variables under each hypothesis $\ell \in \{0, 1\}$ are absolutely continuous with respect to a common distribution and have well-defined probability density functions (pdfs). For every non-empty set $A \subseteq V$, we denote the joint pdf of X_A under H_ℓ by $f_\ell(\cdot; A)$. We also define $\mathsf{T} \in \{\mathsf{H}_0, \mathsf{H}_1\}$ as the true hypothesis and denote the prior probability that hypothesis H_ℓ is true by ϵ_ℓ, where $\epsilon_0 + \epsilon_1 = 1$.

B. Sampling Model

We consider a fully sequential data acquisition mechanism, in which we select and sample one node at-a-time. The objective is to identify an optimal sequence of nodes, such that with the minimum number of samples, the true model $\mathsf{T} \in \{\mathsf{H}_0, \mathsf{H}_1\}$ can be discerned. Samples are collected sequentially, such that at any time t and based on the information accumulated up to that time, the sampling procedure takes one of the following actions.

- A_1) *Exploration:* Due to lack of sufficient confidence, making any decision is deferred, and one more sample is taken from another node in the graph. Under this action, the node to be selected should be specified.
- A_2) *Detection:* The data collection process is terminated, and a reliable decision about the true model of the graph is formed. Under this action, the stopping time and the final decision rule upon stopping will be specified.

The sampling process can be expressed uniquely by the data-adaptive rule for selecting the nodes over time, the stopping rule, and the final detection decision rule. To formalize the information-gathering process (exploration), we define $\psi_n : V \to V$, where $\psi_n(t)$ returns the index of the node observed at time t. Accordingly, we define ψ_n^t as the *ordered* set of nodes selected and sampled up to time t, i.e., $\psi_n^t \triangleq \{\psi_n(1), \ldots, \psi_n(t)\}$. We also define φ_n^t as the set of nodes that are remained unobserved prior to time t and can be observed at t, i.e., $\varphi_n^t \triangleq V\setminus\psi_n^{t-1}$. We denote sample collected at time t by $Y_t \triangleq X_{\psi_n(t)}$, and denote the sequence of samples accumulated up to time t by $Y^t \triangleq \big(Y_1, \ldots, Y_t\big)$. The information accumulated sequentially generate a σ-algebra of $\mathcal{F}$ denoted by $\{\mathcal{F}_t : \ t = 1, 2, \ldots\}$, where

$$\mathcal{F}_t \triangleq \sigma(Y^t; \psi_n^t). \tag{4}$$

We define $\tau_n \in \mathbb{N}$ as the Markov stopping time with respect to the family $\{\mathcal{F}_t\}$, at which the sampling process is terminated and a decision is formed. We also define $\delta_n \in \{0, 1\}$ as an $\mathcal{F}_t$-measurable function as the terminal decision rule, where $\delta_n = \ell$ indicates accepting hypothesis H_ℓ, for $\ell \in \{0, 1\}$. We define the tuple $\Phi_n \triangleq (\tau_n, \delta_n, \psi_n^{\tau_n})$ to uniquely specify the sampling strategy[1] and the decision rules involved. Finally, we define two information measures that are instrumental in formalizing and analyzing various decision rules throughout the paper. Specifically, for any given ψ_n^t and $A \subseteq V\setminus\psi_n^t$ we define

$$\mathscr{J}_0(A, \psi_n^t) \triangleq D_{\mathrm{KL}}\big(f_0(X_A; A \mid \mathcal{F}_t) \parallel f_1(X_A; A \mid \mathcal{F}_t)\big), \tag{5}$$

$$\mathscr{J}_1(A, \psi_n^t) \triangleq D_{\mathrm{KL}}\big(f_1(X_A; A \mid \mathcal{F}_t) \parallel f_0(X_A; A \mid \mathcal{F}_t)\big), \tag{6}$$

where $f_\ell(X_A; A \mid \mathcal{F}_t)$ denotes the conditional pdf of X_A given the data collected up to time t, captured by the σ-algebra $\mathcal{F}_t$, and $D_{\mathrm{KL}}(f \parallel g)$ denotes the Kullback-Leibler (KL) divergence from a statistical model with pdf g to a model with pdf f.

C. Problem Statement

The coupled information-gathering strategy and decision-making processes are uniquely specified by the triplet $\Phi_n = (\tau_n, \delta_n, \psi_n^{\tau_n})$. Designing the optimal sampling strategy for achieving the quickest reliable decision involves resolving the tension between the *quality* and *agility* of the process as two opposing measures (improving one penalizes the other one). The agility of the process is captured by the average delay in reaching a decision, i.e., $\mathbb{E}\{\tau_n\}$, and the decision quality is captured by the frequency of erroneous decisions denoted by

$$\mathsf{P}_n^0 \triangleq \mathbb{P}_0(\delta_n = 1), \qquad \text{and} \qquad \mathsf{P}_n^1 \triangleq \mathbb{P}_1(\delta_n = 0). \tag{7}$$

To formalize the quickest reliable decision, we control the quality of the decision and minimize the average number of samples over all possible combinations of $\Phi_n = (\tau_n, \delta_n, \psi_n^{\tau_n})$.

[1]We remark that the subscript n is included in all decision rules to signify the effect of graph size.

An optimal sampling strategy of interest is a solution to

$$\mathcal{P}(\alpha,\beta) \triangleq \left\{ \begin{array}{ll} \inf_{\Phi_n} & \mathbb{E}\{\tau_n\} \\ \text{s.t.} & \mathsf{P}_n^0 \leq \mathrm{e}^{-n\alpha} \\ & \mathsf{P}_n^1 \leq \mathrm{e}^{-n\beta} \end{array} \right. , \tag{8}$$

where $\alpha, \beta \in \mathbb{R}_+$ control the error probability terms P_n^0 and P_n^1, respectively, and are selected such that the problem $\mathcal{P}(\alpha,\beta)$ is feasible.

III. Network-Guided Active Sampling

The core element in characterizing the decision tuple $\Phi_n = (\tau_n, \delta_n, \psi_n^{\tau_n})$ is the data-adaptive and sequential sampling process $\psi_n(t)$. The structure of this process is strongly shaped by the two MRFs specified under H_0 and H_1. In this section, we characterize a data-adaptive and sequential sampling process and show that this process, in conjunction with a thresholding rule for the stopping time and a likelihood ratio detection rule, constitutes an optimal solution to (8). Optimality properties, performance analysis, and complexity analysis are provided in Section IV.

A. Terminal Decision Rules

Before providing the details of the core process (node selection rule), we briefly discuss the terminal decision rules. For this purpose, define[2]

$$\Lambda_t \triangleq \ln \frac{f_1(Y^t;\psi^t)}{f_0(Y^t;\psi^t)}, \tag{9}$$

as the log-likelihood ratio (LLR) of the samples collected up to time t. It can be readily verified that

$$\Lambda_{t+1} = \Lambda_t + \ln \frac{f_1(Y_{t+1};\psi(t+1)\,|\,\mathcal{F}_t)}{f_0(Y_{t+1};\psi(t+1)\,|\,\mathcal{F}_t)}. \tag{10}$$

Stopping Rule: To specify the stopping rule of the sampling process, we define

$$\gamma_n^{\mathrm{L}} \triangleq -n\beta, \quad \text{and} \quad \gamma_n^{\mathrm{U}} \triangleq n\alpha, \tag{11}$$

and specify the stopping time through the following sequential likelihood ratio test:

$$\tau_n^* \triangleq \inf \big\{ t : \Lambda_t \notin (\gamma_n^{\mathrm{L}}, \gamma_n^{\mathrm{U}}) \text{ or } t = n \big\}. \tag{12}$$

This is a *truncated* sequential probability ratio test (SPRT), in which the delay is bounded by the total number of samples possibly available. If we drop the condition $t = n$, the stopping rule simplifies to that of the canonical SPRT. We note that depending on the context, there exist other variations of truncated SPRT as well [79].

Detection Rule: At the stopping time, we decide on the model according to

$$\delta_n^* \triangleq \left\{ \begin{array}{ll} 0, & \text{if } \Lambda_{\tau_n^*} < 0 \\ 1, & \text{if } \Lambda_{\tau_n^*} \geq 0 \end{array} \right. . \tag{13}$$

Based on (12) and (13), the sampling process resumes as long as $\Lambda_t \in (\gamma_n^{\mathrm{L}}, \gamma_n^{\mathrm{U}})$ and terminates once Λ_t falls outside this band or we exhaust all the samples, i.e., $t = n$. Furthermore, if Λ_t exits this interval from the upper threshold γ_n^{U} the set $\{X_1, \ldots, X_n\}$ is deemed to form a Markov network with model $\mathbb{P}_1$, and if it falls below the lower threshold γ_n^{L} we make a decision in favor of $\mathbb{P}_0$. We remark this decision rule is different from that of the SPRT. Specifically, our thresholds are constants, while those of the SPRT are controlled by the target error probabilities according to

$$\delta_{\mathrm{SPRT}}^* = \left\{ \begin{array}{ll} 0, & \text{if } \Lambda_{\tau_{\mathrm{SPRT}}} < \gamma_n^{\mathrm{L}} \\ 1, & \text{if } \Lambda_{\tau_{\mathrm{SPRT}}} \geq \gamma_n^{\mathrm{U}} \end{array} \right. . \tag{14}$$

We also note that the SPRT continues as long as $\gamma_n^{\mathrm{L}} < \Lambda_t < \gamma_n^{\mathrm{U}}$.

B. Dynamic Sampling

At any time $t \in \{1, \ldots, \tau_n\}$, prior to the stopping time, based on the information accumulated up to time $(t-1)$ the sampling process dynamically identifies and takes a sample from one unobserved node that is expected to provide the most relevant information about the true hypothesis. In this subsection, we provide two approaches to dynamic node selection. First, we provide the design of the selection rule based on Chernoff's principle, as the widely used approach for various controlled (active) sensing decisions. Its widespread use is mainly due to its computational simplicity and the fact that it admits asymptotic optimality in a wide range of settings. Next, we discuss the shortcomings of Chernoff's rule, mainly because it loses its optimality (even in the asymptotic regime) for the problem at hand. Motivated by this, we finally offer an alternative rule to circumvent Chernoff rule's shortcomings. We remark that discussing Chernoff's rule serves a two-fold purpose: it furnishes some of the elements for designing the optimal approach and serves as the baseline for assessing the performance of the proposed rule.

1) Chernoff's Principle: In the context of the problem studied in this paper, at any time t and based on the filtration $\mathcal{F}_t$, Chernoff's rule first forms the maximum likelihood (ML) decision about the true model of the data $\mathsf{T} \in \{\mathsf{H}_0, \mathsf{H}_1\}$. By denoting the ML decision about the true hypothesis at time t by $\delta_{\mathrm{ML}}(t)$ we have

$$\delta_{\mathrm{ML}}(t) \triangleq \left\{ \begin{array}{ll} \mathsf{H}_0, & \text{if } \Lambda_{\mathrm{t}} < 0 \\ \mathsf{H}_1, & \text{if } \Lambda_{\mathrm{t}} \geq 0 \end{array} \right. . \tag{15}$$

Next, based on this decision, Chernoff's rule at time t selects and samples the node whose sample is expected to maximally reinforce that the decision $\delta_{\mathrm{ML}}(t)$ becomes also the decision at time $(t+1)$. We define $\psi_{\mathrm{ch}}(t)$ as the node selected by Chernoff's rule at time t, and accordingly define the *ordered* set $\psi_{\mathrm{ch}}^t = \{\psi_{\mathrm{ch}}(1), \ldots, \psi_{\mathrm{ch}}(t)\}$. To formalize Chernoff's rule in the context of the hypothesis testing problem considered in this paper, and in order to quantify the information gained from each sample, we define the following two measures:

$$D_0^i(t) \triangleq \mathscr{J}_0(\{i\}, \psi_{\mathrm{ch}}^{t-1}), \tag{16}$$

$$\text{and} \quad D_1^i(t) \triangleq \mathscr{J}_1(\{i\}, \psi_{\mathrm{ch}}^{t-1}), \tag{17}$$

[2]For simplicity in notations, throughout the rest of the paper, we omit the subscript n in terms ψ_n^t, $\psi_n(t)$, and φ_n^t.

where $\mathscr{J}_0$ and $\mathscr{J}_1$ are defined in (5) and (6), respectively. Measure $D^i_\ell(t)$ quantifies the information gained by observing node i at time t when the true hypothesis is H_ℓ.

Chernoff's rule selects the node that maximizes the distance between f_ℓ and its alternative when the ML decision is in favor of H_ℓ. Therefore, we obtain the following node selection function:

$$\psi_{\text{ch}}(t) \triangleq \begin{cases} \underset{i\in\varphi^t}{\arg\max}\ D^i_0(t), & \text{if } \delta_{\text{ML}}(t-1)=\mathsf{H}_0 \\ \underset{i\in\varphi^t}{\arg\max}\ D^i_1(t), & \text{if } \delta_{\text{ML}}(t-1)=\mathsf{H}_1 \end{cases}. \tag{18}$$

To avoid any ambiguities, whenever $\arg\max_{i\in\varphi^t_n} D^i_\ell(t)$ is not unique (for instance, at the beginning of the sampling process), we select one node randomly according to a uniform distribution. Chernoff's rule minimizes the average delay in the asymptote of a low rate of erroneous decisions if all the selection actions are *independent* [57], [62], which in the context of this paper translates to testing for two distributions without any correlation structures. In this paper, however, the available actions, i.e., selecting unobserved nodes, are co-dependent due to the underlying MRF's correlation structure. Therefore, Chernoff's rule, which ignores such correlation, naturally fails to leverage the correlation structure in forming the sampling decisions. Specifically, by selecting the best immediate action, Chernoff's rule ignores the perspective of the decisions and the impact of the current decision on the future ones.

We provide an example in Section V-C through which we show that designing the node selection rule based on Chernoff's principle is not optimal (even asymptotically). Our analyses show that incorporating the impact of the decisions on future actions improves the agility of the process significantly. This, in turn, brings about computational complexities, which we will show can be reduced considerably by leveraging the MRF structures. In the context of the problem analyzed in this paper, another disadvantage of Chernoff's rule is for settings in which the MRFs are comprised of multiple disconnected subgraphs. In such cases, the sampling strategy will be trapped in one subgraph until it exhausts all the nodes in that subgraph before moving to another one. This limits the flexibility of the sampling strategy for freely navigating the entire graph. Another shortcoming of Chernoff's rule that penalizes the quickness significantly is when the highly correlated nodes (random variables) are concentrated in a cluster with a size considerably smaller than that of the graph n. In such cases, our proposed selection rule approaches the cluster more rapidly.

2) Active Sampling Rule: We start by introducing information measures that link the node selection decisions over time. This enables dynamically incorporating the impact of the decision at any given time on all possible future ones. We select these measures to facilitate selecting the nodes, the samples of which maximize the combination of immediate information, and future expected information. To this end, at time t and for each node $i\in\varphi^t$ we define the set $\mathcal{R}^i_t$ as the set of all subsets of φ^t that contain i, i.e.,

$$\mathcal{R}^i_t \triangleq \{\mathcal{S}\ :\ \mathcal{S}\subseteq\varphi^t \text{ and } i\in\mathcal{S}\}. \tag{19}$$

Corresponding to the samples collected from the nodes in the set $\mathcal{S}\in\mathcal{R}^i_t$, under H_0 and H_1 we define the following information measures:

$$M^i_0(t,\mathcal{S}) \triangleq \mathscr{J}_0(\mathcal{S},\psi^{t-1}) \tag{20}$$

$$= \mathbb{E}_0\left\{\ln\frac{f_0(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}{f_1(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}\right\}, \tag{21}$$

$$\text{and}\quad M^i_1(t,\mathcal{S}) \triangleq \mathscr{J}_1(\mathcal{S},\psi^{t-1}) \tag{22}$$

$$= \mathbb{E}_1\left\{\ln\frac{f_1(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}{f_0(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}\right\}. \tag{23}$$

The terms $M^i_\ell(t,\mathcal{S})$ capture the information content of $|\mathcal{S}|$ samples. Hence, the normalized terms $\frac{1}{|\mathcal{S}|}M^i_\ell(t,\mathcal{S})$ account for the *average* information content per sample. Based on these two normalized measures, an optimal action is to select the node that maximizes the *average* information over all possible future decisions. Therefore, the node selection function is the solution of the following optimization problem over all combinations of the unobserved nodes:

$$\psi^*(t)=\begin{cases} \underset{i\in\varphi^t}{\arg\max}\ \underset{\mathcal{S}\in\mathcal{R}^i_t}{\max}\ \dfrac{M^i_0(t,\mathcal{S})}{|\mathcal{S}|}, & \text{if } \delta_{\text{ML}}(t-1)=\mathsf{H}_0 \\ \underset{i\in\varphi^t}{\arg\max}\ \underset{\mathcal{S}\in\mathcal{R}^i_t}{\max}\ \dfrac{M^i_1(t,\mathcal{S})}{|\mathcal{S}|}, & \text{if } \delta_{\text{ML}}(t-1)=\mathsf{H}_1 \end{cases}. \tag{24}$$

In this selection rule, an ML decision about the true hypothesis is formed based on the collected data, and the node that maximizes the average information over all possible future sequences of samples is selected. We note that the sets $\mathcal{S}$ are selected such that they i) contain node i, which is a candidate to be observed at time t, and ii) contain possibly additional nodes that will be observed in the future. Mimicking this decomposition of $\mathcal{S}$, for $\mathcal{S}\in\mathcal{R}^i_t$, the information measure $M^i_\ell(t,\mathcal{S})$ for $\ell\in\{0,1\}$ can be also decomposed according to

$$M^i_\ell(t,\mathcal{S}) = \mathscr{J}_\ell(\{i\},\psi^{t-1}) \ +\ \mathscr{J}_\ell(\mathcal{S}\backslash\{i\},\psi^{t-1}). \tag{25}$$

In this expansion, the first term in the decomposition, i.e., $\mathscr{J}_\ell(\{i\},\psi^{t-1})$ defined in (5) and (6), accounts for the information gained by observing node i at time t. Similarly, the second term $\mathscr{J}_\ell(\mathcal{S}\backslash\{i\},\psi^{t-1})$ is the expected information gained from future samples from the nodes contained in $\mathcal{S}\backslash\{i\}$ when $\psi(t)=i$. This second term constitutes the key distinction of the proposed rule compared to Chernoff's rule, which accounts for incorporating every possible future action. Finding the optimal node i and set $\mathcal{S}$ in (24) involves an exhaustive search over all the remaining nodes, which can become computationally prohibitive. In the next subsection, we show that by leveraging the Markov properties of an MRF, and a certain acyclic dependency assumption, the exhaustive search for an optimal $\mathcal{S}\in\mathcal{R}^i_t$ can be simplified significantly. Based on the stopping rule specified in (12), the terminal decision rule given in (13), and the sampling rule specified in (73), Algorithm 1 provides the detailed steps

Algorithm 1 Network-guided active sampling for quickest detection of Markov networks

set $t=0$, $\varphi^1=V$, $\Lambda_0=0$, $\gamma_n^{\mathrm{L}}=-n\beta$, and $\gamma_n^{\mathrm{U}}=n\alpha$
$t \leftarrow t+1$
for $i \in \varphi^t$
  for any $\mathcal{S} \subseteq \mathcal{R}_t^i$
    compute $M_0^i(t,\mathcal{S})$ and $M_1^i(t,\mathcal{S})$ according to (20)-(22)
  end for
end for
find $\psi^*(t)$ based on (24)
$\varphi^{t+1} \leftarrow \varphi^{t+1} \setminus \psi(t)$
compute Λ_t according to (10)
if $\gamma_n^{\mathrm{L}} < \Lambda_t < \gamma_n^{\mathrm{U}}$ and $t<n$
    go to step 2
else if $\Lambda_t < 0$
    set $\delta_n^*=0$ and $\tau_n^*=t$
else
    set $\delta_n^*=1$ and $\tau_n^*=t$
end if

for detecting a Markov network with a certain correlation structure.

IV. Main Results

In this section, we provide performance guarantees for the proposed network-guided active sampling procedure in Algorithm 1. Specifically, we analyze the accuracy of the decision in Section IV-A; the delay (sample complexity) of the algorithm in Section IV-B; the error exponents in Section IV-C; and the complexity of the node-selection rule in Section IV-D.

A. Decision Reliability

Problem $\mathcal{P}(\alpha,\beta)$ by design faces a hard constraint on the number of available samples n. This, in turn, acts as a *hard* constraint on the stopping time τ_n. Under such a constraint, the error probabilities P_n^0 and P_n^1 cannot necessarily be made arbitrarily small simultaneously. Hence, a decision algorithm provides a *feasible* solution to $\mathcal{P}(\alpha,\beta)$ only if it satisfies the constraints enforced on P_n^0 and P_n^1 while not requiring more than n samples.

Definition 1 ((α,β)-Accuracy): We say that a decision tuple $\Phi_n \triangleq (\tau_n, \delta_n, \psi_n^{\tau_n})$ is (α,β)-accurate if it ensures $\mathsf{P}_n^0 \leq \mathrm{e}^{-n\alpha}$ and $\mathsf{P}_n^1 \leq \mathrm{e}^{-n\beta}$. First, we establish that the decisions generated by Algorithm 1 satisfy the performance guarantees of the problem.

In this subsection, we examine the problem (8) in both the asymptotic and non-asymptotic regime with respect to the size of the network n, and characterize conditions on α and β under which Algorithm 1 is guaranteed to generate (α,β)-accurate decisions. To this end, note that the sampling process terminates if i) the LLR Λ_t exits the band $(\gamma_n^{\mathrm{L}}, \gamma_n^{\mathrm{U}})$ at some $t \in V$, or ii) we exhaust all the samples, i.e., $\tau_n = n$. For establishing the conditions for ensuring (α,β)-accuracy, in the first step, we show that if the process terminates by exiting the band $(\gamma_n^{\mathrm{L}}, \gamma_n^{\mathrm{U}})$, then the decision is (α,β)-accurate. In the second step, we evaluate the probability of Λ_t exiting the band $(\gamma_n^{\mathrm{L}}, \gamma_n^{\mathrm{U}})$ prior to exhaustive all n samples. These two steps, collectively, establish a sufficient condition for ensuring (α,β)-accuracy of Algorithm 1. For this purpose, we denote the Bhattacharyya coefficient, as a measure of similarity of the two distributions, by

$$\mathsf{B}_n(f_0,f_1) \triangleq \int \sqrt{f_0(x;V) f_1(x;V)}\,\mathrm{d}x. \tag{26}$$

Accordingly, we denote the *normalized* Bhattacharyya distance by

$$\kappa(f_0,f_1) \triangleq -\lim_{n\to\infty} \frac{1}{n} \ln \mathsf{B}_n(f_0,f_1). \tag{27}$$

The following theorem establishes a sufficient condition under which Algorithm 1 generates (α,β)-accurate solutions.

Theorem 1 (Non-Asymptotic (α,β)-Accuracy): For a given network size n, Algorithm 1 generates an (α,β)-accurate solution with a probability at least

$$1-\mathsf{B}_n(f_0,f_1)\left[\epsilon_0 \exp\left(\frac{n\beta}{2}\right)+\epsilon_1 \exp\left(\frac{n\alpha}{2}\right)\right]. \tag{28}$$

Proof: See Appendix A. ∎

We will evaluate the probability term in (28) in Section V-A through an illustrative example and in Section VI through numerical evaluations. We will show that for widely used MRF models (e.g., Gaussian MRFs), this probability approaches 1 in all practical ranges of n and error probabilities, rendering the Algorithm 1 (α,β)-accurate almost surely even in the non-asymptotic regime. By leveraging the result of Theorem 1, we can readily provide a sufficient condition for (α,β)-accuracy in the asymptote of large network sizes.

Corollary 1 (Asymptotic (α,β)-Accuracy): Algorithm 1 generates (α,β)-accurate solutions *almost surely* in the asymptote of large networks if

$$\max\{\alpha,\beta\} < 2\kappa(f_0,f_1). \tag{29}$$

Proof: The proof follows from finding a sufficient condition that ensures probability in (28) approaches 1 as $n\to\infty$. ∎

B. Delay Analysis

In this subsection, we analyze the performance of the proposed selection rule in the asymptote of large networks sizes, i.e., when $n\to\infty$, i.e., $V=\mathbb{N}$. We note that the proposed network-guided node selection rule capitalizes on the discrepancies in the information measures corresponding to selecting different nodes. In general, a wider range of information measures leads to more effectively distinguishing the most informative nodes to sample. This, in turn, reduces the average delay for reaching a sufficiently confident decision. In order to analyze the performance, corresponding to any subset of nodes $U \subseteq \mathbb{N}$ we define normalized LLR measures as follows:

$$\mathsf{nLLR}_0(X_U;U) \triangleq \frac{1}{|U|}\ln\frac{f_0(X_U;U)}{f_1(X_U;U}, \quad \text{when } (X_1,\ldots,X_n)\sim\mathbb{P}_0, \tag{30}$$

and

$$\mathsf{nLLR}_1(X_U;U) \triangleq \frac{1}{|U|}\ln\frac{f_1(X_U;U)}{f_0(X_U;U)}, \quad \text{when } (X_1,\ldots,X_n)\sim\mathbb{P}_1. \tag{31}$$

These log-likelihood ratios play pivotal roles in characterizing the performance of sequential methods. When the

random variables $\{X_i : i \in V\}$ are independent and identically distributed (i.i.d.), according to the strong law of large numbers, the measures $\mathsf{nLLR}_\ell(Y^t;\psi^t)$ converge almost surely to the KL divergence terms as $|U| \to \infty$. While in an i.i.d. setting these measures are well-defined and can have tangible interpretations (e.g., being random walks), in a non-i.i.d. setting, they are not as well-defined, and their convergence can be guaranteed only under stronger conditions. A relevant notion of convergence for non-i.i.d. settings that is especially widely used in sequential detection is *complete* convergence (introduced in [80], a good overview in [81], and used in the context of sequential detection in [31] and [83]).[3] For this purpose, corresponding to the set of nodes V we define

$$\mathcal{S}(V) \triangleq \{\forall A \subseteq V \ : \ |A| \geq g(n)\}, \tag{32}$$

where $n = |V|$ and $g(x)$ is an arbitrary function that satisfies $g(x) \xrightarrow{x\to\infty} \infty$. Hence, $\mathcal{S}(V)$ is the collection of all subsets of V whose cardinality is at least $g(n)$.

Definition 2 (Complete Convergence): Corresponding to any possible sampling sequence $\psi^\infty \in \mathcal{S}(\mathbb{N})$, we say that the normalized log-likelihood ratios $\mathsf{nLLR}_\ell(Y^t;\psi^t)$ converge *completely* to a constant $I_\ell(\psi^\infty)$ when

$$\sum_{t=1}^{\infty} \mathbb{P}_\ell\Big\{\Big|\mathsf{nLLR}_\ell(Y^t;\psi^t) - I_\ell(\psi^\infty)\Big| > h\Big\} < +\infty, \quad \forall h > 0. \tag{33}$$

It can be readily verified that the condition in (33) is equivalent to

$$\mathbb{E}_\ell[T_\ell(h,\psi^\infty)] < +\infty, \quad \forall h > 0, \tag{34}$$

where we have defined

$$\begin{aligned} &T_\ell(h,\psi^\infty) \\ &\quad \triangleq \sup\ \Big\{t \in \mathbb{N} \ : \ \Big|\mathsf{nLLR}_\ell(X_{U^t};U^t) - I_\ell(\psi^\infty)\Big| \geq h\Big\}. \end{aligned} \tag{35}$$

The term $T_\ell(h,\psi^\infty)$ denotes the last time that the sequence $\{\mathsf{nLLR}_\ell(Y^t;\psi^t)\}$ leaves the interval

$$[I_\ell(\psi^\infty) - h,\ I_\ell(\psi^\infty) + h]. \tag{36}$$

Next, we define two types of networks, depending on how the LLR sequences converge.

Definition 3 (Homogeneous Network): We say that an MRF is *homogeneous* when $I_\ell(\psi^\infty)$ exists and it is the same for all possible sets ψ^∞. When we have a homogeneous structure, we replace $I_\ell(\psi^\infty)$ by the shorthand I_ℓ, which emphasizes a lack of dependence on ψ^∞.

The critical property of homogeneous networks is that observing any subsequence of nodes provides the same average amount of information in the long run.

Example 1: Consider a setting in which the samples are i.i.d. under H_0 and they form a Gauss-Markov random field (GMRF) under H_1 with the same marginal distributions as the ones under H_0. If under H_1 the nodes form a line graph with correlation coefficients $a \neq \pm 1$, then we have a homogeneous network in which

$$I_0 = \ln(1-a^2) + \frac{2a^2}{1-a^2}, \quad \text{and} \quad I_1 = \ln\frac{1}{1-a^2}. \tag{37}$$

Definition 4 (Heterogeneous Network): We say that an MRF is *heterogeneous* when the two information measures $I_\ell(\psi^\infty)$ exist and vary for different permutations ψ^∞. In such networks, we define

$$I_\ell^* \triangleq \sup_{\psi^\infty \in \mathcal{S}(\mathbb{N})} I_\ell(\psi^\infty), \qquad \text{for } \ell \in \{0,1\}. \tag{38}$$

Example 2: Consider a setting in which the samples are i.i.d. under H_0, and they form a GMRF under H_1 with the same marginal distributions as the ones under H_0. Under H_1 the dependency graph consists of two line subgraphs defined over two distinct sets of nodes denoted by

$$\psi_1^\infty = \{2k-1 : k \in \mathbb{N}\} \quad \text{and} \quad \psi_2^\infty = \{2k : k \in \mathbb{N}\}, \tag{39}$$

where the elements in ψ_i^∞ have constant correlation coefficients $a_i \neq \pm 1$. Assuming $|a_1| > |a_2|$ we have

$$I_0(\psi_i^\infty) = \ln(1-a_i^2) + \frac{2a_i^2}{1-a_i^2}, \tag{40}$$

$$I_1(\psi_i^\infty) = \ln\frac{1}{1-a_i^2}, \tag{41}$$

and for the supremum of these two measures we have

$$I_0^* = \ln(1-a_1^2) + \frac{2a_1^2}{1-a_1^2}, \tag{42}$$

$$I_1^* = \ln\frac{1}{1-a_1^2}. \tag{43}$$

We remark that, in general, GMRFs have heterogeneous structures. One well-known example is in social networks where it has been shown that both weak and strong ties exist and they play very different roles in the dynamics of the network [84]. More generally, it has been shown that a wide range of connection strengths among members of a network is possible [85], [86]. Based on the measures I_ℓ and I_ℓ^* in homogeneous and heterogeneous settings, respectively, next, we analyze the average stopping time. We first focus on the homogeneous setting and establish the optimality of stopping and terminal decision rules characterized in (11)–(13) and then generalize the results to the heterogeneous setting. The following lemma will be instrumental in evaluating the average stopping time.

Lemma 1: For the choices of α and β that satisfy (29), in the homogeneous and heterogeneous networks we almost surely have

$$\max\{\alpha,\beta\} \leq \min\{I_0, I_1\}, \tag{44}$$

$$\text{and} \qquad \max\{\alpha,\beta\} \leq \min\{I_0^*, I_1^*\}. \tag{45}$$

Proof: See Appendix B. ∎

This is in accordance with the results from the binary hypothesis testing literature for both i.i.d. and non-i.i.d. samples, where it has been shown that the error exponents of type I and type II errors are identical to the KL divergence from one distribution to the other [29], [32]. The following theorem

[3] In some literature it is also called 1-quick convergence (see [83]) with generalizations to stronger r-quickness convergence in [31].

provides a universal (algorithm-independent) lower bound on the average delay for any feasible solution to problem (8) when the network has a homogeneous dependency structure.

Theorem 2 (Homogeneous Structures – Delay Converse): In a homogeneous network with information constants I_0 and I_1, any feasible solution of problem (8) with the stopping time τ_n satisfies

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n\}}{n}\geq\frac{\beta}{I_0},\quad\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n\}}{n}\geq\frac{\alpha}{I_1}. \tag{46}$$

Proof: See Appendix C. ∎

We show that any selection rule combined with the likelihood ratio test given in (11)–(13) achieves these lower bounds.

Theorem 3 (Homogeneous Structures – Delay Achievability): In a homogeneous network, for the stopping and terminal decision rules specified in (11)–(13) and an arbitrary sampling rule, in the asymptote of large n we have

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n^*\}}{n}\leq\frac{\beta}{I_0},\quad\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n^*\}}{n}\leq\frac{\alpha}{I_1}. \tag{47}$$

Proof: See Appendix D. ∎

The last two theorems, collectively, establish that when the network has a homogeneous structure, irrespectively of how the nodes are selected and sampled over time, the stopping and terminal decision rules specified in (11)–(13) render asymptotically optimal decisions. The optimality of the decisions being independent of the node selection rule signifies that in homogeneous structures, all sequences of nodes, asymptotically, contain the same average amount of information, and the overall performance does not critically depend on the sampling path. Next, we show that the observation above is not necessarily valid for the networks with heterogeneous structures, and the optimality of decisions in those networks critically depends on the sampling path. By leveraging Theorem 2, in the next corollary, we first provide algorithm-independent lower bounds on the average delay in heterogeneous networks.

Corollary 2 (Heterogeneous Structures – Delay Converse): In a heterogeneous network with information constants I_0^* and I_1^*, any feasible solution of problem (8) with the stopping time τ_n satisfies

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n\}}{n}\geq\frac{\beta}{I_0^*},\quad\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n\}}{n}\geq\frac{\alpha}{I_1^*}. \tag{48}$$

Proof: By following the same line of argument as in the proof of Theorem 2 we can show that for any arbitrary sampling path $\psi^\infty\in\mathbb{N}$ we have

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n\}}{n}\geq\frac{\beta}{I_0(\psi^\infty)}, \tag{49}$$

$$\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n\}}{n}\geq\frac{\alpha}{I_1(\psi^\infty)}. \tag{50}$$

Since (49) is true for any set ψ^∞, subsequently, we have

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n\}}{n}\geq\inf_{\psi^\infty\in\mathbb{N}}\frac{\beta}{I_0(\psi^\infty)}\overset{(38)}{=}\frac{\beta}{I_0^*}, \tag{51}$$

$$\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n\}}{n}\geq\inf_{\psi^\infty\in\mathbb{N}}\frac{\alpha}{I_1(\psi^\infty)}\overset{(38)}{=}\frac{\alpha}{I_1^*}. \tag{52}$$

∎

Next, we provide the proof for the optimality of the decisions produced by Algorithm 1, and especially the optimality of the proposed dynamic node selection rule when facing heterogeneous networks. This result will also be instrumental in characterizing the performance gap between the proposed sampling strategy and Chernoff's rule. By characterizing this gap, through an example in Section V-C, we will show that Chernoff's rule loses its optimality for the correlation detection problem in networks. To prove the upper bounds on the average delay, we define the random variable $\hat{\tau}_n$ as the first time instant after which the ML decision about the true hypothesis specified in (15) is always correct, i.e.,

$$\hat{\tau}_n\triangleq\inf\{u\ :\ \delta_{\mathrm{ML}}(t)=\mathsf{T},\ \forall t\geq u\}, \tag{53}$$

where we adopt the convention that the infimum of an empty set is $+\infty$. We emphasize that $\hat{\tau}_n$ is not a stopping time, but rather a term that, as we will show, is dominated by the stopping time. In order to establish the desired upper bounds, we show the following two properties for $\hat{\tau}_n$:

1) $\mathbb{E}_i\{\hat{\tau}_n\}$ is upper bounded by a constant.
2) $\frac{1}{n}\mathbb{E}_i\{\tau_n^*-\hat{\tau}_n\}$ is upper bounded according to

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n^*-\hat{\tau}_n\}}{n}\leq\frac{\beta}{I_0^*}, \tag{54}$$

$$\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n^*-\hat{\tau}_n\}}{n}\leq\frac{\alpha}{I_1^*}. \tag{55}$$

In order to prove that $\mathbb{E}_i\{\hat{\tau}_n\}$ is finite, we first provide the following lemma, which establishes that the probability $\mathbb{P}_i(\hat{\tau}_n\geq t)$ decays exponentially with respect to time t.

Lemma 2: $\mathbb{E}_i\{\hat{\tau}_n\}$ is upper bounded by a constant.

Proof: See Appendix E. ∎

Next, in order to prove (54), we define

$$U^\infty\triangleq\arg\max_{\psi^\infty\in\mathcal{S}(\mathbb{N})}I_1(\psi^\infty), \tag{56}$$

corresponding to which we have $I_1(U^\infty)=I_1^*$. When there are more than one choice for U^∞, we select it to be the largest such set. Based on this definition, we provide the following lemma showing that the number of times that we sample from a set other than U^∞ is, on average, finite. This property follows from the assumption of complete convergence in heterogeneous networks.

Lemma 3: Let us define

$$\mathcal{H}_t\triangleq\{s\in\{\hat{\tau}_n+1,\ldots,t\}\ :\ \psi^*(s)\notin U^\infty\}. \tag{57}$$

Then, $\lim_{t\to\infty}\frac{1}{t}|\mathcal{H}_t|=0$.

Proof: See Appendix F. ∎

This establishes that by the stopping time, the samples collected are taken dominantly from the set U^∞. By leveraging Lemma 3, we next provide the final ingredient for characterizing the achievable average delay.

Lemma 4: $\frac{1}{n}\mathbb{E}_i\{\tau_n^*-\hat{\tau}_n\}$ is upper bounded according to (54).

Proof: See Appendix G. ∎

Theorem 4 (Heterogeneous Structures – Delay Achievability): Algorithm 1 generates decisions that are asymptotically

optimal solutions to problem (8). Specifically

$$\lim_{n\to\infty}\frac{\mathbb{E}_0\{\tau_n^*\}}{n}\leq\frac{\beta}{I_0^*},\quad\text{and}\quad\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n^*\}}{n}\leq\frac{\alpha}{I_1^*}.\tag{58}$$

Proof: By combining the results of Lemma 2 and Lemma 4 we obtain

$$\frac{\alpha}{I_1^*}\overset{(54)}{\geq}\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n^*-\hat{\tau}_n\}}{n}\tag{59}$$

$$\geq\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n^*\}}{n}-\lim_{n\to\infty}\frac{B}{n(1-\mathrm{e}^{-c})}\tag{60}$$

$$=\lim_{n\to\infty}\frac{\mathbb{E}_1\{\tau_n^*\}}{n},\tag{61}$$

which concludes the proof for the upper bound on $\frac{1}{n}\mathbb{E}_1\{\tau_n^*\}$. The proof of the upper bound on $\frac{1}{n}\mathbb{E}_0\{\tau_n^*\}$ follows the same line of argument. ∎

C. Error Exponents

In this subsection, we characterize the gain obtained from the data-adaptive stopping time. To this end, we compare the performance of sequential sampling procedures with that of the fixed-sample-size setting in terms of their associated error exponents. In the fixed-sample-size counterpart of the binary testing problem considered in this paper, the optimal decision rule is the Neyman-Pearson (NP) rule, where its associated error exponents are characterized in [21]. By denoting the NP decision rule by δ_{NP}, we define

$$\mathsf{P}_{\mathrm{NP}}^0\triangleq\mathbb{P}_0(\delta_{\mathrm{NP}}=1),\quad\text{and}\quad\mathsf{P}_{\mathrm{NP}}^1\triangleq\mathbb{P}_1(\delta_{\mathrm{NP}}=0),\tag{62}$$

as the frequencies of erroneous decisions by the NP test based on n samples. Accordingly, we define

$$E_{\mathrm{NP}}^0\triangleq-\lim_{n\to\infty}\frac{1}{n}\ln\mathsf{P}_{\mathrm{NP}}^0,\tag{63}$$

$$\text{and}\qquad E_{\mathrm{NP}}^1\triangleq-\lim_{n\to\infty}\frac{1}{n}\ln\mathsf{P}_{\mathrm{NP}}^1,\tag{64}$$

as the associated error exponents. Similarly, we define

$$E_n^0\triangleq-\lim_{n\to\infty}\frac{1}{r_1}\ln\mathsf{P}_n^0(r_1),\tag{65}$$

$$\text{and}\qquad E_n^1\triangleq-\lim_{n\to\infty}\frac{1}{r_0}\ln\mathsf{P}_n^1(r_0),\tag{66}$$

as the error exponents of the sequential detection approach, where $\mathsf{P}_n^0(r_1)$ and $\mathsf{P}_n^1(r_0)$ are the error probabilities of sequential sampling when the average number of samples (i.e., the stopping time) is $r_\ell\triangleq\mathbb{E}_\ell\{\tau_n^*\}$. The connections between the error exponents of the NP test and sequential sampling strategies are established in the following theorem.

Theorem 5 (Gain of Adaptivity): The error exponents of the decision rules in Algorithm 1 are related to those of the NP rule through

$$E_n^1=I_0\qquad\text{and}\qquad E_n^0=I_1,\tag{67}$$

$$E_{\mathrm{NP}}^1=I_0\qquad\text{and}\qquad E_{\mathrm{NP}}^0=0.\tag{68}$$

Proof: See Appendix H. ∎

D. Search Complexity Analysis

In this subsection, we show that under certain connectivity structures for the given MRFs, by judiciously leveraging the structures, the complexity of the search for the optimal node selection path over time can be reduced significantly. For this purpose, based on the given graphs $\mathcal{G}_0(V,E_1)$ and $\mathcal{G}_1(V,E_2)$ we construct the graph $\mathcal{G}(V,E)$ such that

$$E\triangleq E_0\cup E_1.\tag{69}$$

Based on this, we define the neighborhood of node $i\in V$ according to

$$\mathcal{N}_i\triangleq\{j\in V\,:\,j\neq i\,,\,(i,j)\in E\}.\tag{70}$$

We will show that when $\mathcal{G}$ is acyclic, for each node i, the optimal set $\mathcal{S}$ is restricted to only contain the neighbors of i that are not observed prior to time t, i.e., $\mathcal{S}\subseteq\mathcal{L}_t^i$ where

$$\mathcal{L}_t^i\triangleq\{i\}\cup\{\mathcal{N}_i\cap\varphi^t\}.\tag{71}$$

This indicates that for determining the node to select at each time, it is sufficient to consider a significantly shorter future sampling path for each node. The cardinality of the set of subsets of $\mathcal{L}_t^i$ is significantly smaller than that of φ^t, which translates to a substantial reduction in the complexity of characterizing the optimal selection functions. This observation is formalized in the following theorem.

Theorem 6: For an acyclic dependency graph $\mathcal{G}$, at each time t and for $\ell\in\{0,1\}$ we have

$$\arg\max_{i\in\varphi^t}\max_{\mathcal{S}\in\mathcal{R}_t^i}\frac{M_\ell^i(t,\mathcal{S})}{|\mathcal{S}|}=\arg\max_{i\in\varphi^t}\max_{\mathcal{S}\subseteq\mathcal{L}_t^i}\frac{M_\ell^i(t,\mathcal{S})}{|\mathcal{S}|}.\tag{72}$$

Proof: See Appendix I. ∎

Based on this theorem, the selection function given in (24) simplifies to

$$\psi^*(t)=\begin{cases}\arg\max\limits_{i\in\varphi^t}\max\limits_{\mathcal{S}\subseteq\mathcal{L}_t^i}\dfrac{M_0^i(t,\mathcal{S})}{|\mathcal{S}|}, & \text{if }\delta_{\mathrm{ML}}(\mathrm{t}-1)=\mathsf{H}_0\\ \arg\max\limits_{i\in\varphi^t}\max\limits_{\mathcal{S}\subseteq\mathcal{L}_t^i}\dfrac{M_1^i(t,\mathcal{S})}{|\mathcal{S}|}, & \text{if }\delta_{\mathrm{ML}}(\mathrm{t}-1)=\mathsf{H}_1\end{cases}.\tag{73}$$

By further leveraging the Markov property, computing

$$\max_{\mathcal{S}\subseteq\mathcal{L}_t^i}\frac{M_\ell^i(t,\mathcal{S})}{|\mathcal{S}|}\tag{74}$$

can be further simplified. Specifically, by recalling the definition of $M_\ell^i(t,\mathcal{S})$ given in (20) and (22) we have

$$M_\ell^i(t,\mathcal{S})\tag{75}$$

$$=D_{\mathrm{KL}}\big(f_\ell(X_{\mathcal{S}}|\mathcal{F}_{t-1})\parallel f_{1-\ell}(X_{\mathcal{S}}|\mathcal{F}_{t-1})\big)\tag{76}$$

$$=D_{\mathrm{KL}}\big(f_\ell(X_i|\mathcal{F}_{t-1})\parallel f_{1-\ell}(X_i|\mathcal{F}_{t-1})\big)\tag{77}$$

$$+\sum_{j\in\mathcal{S}\setminus\{i\}}D_{\mathrm{KL}}\big(f_\ell(X_j|X_i,\mathcal{F}_{t-1})\parallel f_{1-\ell}(X_j|X_i\mathcal{F}_{t-1})\big)$$

$$=D_{\mathrm{KL}}\big(f_\ell(X_i|X_{\psi^*(t-1)})\parallel f_{1-\ell}(X_i|X_{\psi^*(t-1)})\big)\tag{78}$$

$$+\sum_{j\in\mathcal{S}\setminus\{i\}}D_{\mathrm{KL}}\big(f_\ell(X_j|X_i)\parallel f_{1-\ell}(X_j|X_i)\big),\tag{79}$$

where the transition from (77) to (78) is due to the graph being acyclic and Markov. Hence, for computing the information measures $M_\ell^i(t,\mathcal{S})$ we need to compute only the marginal distributions of the form $f_\ell(X_i|X_j)$.

V. Special Cases and Illustrative Examples

In this section, we consider a few special cases, for each of which we present more specialized results. First, for gaining further insight into the tightness of the probabilistic (α,β)-accuracy guarantee in the non-asymptotic regime (Theorem 1), we provide an illustrative example showing the achievable ranges of error probabilities for a given network size. Next, we consider the setting in which both distributions are Gaussian and characterize measures defined for designing the sampling strategy in terms of the covariance matrices of the distributions. Built on these results, next, we provide a counterexample establishing that Chernoff's rule is not asymptotically optimal for carrying out the detection decisions in the MRFs considered in this paper. Finally, we consider detecting whether a given MRF contains a cluster of nodes whose data form a given correlation model. In all the special cases, we quantify the performance gaps between our network-guided active sampling strategy and Chernoff's rule.

A. Non-Asymptotic Detection Performance

In this subsection, we provide an illustrative example to assess the sufficient condition for (α,β)-accuracy of Algorithm 1 in the non-asymptotic regime, which was established in Theorem 1. We consider testing correlation versus independence when both distributions are Gaussian, i.e.,

$$\begin{aligned}\mathsf{H}_0 \ &: \ (X_1,\ldots,X_n)\sim\mathcal{N}(\boldsymbol{\theta},\mathbf{I}),\\ \mathsf{H}_1 \ &: \ (X_1,\ldots,X_n)\sim\mathcal{N}(\boldsymbol{\theta},\boldsymbol{\Sigma}),\end{aligned}\tag{80}$$

where $\mathbf{I}$ is the identity matrix and $\boldsymbol{\Sigma}$ is an arbitrary correlation matrix such that $\Sigma_{ii}=1$. Hence, the Bhattacharyya distance, which we denote by $\kappa_n(f_0,f_1)$, is given by

$$\kappa_n(f_0,f_1) \triangleq -\ln \mathsf{B}_n(f_0,f_1) \tag{81}$$

$$= \frac{1}{2}\ln\frac{1}{\sqrt{\det\boldsymbol{\Sigma}}}\cdot\det\left(\frac{\mathbf{I}+\boldsymbol{\Sigma}}{2}\right) \tag{82}$$

$$= \frac{1}{2}\ln\prod_{i=1}^{n}\frac{1+\lambda_i}{2\sqrt{\lambda_i}}, \tag{83}$$

where $\{\lambda_i\}_{i=1}^n$ are the distinct eigenvalues of the symmetric positive definite matrix $\boldsymbol{\Sigma}$. Accordingly, the Bhattacharyya coefficient is given by

$$\mathsf{B}_n(f_0,f_1) \ = \ \exp\left(-\kappa_n(f_0,f_1)\right) \ = \ \prod_{i=1}^{n}\sqrt{\frac{2\sqrt{\lambda_i}}{1+\lambda_i}}. \tag{84}$$

By noting that $\Sigma_{ii}=1$ for all $i\in V$, according to Gershgorin circle theorem all the eigenvalues $\{\lambda_i\}_{i=1}^n$ lie in closed discs centered at 1. Select $\xi>0$ such that at least half of the eigenvalues $\{\lambda_i\}_{i=1}^n$ lie outside the interval

$$\left[\left(\sqrt{1+\xi}-\sqrt{\xi}\right)^2,\ \left(\sqrt{1+\xi}+\sqrt{\xi}\right)^2\right]. \tag{85}$$

It can be readily verified that if

$$\lambda_i \notin \left[\left(\sqrt{1+\xi}-\sqrt{\xi}\right)^2,\ \left(\sqrt{1+\xi}+\sqrt{\xi}\right)^2\right], \tag{86}$$

then

$$\frac{2\sqrt{\lambda_i}}{1+\lambda_i} \ < \ \frac{1}{\sqrt{1+\xi}}. \tag{87}$$

Hence, we have the following upper bound on the Bhattacharyya coefficient:

$$\mathsf{B}_n(f_0,f_1) \ \le \ (1+\xi)^{-\frac{n}{8}}. \tag{88}$$

Therefore, for all the error probability exponents α and β that satisfy

$$\frac{1}{4}\ln(1+\xi) \ > \ \max\{\alpha,\beta\}, \tag{89}$$

according to Theorem 1 Algorithm 1 is (α,β)-accurate almost surely in the non-asymptotic regime. For instance, for $n=200$, $\xi=0.2$, $\alpha=\beta=0.02$, Algorithm 1 is (α,β)-accurate with probability at least 0.999.

B. Gauss-Markov Random Fields

In this subsection, we specialize the general results to GMRF, where we assume that

$$\begin{aligned}\mathsf{H}_0 \ &: \ (X_1,\ldots,X_n)\sim\mathcal{N}(\boldsymbol{\theta},\mathbf{I}),\\ \mathsf{H}_1 \ &: \ (X_1,\ldots,X_n)\sim\mathcal{N}(\boldsymbol{\theta},\boldsymbol{\Sigma}),\end{aligned}\tag{90}$$

where $\Sigma_{ii}=1$ for all $i\in V$. This test is generally known as the problem of testing against independence. The graphical model associated with H_0 consists of n nodes without any edges, and we denote the graphical model associated with H_1 by $\mathcal{G}(V,E)$. A GMRF with covariance matrix $\boldsymbol{\Sigma}$ is non-degenerate if $\boldsymbol{\Sigma}$ is positive-definite, in which case, the potential matrix associated with the GMRF is denoted by $\boldsymbol{J}\triangleq\boldsymbol{\Sigma}^{-1}$. The non-zero elements of the potential matrix have a one-to-one correspondence with the edges of the dependency graph in the sense that

$$J_{uv}=0 \quad\Leftrightarrow\quad (u,v)\notin E. \tag{91}$$

In a GMRF, the properties of the network are strongly influenced by the structure of the underlying dependency graph. GMRFs with acyclic dependency represent an important class of GMRFs in which there exists at most one path between any pair of nodes, and consequently, the cross-covariance value between any two non-neighbor nodes in the graph is related to the cross-covariance values of the nodes connecting them. Specifically, corresponding to any two edges $(i,j)\in E$ and $(i,k)\in E$, which share node $i\in V$, we have

$$\Sigma_{jk}=\Sigma_{ji}\Sigma_{ii}^{-1}\Sigma_{ik}, \quad \text{for all } \{j,k\}\subseteq\mathcal{N}_i. \tag{92}$$

In a GMRF with an acyclic graph, the elements and the determinant of the potential matrix can be expressed explicitly in terms of the elements of the covariance matrix.

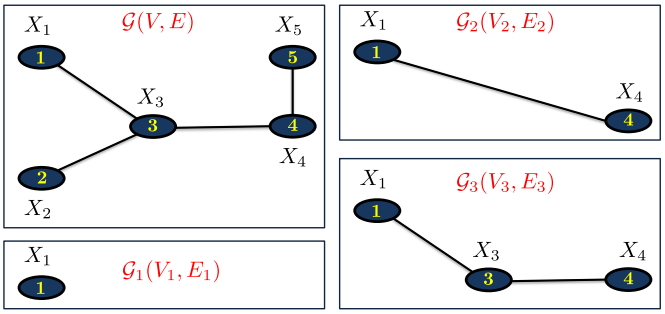

Fig. 2. Toy example for the evolution of $\mathcal{G}_t(V_t,E_t)$ over time for $\psi^3=\{1,4,3\}$.

Theorem 7 ([21], Theorem 1): For a GMRF with an acyclic dependency graph $\mathcal{G}=(V,E)$ and covariance matrix $\mathbf{\Sigma}$, the elements of the potential matrix are given by

$$J_{ii}=\frac{1}{\Sigma_{ii}}\left(1+\sum_{j\in\mathcal{N}_i}\frac{\Sigma_{ij}^2}{\Sigma_{ii}\Sigma_{jj}-\Sigma_{ij}^2}\right),\quad \forall i\in V, \tag{93}$$

and

$$J_{ij}=\begin{cases}\dfrac{-\Sigma_{ij}}{\Sigma_{ii}\Sigma_{jj}-\Sigma_{ij}^2} & \text{if } (i,j)\in E\\ 0 & \text{if } (i,j)\notin E\end{cases}. \tag{94}$$

Furthermore, the determinant of the potential matrix is also given by

$$\det(\boldsymbol{J})=\prod_{i\in V}\Sigma_{ii}^{\deg(i)-1}\prod_{(i,j)\in E}[\Sigma_{ii}\Sigma_{jj}-\Sigma_{ij}^2]^{-\frac{1}{2}}, \tag{95}$$

where $\deg(i)$ is the degree of node i.

We leverage the properties of the GMRFs to obtain closed-form expressions for the information measures defined in (16) and (20)–(22), as well as the node selection rules characterized in Section III-B. In order to describe the effect of the sequential sampling process on different measures that we use, we sequentially construct the sequence of graphs $\{\mathcal{G}_t(V_t,E_t):t\in\{1,\ldots,\tau_n^*\}\}$ such that the graph $\mathcal{G}_t(V_t,E_t)$ at time t is adapted to the nodes observed up to time t. Specifically, we set $V_t=\psi^t$, and for each pair of nodes $i,j\in V_t$ we include an edge $(i,j)\in E_t$ if and only if either $(i,j)\in E$, or there exists a path between nodes i and j in the original graph $\mathcal{G}(V,E)$ such that none of the nodes on this path has been observed up to time t (except for i and j). Figure 2 depicts a toy example on the evolution of $\mathcal{G}_t(V_t,E_t)$ over time for $t\in\{1,2,3\}$ corresponding to an underlying graph $\mathcal{G}(V,E)$. Furthermore, for any $(i,j)\in E_t$ we define

$$\mathsf{LLR}(i,j)\triangleq\frac{1}{2}\Big[\ln\frac{1}{1-\Sigma_{ij}^2}-\frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\big(X_i^2+X_j^2\big)\Big]+\frac{\Sigma_{ij}}{1-\Sigma_{ij}^2}X_iX_j. \tag{96}$$

Under these definitions and by assuming that $\mathcal{G}_t(V_t,E_t)$ remains acyclic at time t, for the LLR of the samples up to time t defined in (9) we have

$$\Lambda_t=\sum_{i\in V_t}\sum_{j\in\mathcal{N}_i^t}\mathsf{LLR}(i,j), \tag{97}$$

where X_i is the sample taken from node i and $\mathcal{N}_i^t\triangleq\{j\in V_t:(i,j)\in E_t\}$. Next, by invoking the GMRF structure and leveraging the results in Theorem 7, the information measures defined for Chernoff's rule in (16) for any $i\in\varphi^t$ can be further simplified and expressed in terms of the correlation coefficients. Specifically, corresponding to Chernoff's rule and its associated sampling sequence $\psi_{\mathrm{ch}}^{\tau_c}$ we have[4]

$$D_0^i(t)=\frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\Big[\ln(1-\Sigma_{ij}^2)+\frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\big(X_j^2+1\big)\Big], \tag{98}$$

$$\text{and}\quad D_1^i(t)=\frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\Big[\ln\frac{1}{1-\Sigma_{ij}^2}+\Sigma_{ij}^2\big(X_j^2-1\big)\Big]. \tag{99}$$

Furthermore, by defining

$$\Delta_t^i\triangleq\{(j,k):\ j,k\in\mathcal{N}_i^t\}, \tag{100}$$

from (5) and (6) and by leveraging the results in Theorem 7, for the proposed node selection rule we have[5]

$$\begin{aligned}\mathscr{J}_0(\{i\},\psi^{t-1})&=\frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\ln(1-\Sigma_{ij}^2)+\frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\Big(X_j^2+1\Big)\\&\quad+\sum_{(j,k)\in\Delta_t^i}\mathsf{LLR}(j,k),\end{aligned} \tag{101}$$

and

$$\begin{aligned}\mathscr{J}_1(\{i\},\psi^{t-1})&=\frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\ln\frac{1}{1-\Sigma_{ij}^2}-\frac{1}{2}\sum_{(j,k)\in\Delta_t^i}\ln\frac{1}{1-\Sigma_{jk}^2}\\&\quad+\frac{1}{2}\Big[\sum_{j\in\mathcal{N}_i^t}\frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\Big(X_j^2-1\Big)+\sum_{(j,k)\in\Delta_t^i}\mathsf{LLR}(j,k)\Big]\\&\quad\times\frac{\prod_{j\in\mathcal{N}_i^t}(1-\Sigma_{ij}^2)}{\prod_{(j,k)\in\Delta_t^i}(1-\Sigma_{jk}^2)}.\end{aligned} \tag{102}$$

Similarly, by leveraging the result of Theorem 6, for any $\mathcal{S}\in\mathcal{L}_t^i$ we find

$$\begin{aligned}\mathscr{J}_0(\mathcal{S}\backslash\{i\},\psi^{t-1})&=\frac{1}{2}\sum_{j\in\mathcal{S}\backslash\{i\}}\Big[\ln(1-\Sigma_{ij}^2)+\frac{2\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\Big],\\\mathscr{J}_1(\mathcal{S}\backslash\{i\},\psi^{t-1})&=\frac{1}{2}\sum_{j\in\mathcal{S}\backslash\{i\}}\Big[\ln\frac{1}{1-\Sigma_{ij}^2}\Big].\end{aligned} \tag{103}$$

Subsequently, based on (25), the closed-form expression of $M_\ell^i(t,\mathcal{S})$ for $\ell\in\{0,1\}$ is obtained from

$$M_\ell^i(t,\mathcal{S})=\mathscr{J}_\ell(\{i\},\psi^{t-1})+\mathscr{J}_\ell(\mathcal{S}\backslash\{i\},\psi^{t-1}). \tag{104}$$

These closed-form expressions of the information measures in terms of the covariance matrix entries and the dependency graph structure substantially reduces the computational complexities involved in calculating these measures from the expected values in (16) and (20)–(22).

[4]Derivations of $D_0^i(t)$ and $D_1^i(t)$ are provided in Appendix L.

[5]Derivations of $\mathscr{J}_0(\{i\},\psi^{t-1})$ and $\mathscr{J}_1(\{i\},\psi^{t-1})$ are provided in Appendix L.

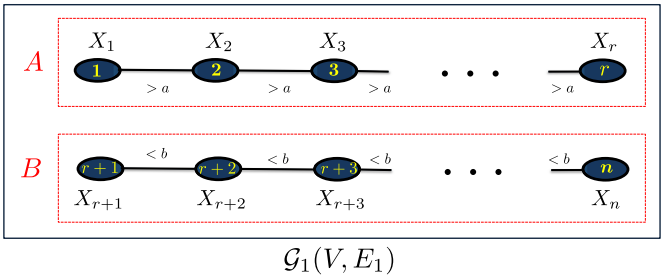

Fig. 3. A GMRF consisting of two line graphs.

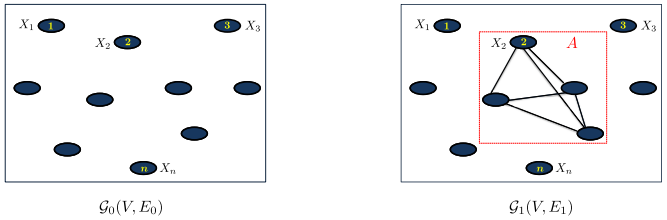

Fig. 4. Independence versus a MRF consisting of one cluster and independent random variables.

C. Counter Example for the Optimality of Chernoff's Rule

Building on the results for the GMRF, in this subsection, we provide an example of a heterogeneous network for which Chernoff's rule is not asymptotically optimal, and quantify the gap between its performance and that of our proposed rule. For this purpose, we consider a setting in which the random variables $X_V = \{X_i \ : \ i \in V\}$ are independent under H_0, while under H_1 they form a GMRF with covariance matrix $\mathbf{\Sigma}$. As depicted in Fig. 3, the dependency graph of the GMRF consists of two disjoint line graphs corresponding to the nodes in sets A and $B = V \setminus A$. By denoting the covariance matrix of the random variables generated by sets A and B by $\mathbf{\Sigma}^A$ and $\mathbf{\Sigma}^B$, respectively, we assume that for any $(i, j) \in E$ we have

$$|\Sigma^A_{ij}| > a, \quad \text{and} \quad |\Sigma^B_{ij}| < b, \tag{105}$$

where $a > b$. This means that the random variables generated by the nodes in set A are more strongly correlated than those generated by the nodes in set B. For such a network, the performance gap between the proposed rule and Chernoff's rule is established in terms of a and b in the following theorem.

Theorem 8: Consider testing independence in (90), where the GMRF consists of two disjoint line graphs corresponding to the sets of nodes in A and B. If the correlation coefficient values between the neighbors in set A are greater than a, while in set B they are less than b and $|A| = p = o(n)$, then as n grows for $\ell \in \{0, 1\}$

$$\lim_{n\to\infty} \frac{\mathbb{E}_\ell\{\tau_{\mathrm{c}}\}}{\mathbb{E}_\ell\{\tau^*_n\}} = \frac{I_\ell(A)}{I_\ell(B)} \geq \left(\frac{a}{b}\right)^2 > 1, \tag{106}$$

where τ_{c} and τ^*_n are the stopping times of the strategies based on Chernoff's rule and Algorithm 1, respectively.

Proof: See Appendix J. ∎

This theorem establishes that Chernoff's rule is not necessarily an asymptotically optimal sampling strategy when selection decisions are statistically dependent.

D. Cluster Detection

In this subsection, we analyze cases in which the two statistical models under H_0 and H_1 are all similar except for a small cluster of nodes that exhibit two different correlation models. Specifically, we first consider a model in which there is a subset of nodes $B \subseteq V$ such that random variables $X_B \triangleq \{X_i : i \in B\}$ are statistically independent under both models H_0 and H_1. This indicates that the correlation models under H_0 and H_1 differ only in their distributions over the random variables from nodes $A \triangleq V \setminus B$, as depicted in Fig. 4. Also, we assume that the random variables $X_A \triangleq \{X_i \ : \ i \in A\}$ form a homogeneous correlation structure, which means that observing any subsequence of the nodes in set A, on average, provides the same amount of information. Clearly, for any set of nodes $U \subseteq B$, we have

$$\forall U \subseteq B \ : \quad I_\ell(U) = 0. \tag{107}$$

In this setting, we show that there is a constant gap between the expected stopping times of the proposed rule and Chernoff's rule. This gap stems from the fact that our proposed approach directly starts from sampling the nodes in A, and does not waste any sampling time by taking samples from set B. However, Chernoff's rule, on average, takes a number of samples from B before sampling from A. The gap between the stopping times is formulated in the next theorem.

Theorem 9: In a network of size n, when there exists a subset of nodes A with size p forming an MRF with a connected graph, while the rest of the network generate independent random variables, we have

$$0 \leq \mathbb{E}_\ell\{\tau_{\mathrm{c}}\} - \mathbb{E}_\ell\{\tau^*_n\} = \Theta\Big(\frac{n}{p}\Big), \quad \text{for } \ell \in \{0, 1\}, \tag{108}$$

where τ_{c} and τ^*_n are the stopping times of the strategies based on Chernoff's rule and the proposed selection rule, respectively.

Proof: See Appendix K. ∎

This theorem establishes the zero-order asymptotic gain of the proposed strategy over Chernoff's rule in a special setting. Note that as p (the size of A) becomes smaller, which leads to more similar and less distinguishable models under H_0 and H_1, the performance gap increases according to $\frac{n}{p}$. Next, we further generalize the above setting to one in which under H_1, besides X_A, random variables X_B also form a homogeneous correlation structure (not independent anymore) with a connected dependency graph, i.e., for $\ell \in \{0, 1\}$ and $\forall U \subseteq B$ we have $I_\ell(U) = I_\ell(B)$. This setting is depicted in Fig. 5. If for set A we have $|A| = o(n)$, then Chernoff's rule starts the sampling process from set B almost surely, and it remains in set B until it exhausts all the nodes of B, while our rule always identifies the most informative nodes to sample. The following theorem characterizes the performance gap between Chernoff's and our rule in this setting.

Theorem 10: Consider a network of size n partitioned into sets A and B specified in Fig. 5. In the asymptote of large n, if the dependency graphs of the nodes in both A and B are

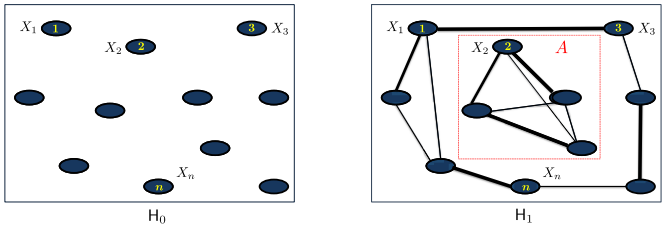

Fig. 5. Independence versus a MRF consisting of two clusters.

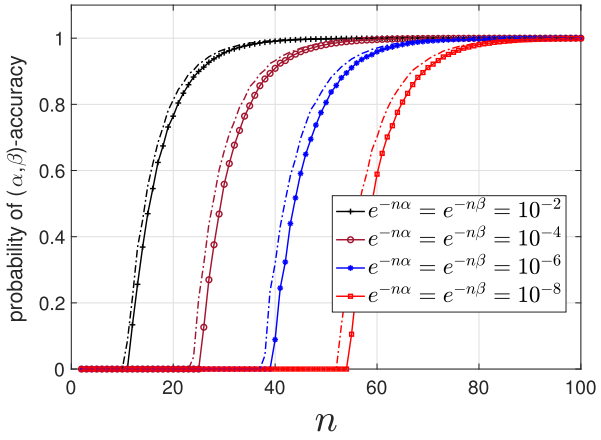

Fig. 6. Lower bound on the probability of (α, β)-accuracy.

connected and $|A| = o(n)$, then

$$\lim_{n\to\infty} \frac{\mathbb{E}_0\{\tau_c\}}{\mathbb{E}_0\{\tau_n^*\}} = \frac{\max\{I_0(A), I_0(B)\}}{I_0(B)}, \tag{109}$$

and

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_c\}}{\mathbb{E}_1\{\tau_n^*\}} = \frac{\max\{I_1(A), I_1(B)\}}{I_1(B)}. \tag{110}$$

Proof: When $p = o(n)$ Chernoff's rule starts the sampling process from set B with probability 1. By invoking the results of Theorem 3, Corollary 2, and Theorem 4 we conclude the proof. ∎

According to the theorem above, when the size of A is sufficiently small such that most of the time, Chernoff's rule starts the sampling process from set B, it loses its first-order asymptotic optimality, as shown in the counterexample in Section V-C. The settings discussed in this subsection highlight the advantages of the proposed selection rule by quantifying two main gains; the gain of selecting the best node at the beginning of the sampling process, and the gain obtained from freely navigating throughout the entire network by jumping across subgraphs in order to find the most informative nodes. Although these settings are special cases, the gain of the proposed rule for a general network is a combination of these two gains.

VI. Numerical Evaluations

In this section, we evaluate the performance of the proposed sampling strategy by comparing it with those of the existing approaches through simulations. First, we examine the (α, β)-accuracy conditions. We consider Gaussian distributions $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_1)$ under models H_0 and H_1, respectively. The covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ have all their diagonal elements equal to 1, and the off-diagonal elements randomly take values in the range $[-1, 1]$, such that the overall combinations constitute valid covariance matrices. Figure 6 shows the variations of the lower bound on the (α, β)-accuracy probability established in Theorem 1 with respect to increasing network size n for four different levels of reliability constraints. It is observed that for reliabilities as small as 10^{-8}, (α, β)-accuracy is guaranteed almost surely when the network size is as small as 100 nodes. We remark that for each reliability level, we evaluate two distinct settings. in one setting, the covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are generated completely randomly (solid curves) and in the other setting half of the n Gaussian random variables, i.e., $\{X_1, \ldots, X_{\frac{n}{2}}\}$ have the same joint distribution (dashed curves).

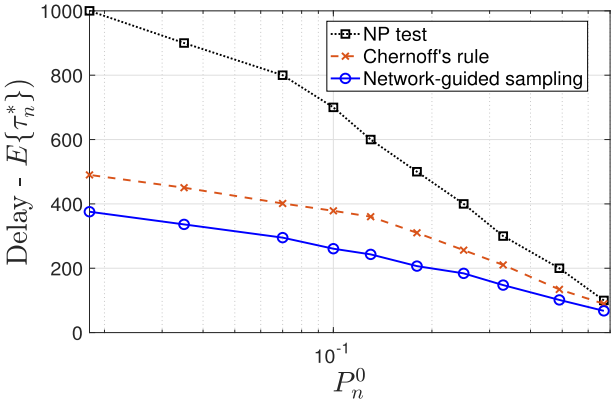

Fig. 7. Average delay versus error probability in a homogeneous network.

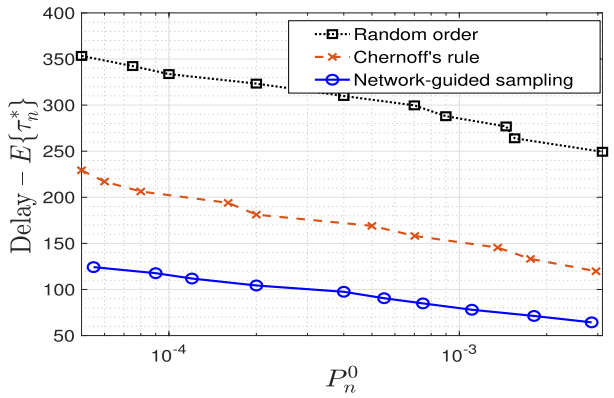

Fig. 8. Average delay versus error probability in a heterogeneous network.

For the rest of the numerical evaluations and simulations, we use the NP test as the fixed sample-size approach, and for the sequential sampling, we consider random (non-adaptive) sampling order and Chernoff's rule. We consider zero-mean Gaussian distributions for data, and test covariance matrix under H_1 versus $\mathbf{I}_n$ under H_0. We also set $\epsilon_0 = \epsilon_1 = 0.5$. As the first comparison, we consider the nearest neighbor dependency graph for uniformly distributed nodes in a two-dimensional space, for which the cross-covariance value between two nearest neighbors is a function of their distance. We denote the distance between nodes i and j by R_{ij} and set the correlation coefficient between nodes i and j to $\Sigma_{ij} =$

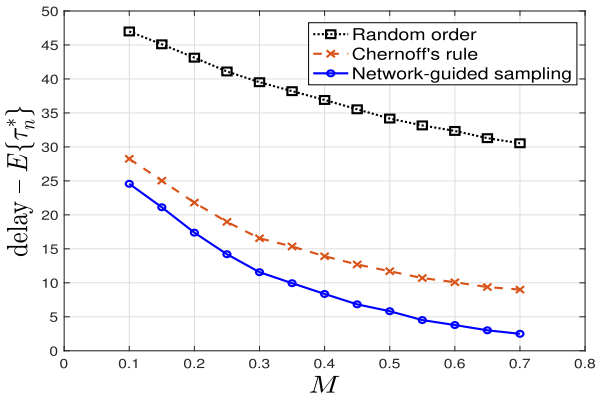

Fig. 9. Average delay versus M.

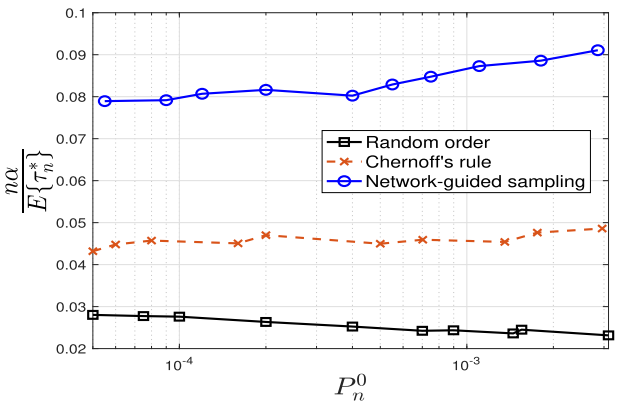

Fig. 10. Error exponent vs. error probability.

$Me^{-aR_{ij}}$, where $a, M \in \mathbb{R}_+$. Under H_0 we set $M = 0$, which corresponds to independent samples. Under H_1 as M increases the KL divergence between the distributions corresponding to f_0 and f_1 grows. In Fig. 7, we set $M = 0.1$, $a = 0$, and $\beta_n = e^{-n\beta} = 0.1$ and compare the performance of different approaches. To this end, for different values of n we find P_n^0 associated with the NP test (i.e., the false alarm probability), based on which we design the sequential sampling strategy for Chernoff's and proposed selection rules and find the average delay. It is observed that the proposed sampling procedure outperforms both the NP test and Chernoff's rule in terms of the reliability-agility trade-off. We also compare the performance of the proposed strategy with that of Chernoff's rule and the random selection rule in a heterogeneous network. For this purpose, we generate a subgraph with three nodes and two edges, in which the cross-covariance values between the neighbors are 0.5 and 0.1. We use 500 copies of this subgraph as the building block for a network consisting of 1500 nodes. For such a network, the optimal rule is to select the nodes with larger cross-covariance values. Figure 8 demonstrates the average delay before reaching a confident decision for different target accuracies and the selection rules when $\alpha = \beta$. By comparing Fig. 7 and Fig. 8, it is observed that in heterogeneous networks, the proposed strategy improves significantly compared to Chernoff's rule. The reason is the larger discrepancy in the amount of information gained from different nodes.

In order to compare the performance of different selection rules for different levels of correlation strength, Fig. 9 compares the average delays incurred by the proposed approach, Chernoff's rule, and a random selection rule for different values of M when $n = 1000$, $\alpha = \beta = 1.6 \times 10^{-3}$, and $a = 1$. It is observed that both Chernoff's rule and the proposed approach outperform the random selection rule, and as the KL divergence grows by increasing M, the improvement is more significant. Furthermore, in Fig. 10 the error exponents are compared where it is observed that the proposed strategy has an error exponent twice as large as that of Chernoff's rule and both of them outperform the strategy based on a random selection of nodes.

In order to verify the results of Theorem 9, we consider a network with $n = 30000$ nodes, in which only a subset

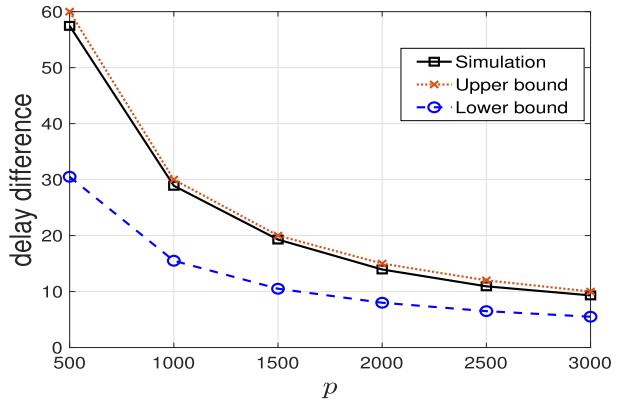

Fig. 11. The average delay difference between Chernoff's rule and tnetwork-guided active sampling.

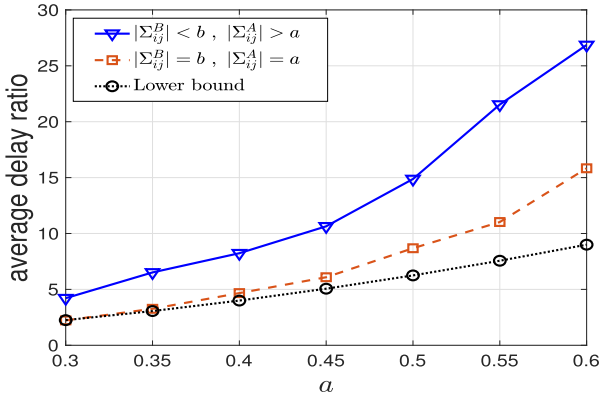

Fig. 12. The ratio of the expected delay of Chernoff's rule and network-guided active sampling.

A consisting of p nodes generate correlated random variables under one of the two hypotheses, while the random variables generated by the rest of the nodes are independent under both hypotheses. Figure 11 demonstrates the average delay of Chernoff's rule in taking its first sample from set A. The upper bound and lower bound obtained in Theorem 9 are also shown for comparison. It is observed that the delay difference is always between the obtained bounds, which confirms that it is $\Theta(\frac{n}{p})$.

Finally, we consider a network with 10000 nodes, from which 50 nodes, denoted by set A, are strongly correlated, i.e., the cross-covariance values between the neighbors in set A, denoted by Σ_{ij}^A, are greater than $a \in (0.3, 0.6)$, while the rest of the nodes, denoted by set B, also form a connected graph with cross-covariance values Σ_{ij}^B less than $b = 0.2$. In Fig. 12 the ratio between the average delay of the proposed sampling strategy is compared with the lower bound $(\frac{a}{b})^2$ obtained in Theorem 8 for different values of a. We also include the ratio between the average delays for the setting in which the cross-covariance values in sets A and B are equal to a and b, respectively, for which it is observed that the lower bound is tighter.

VII. Conclusion

We have considered the quickest detection of a correlation structure in a Markov network, with the objective of determining the true model governing the samples generated by different nodes in the network. After discussing the widely used Chernoff's rule and its shortcomings, we have designed a sequential and data-adaptive sampling strategy to determine the true correlation structure with the fewest average number of samples while, in parallel, the final decision is controlled to meet target reliability. The proposed sampling strategy, which judiciously incorporates the network's correlation structure into its decision rules, involves dynamically deciding whether to terminate the sampling process or to continue collecting further evidence, and prior to terminating the process, which node to observe at each time. We have established the optimality properties of the proposed sampling strategy and leveraged the Markov properties of the network to reduce the computational complexities involved in its implementation. We have provided an example for which Chernoff's rule is not optimal. Finally, we have quantified the advantages of the proposed rule over Chernoff's rule for some special cases.

Notation Used Throughout the Proofs

For convenience, throughout the proofs we drop the subscript n in φ_n^t, ψ_n^t, γ_n^{L}, γ_n^{U}.

Appendix A
Proof of Theorem 1

We start by showing that if

$$\exists t \in V \quad \text{such that} \quad \Lambda_t \notin (\gamma_n^{\mathrm{L}}, \gamma_n^{\mathrm{U}}), \tag{111}$$

then any sequential decision algorithm with the stopping rule τ_n^* and the detection rule δ_n^* specified in (12) and (13), respectively, is (α, β)-accurate. Given the structure of the stopping time, according to which the sampling process terminates as soon as Λ_t exits the band $(\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})$, the assumption in (111) is equivalent to having

$$\Lambda_{\tau^*} \notin (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}}). \tag{112}$$

Therefore, for P_n^0 we have

$$\begin{aligned}
\mathsf{P}_n^0 &= \mathbb{P}_0(\delta_n^* = 1) && (113)\\
&= \sum_{k=1}^{n} \mathbb{P}_0(\delta_n^* = 1, \tau_n^* = k) && (114)\\
&\overset{(13),\,(112)}{=} \sum_{k=1}^{n} \mathbb{P}_0(\Lambda_{\tau_n^*} \geq \gamma^{\mathrm{U}}, \tau_n^* = k) && (115)\\
&\overset{(112)}{=} \sum_{k=1}^{n} \mathbb{P}_0(\Lambda_{\tau_n^*} \geq \gamma^{\mathrm{U}}, \tau_n^* = k) && (116)\\
&= \sum_{k=1}^{n} \int_{(\Lambda_k \geq \gamma^{\mathrm{U}}, \tau_n^* = k)} f_0(Y^k; \psi^k)\, \mathrm{d}Y^k && (117)\\
&\overset{(9)}{=} \sum_{k=1}^{n} \int_{(\Lambda_k \geq \gamma^{\mathrm{U}}, \tau_n^* = k)} \exp(-\Lambda_k) f_1(Y^k; \psi^k)\, \mathrm{d}Y^k && (118)\\
&\leq \sum_{k=1}^{n} \int_{(\Lambda_k \geq \gamma^{\mathrm{U}}, \tau_n^* = k)} \exp(-\gamma^{\mathrm{U}}) f_1(Y^k; \psi^k)\, \mathrm{d}Y^k && (119)\\
&\overset{(11)}{=} \mathrm{e}^{-n\alpha} \sum_{k=1}^{n} \int_{(\Lambda_k \geq \gamma^{\mathrm{U}}, \tau_n^* = k)} f_1(Y^k; \psi^k)\, \mathrm{d}Y^k && (120)\\
&= \mathrm{e}^{-n\alpha} \sum_{k=1}^{n} \mathbb{P}_1(\delta_n^* = 1, \tau_n^* = k) && (121)\\
&= \mathrm{e}^{-n\alpha} \cdot \mathbb{P}_1(\delta_n^* = 1) && (122)\\
&\leq \mathrm{e}^{-n\alpha}, && (123)
\end{aligned}$$

where (115) holds according to the definition of the terminal decision rule in (13), (116) holds due to the assumption in (112), (118) holds due to the definition of LLR in (9), and (119) holds due to the structure of the region over which the integral is computed. Finally (121) holds by noting that the decision rule $\delta_n^* = 1$ specifies that $\Lambda_{\tau_n^*} > 0$, which by taking into account (112) and the fact that $\gamma^{\mathrm{L}} < 0$, becomes equivalent to $\Lambda_{\tau_n^*} \geq \gamma^{\mathrm{U}}$. By following the same line of argument for P_n^1 we obtain

$$\begin{aligned}
\mathsf{P}_n^1 &\leq \mathrm{e}^{\gamma^{\mathrm{L}}} \cdot \mathbb{P}_1(\delta_n^* = 0) && (124)\\
&= \mathrm{e}^{-n\beta} \cdot \mathbb{P}_1(\delta_n^* = 0) && (125)\\
&\leq \mathrm{e}^{-n\beta}. && (126)
\end{aligned}$$

Next, we analyze the likelihood of the condition in (111) being valid, which establishes a probabilistic guarantee for Algorithm 1 generating (α, β)-accurate solutions to $\mathcal{P}(\alpha, \beta)$.

$$\begin{aligned}
1 - \mathbb{P}\big(\exists t \in V \quad \text{s.t.} \quad \Lambda_t \notin (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})\big) &&(127)\\
= \mathbb{P}\left(\Lambda_t \in (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}}), \quad \forall t \in V\right) &&(128)\\
\leq \mathbb{P}\left(\Lambda_n \in (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})\right) &&(129)\\
= \sum_{i=0}^{1} \epsilon_i \mathbb{P}_i\left(\Lambda_n \in (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})\right). &&(130)
\end{aligned}$$

Next, for the probability terms in the right hand side we have

$$\begin{aligned}
\mathbb{P}_0(\Lambda_n \in (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})) &\leq \mathbb{P}_0(\Lambda_n > \gamma^{\mathrm{L}}) && (131)\\
&\leq \frac{1}{\sqrt{\exp(\gamma^{\mathrm{L}})}} \cdot \mathbb{E}_0\{\sqrt{\exp(\Lambda_n)}\} && (132)
\end{aligned}$$

$$\stackrel{(26)}{=} \frac{1}{\sqrt{\exp(\gamma^{\mathrm{L}})}} \cdot \mathsf{B}_n(f_0, f_1) \tag{133}$$

$$\stackrel{(11)}{=} \exp\left(\frac{n\beta}{2}\right) \cdot \mathsf{B}_n(f_0, f_1), \tag{134}$$

where (132) follows the Markov inequality. By following a similar line of argument we obtain

$$\mathbb{P}_1(\Lambda_n \in (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})) \leq \exp\left(\frac{n\alpha}{2}\right) \cdot \mathsf{B}_n(f_0, f_1). \tag{135}$$

Hence, from (130), (134), and (135) we obtain

$$\mathbb{P}\big(\exists t \in V \quad \text{s.t.} \quad \Lambda_t \notin (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})\big) \tag{136}$$

$$\geq\ 1 - \mathsf{B}_n(f_0, f_1) \left[\epsilon_0 \exp\left(\frac{n\beta}{2}\right) + \epsilon_1 \exp\left(\frac{n\alpha}{2}\right)\right]. \tag{137}$$

APPENDIX B
PROOF OF LEMMA 1

From the definition of $\kappa(f_0, f_1)$ in (27) we have

$$2\kappa(f_0, f_1) = -\lim_{n\to\infty} \frac{2}{n} \ln \mathsf{B}_n(f_0, f_1) \tag{138}$$

$$\stackrel{(26)}{=} -\lim_{n\to\infty} \frac{2}{n} \ln \int \sqrt{f_0(x;V) f_1(x;V)}\, \mathrm{d}x \tag{139}$$

$$= -\lim_{n\to\infty} \frac{2}{n} \ln \int \sqrt{\frac{f_0(x;V)}{f_1(x;V)}}\, f_1(x;V)\, \mathrm{d}x \tag{140}$$

$$\leq -\lim_{n\to\infty} \frac{2}{n} \int \ln \left(\sqrt{\frac{f_0(x;V)}{f_1(x;V)}}\right) f_1(x;V)\, \mathrm{d}x \tag{141}$$

$$= \lim_{n\to\infty} \int \frac{1}{n} \ln \left(\frac{f_1(x;V)}{f_0(x;V)}\right) f_1(x;V)\, \mathrm{d}x \tag{142}$$

$$\stackrel{(31)}{=} \lim_{n\to\infty} \mathbb{E}_1\left[\mathsf{nLLR}_1(X_V;V)\right], \tag{143}$$

where (141) holds due to Jensen's inequality. By definition, in a homogeneous network, when the limit exists, the term $\mathsf{nLLR}_1(X_V;V)$ converges completely to I_1. This, in turn, implies that $\mathbb{E}_1\left[\mathsf{nLLR}_1(X_V;V)\right]$ also converges completely to I_1, which, subsequently, converges almost surely to I_1 (complete convergence implies almost sure convergence [81]). Hence, in homogeneous networks

$$2\kappa(f_0, f_1) \leq \lim_{n\to\infty} \mathbb{E}_1\left[\mathsf{nLLR}_1(X_V;V)\right] \xrightarrow{\text{a.s.}} I_1. \tag{144}$$

For the heterogeneous networks, we will follow the same line of argument to show that

$$2\kappa(f_0, f_1) \leq \lim_{n\to\infty} \mathbb{E}_1\left[\mathsf{nLLR}_1(X_V;V)\right] \xrightarrow{\text{a.s.}} I_1(V) \stackrel{(38)}{\leq} I_1^*. \tag{145}$$

A similar line of argument also shows that almost surely $2\kappa(f_0, f_1) \leq I_0$ and $2\kappa(f_0, f_1) \leq I_0^*$ in homogeneous and heterogeneous networks. By noting the assumption $\max\{\alpha, \beta\} \leq 2\kappa(f_0, f_1)$, the desired conclusion is established.

APPENDIX C
PROOF OF THEOREM 2

In order to prove (46), we show that for any feasible solution to (8) and for all $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}_1\left(\frac{\tau_n}{n} > \frac{\alpha}{I_1 + \epsilon}\right) = 1. \tag{146}$$

This property, in turn, establishes the desired result in (46). Specifically, by applying the Markov inequality we obtain

$$\lim_{n\to\infty} \mathbb{E}_1\left\{\frac{\tau_n}{n} \cdot \frac{I_1}{\alpha}\right\} \geq \lim_{n\to\infty} \frac{I_1}{I_1 + \epsilon} \cdot \mathbb{P}_1\left(\frac{\tau_n}{n} \cdot \frac{I_1}{\alpha} > \frac{I_1}{I_1 + \epsilon}\right)$$

$$\stackrel{(146)}{=} \frac{I_1}{I_1 + \epsilon}, \qquad \forall \epsilon > 0. \tag{147}$$

Since the inequality in (147) is valid for all $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{E}_1\left\{\frac{\tau_n}{n} \cdot \frac{I_1}{\alpha}\right\} \geq \sup_{\epsilon>0} \frac{I_1}{I_1 + \epsilon} = 1, \tag{148}$$

which concludes (46). To prove (146), for $i \in \{0, 1\}$ and $L \in \{2, \ldots, n-1\}$, and corresponding to any (α, β)-accurate algorithm with stopping time τ_n and decision rule δ_n let us define the event

$$\mathcal{A}(i, L) \stackrel{\triangle}{=} \{\delta_n = i,\, \tau_n \leq L\}. \tag{149}$$

Then, for any $\zeta > 0$,for the error probability term P_n^0 when the stopping time is τ_n and the decision rule is δ_n, we have

$$\mathsf{P}_n^0 = \mathbb{P}_0(\delta_n = 1) \tag{150}$$

$$= \mathbb{E}_0\{\mathbb{1}_{\{\delta_n=1\}}\} \tag{151}$$

$$= \mathbb{E}_1\{\mathbb{1}_{\{\delta_n=1\}} \exp(-\Lambda_{\tau_n})\} \tag{152}$$

$$\geq \mathbb{E}_1\{\mathbb{1}_{\{\mathcal{A}(1,L), \Lambda_{\tau_n} < \zeta\}} \exp(-\Lambda_{\tau_n})\} \tag{153}$$

$$\geq \mathrm{e}^{-\zeta}\, \mathbb{P}_1(\mathcal{A}(1, L), \Lambda_{\tau_n} < \zeta) \tag{154}$$

$$\geq \mathrm{e}^{-\zeta}\, \mathbb{P}_1\Big(\mathcal{A}(1, L),\, \sup_{t<L} \Lambda_t < \zeta\Big) \tag{155}$$

$$\geq \mathrm{e}^{-\zeta} \left[\mathbb{P}_1(\mathcal{A}(1, L)) - \mathbb{P}_1\big(\sup_{t<L} \Lambda_t \geq \zeta\big)\right] \tag{156}$$

$$\geq \mathrm{e}^{-\zeta} \left[\mathbb{P}_1\delta_n = 1) - \mathbb{P}_1(\tau_n > L) - \mathbb{P}_1\big(\sup_{t<L} \Lambda_t \geq \zeta\big)\right], \tag{157}$$

where (152) holds by changing the probability measure, (153) holds by noting that the event $\{\mathcal{A}(1, L), \Lambda_{\tau_n} < \zeta\}$ is a subset of the event $\{\delta_n = 1\}$, and (156) and (157) hold due to basic set operations properties. By rearranging the terms in (150) and (157) and invoking $\mathbb{P}_0(\delta_n = 1) \leq \mathrm{e}^{-n\alpha}$ and $\mathbb{P}_1(\delta_n = 0) \leq \mathrm{e}^{-n\beta}$ (the decision rules are (α, β)-accurate) we obtain

$$\mathbb{P}_1(\tau_n > L) \ \geq\ \mathbb{P}_1(\delta_n = 1) - e^{\zeta}\, \mathbb{P}_0(\delta_n = 1) - \mathbb{P}_1\big(\sup_{t<L} \Lambda_t \geq \zeta\big) \tag{158}$$

$$= 1 - \mathsf{P}_n^1 - e^{\zeta}\, \mathsf{P}_n^0 - \mathbb{P}_1\big(\sup_{t<L} \Lambda_t \geq \zeta\big) \tag{159}$$

$$\stackrel{(8)}{\geq}\ 1 - \mathrm{e}^{-n\beta} - e^{\zeta}\, \mathrm{e}^{-n\alpha} - \mathbb{P}_1\big(\sup_{t<L} \Lambda_t \geq \zeta\big). \tag{160}$$

Note that (160) holds for any $\zeta > 0$. Next, we set $\zeta \triangleq cLI_1$ where

$$c \triangleq 1 + \frac{\epsilon}{2I_1}. \tag{161}$$

Hence, for any $K \in \{2, \ldots, L-1\}$ for the last term in (160) we have

$$\mathbb{P}_1\Big(\sup_{t<L} \Lambda_t \geq \zeta\Big) \tag{162}$$

$$= \mathbb{P}_1\Big(\sup_{t<L} \Lambda_t \geq cLI_1\Big) \tag{163}$$

$$\leq \mathbb{P}_1\Big(\sup_{t<K} \Lambda_t + \sup_{K\leq t<L} \Lambda_t \geq cLI_1\Big) \tag{164}$$

$$\leq \mathbb{P}_1\Big(\sup_{t<K} \Lambda_t + \sup_{K\leq t<L} \Big\{\frac{L}{t}\Lambda_t\Big\} \geq cLI_1\Big) \tag{165}$$

$$= \mathbb{P}_1\Big(\frac{1}{L}\sup_{t<K} \Lambda_t + \sup_{K\leq t<L} \Big\{\frac{\Lambda_t}{t} - I_1\Big\} \geq (c-1)I_1\Big) \tag{166}$$

$$\leq \mathbb{P}_1\Big(\frac{1}{L}\sup_{t<K} \Lambda_t + \sup_{t\geq K} \Big|\frac{\Lambda_t}{t} - I_1\Big| \geq (c-1)I_1\Big) \tag{167}$$

$$\overset{(161)}{=} \mathbb{P}_1\Big(\frac{1}{L}\sup_{t<K} \Lambda_t + \sup_{t\geq K} \Big|\frac{\Lambda_t}{t} - I_1\Big| \geq \frac{\epsilon}{2}\Big) \tag{168}$$

$$\leq \mathbb{P}_1\Big(\frac{1}{L}\sup_{t<K} \Lambda_t \geq \frac{\epsilon}{4}\Big) + \mathbb{P}\Big(\sup_{t>K} \Big|\frac{\Lambda_t}{t} - I_1\Big| > \frac{\epsilon}{4}\Big). \tag{169}$$

We show that both probability terms in (169) diminish as n grows. From the second term in (169) note that from the definition of $T_\ell(h, \psi^\infty)$ in (35) we know that corresponding to any given sampling path ψ^∞ we have

$$\forall t \ \geq T_\ell\left(\frac{\epsilon}{4}, \psi^\infty\right): \quad \Big|\frac{\Lambda_t}{t} - I_1\Big| \leq \frac{\epsilon}{4} \quad . \tag{170}$$

This indicates that by setting $K = T_\ell(\frac{\epsilon}{4}, \psi^\infty)$, it can be readily verified that

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{t>K} \Big|\frac{\Lambda_t}{t} - I_1\Big| > \frac{\epsilon}{4}\Big) = 0. \tag{171}$$

As a result, for $K = T_\ell(\epsilon/4, \psi^\infty)$ from (162)-(169) we obtain

$$\lim_{n\to\infty} \mathbb{P}_1\Big(\sup_{t<L} \Lambda_t \geq cLI_1\Big) \leq \lim_{n\to\infty} \mathbb{P}_1\Big(\frac{1}{L}\sup_{t\leq K} \Lambda_t \geq \frac{\epsilon}{4}\Big). \tag{172}$$

For the right hand side of (172) we find that for any $\epsilon > 0$

$$\mathbb{P}_1\Big(\frac{1}{L}\sup_{t<K} \Lambda_t \geq \frac{\epsilon}{4}\Big) \ \leq \ \mathbb{P}_1\Big(\frac{1}{L}\sum_{t=1}^{K} \Lambda_t \geq \frac{\epsilon}{4}\Big) \tag{173}$$

$$\leq \frac{4}{\epsilon}\cdot\frac{1}{L}\mathbb{E}_1\left[\sum_{t=1}^{K} \Lambda_t\right] \tag{174}$$

$$= \ \frac{4}{\epsilon}\cdot\frac{1}{L}\mathbb{E}_1\left[\sum_{t=1}^{K} \mathbb{E}_1[\Lambda_t]\right] \tag{175}$$

$$\leq \ \frac{4}{\epsilon}\cdot\frac{1}{L}\mathbb{E}_1[K] \max_{1\leq t\leq K} \mathbb{E}_1[\Lambda_t], \tag{176}$$

where (174) holds due to Markov's inequality and (175) follows from Wald's identity (general form). Next, we set

$$L = \left\lceil \frac{n\alpha}{I_1 + \epsilon} \right\rceil. \tag{177}$$

By recalling Lemma 1 we know that for sufficiently large n, we have $L \leq n$. Hence, based on (172) and (176) we get

$$\lim_{n\to\infty} \mathbb{P}_1\Big(\sup_{t<L} \Lambda_t \geq cLI_1\Big)$$

$$\leq \lim_{n\to\infty} \frac{4}{\epsilon}\cdot\frac{1}{L}\mathbb{E}_1[K] \max_{1\leq t\leq K} \mathbb{E}_1[\Lambda_t] \tag{178}$$

$$\overset{(177)}{\leq} \frac{4}{\epsilon}\cdot\frac{I_1+\epsilon}{\alpha} \lim_{n\to\infty} \frac{1}{n}\, \mathbb{E}_1[K] \max_{1\leq t\leq K} \mathbb{E}_1[\Lambda_t] \tag{179}$$

$$= 0, \tag{180}$$

where the last step holds by noting that $\mathbb{E}_\ell[K] = \mathbb{E}_\ell[T_\ell(\epsilon/4, \psi^\infty)] < +\infty$ specified in (34). Subsequently, from (162), (169), (171), (178), and (180) we have

$$\lim_{n\to\infty} \mathbb{P}_1\Big(\sup_{t<L} \Lambda_t \geq \zeta\Big) = 0. \tag{181}$$

As a result, from (158)-(160) we obtain

$$\lim_{n\to\infty} \mathbb{P}_1\left(\frac{\tau_n}{n} > \frac{\alpha}{I_1+\epsilon}\right)$$

$$\overset{(177)}{=} \lim_{n\to\infty} \mathbb{P}_1\left(\tau_n > L\right) \tag{182}$$

$$\overset{(160)}{\underset{(181)}{=}} \lim_{n\to\infty} \left[1 - \exp(-n\beta) - \exp\Big(-n\alpha\cdot\frac{\epsilon}{2I_1+\epsilon}\Big)\right] \tag{183}$$

$$= 1, \tag{184}$$

which proves (146). Since this is always valid irrespectively of the sampling procedure and the stopping rule, we conclude that (46) is always valid, establishing

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n\}}{n} \geq \frac{\alpha}{I_1}. \tag{185}$$

We can prove

$$\lim_{n\to\infty} \frac{\mathbb{E}_0\{\tau_n\}}{n} \geq \frac{\beta}{I_0}, \tag{186}$$

by following the same line of argument.

Appendix D
Proof of Theorem 3

Following the definition of $T_1(h, \psi^\infty)$ in (35), we provide a truncated counterpart of it for a network with n nodes (non-asymptotic regime) as follows.

$$R_1(h, \psi^n) \triangleq \sup\ \Big\{t \leq n\ :\ \Big|\frac{\Lambda_t}{t} - I_1\Big| > h\Big\}, \quad \forall h > 0, \tag{187}$$

where we adopt the convention that the supremum of an empty set is $+\infty$. Obviously,

$$\lim_{n\to\infty} R_1(h, \psi^n) = T_1(h, \psi^\infty). \tag{188}$$

According to the definition of the stopping time in (12), at the instance prior to stopping, i.e., at time $\tau_n^* - 1$, we always

have $\Lambda_{\tau_n^*-1} \in (\gamma^{\mathrm{L}}, \gamma^{\mathrm{U}})$. We start the proof by comparing $\Lambda_{\tau_n^*-1}$ with these two bounds. First, consider the following relationship

$$\Lambda_{\tau_n^*-1} < \gamma^{\mathrm{U}}. \tag{189}$$

Based on the definition of $R_1(h,\psi^n)$ in (187), if $R_1(h,\psi^{\tau_n^*}) < \tau_n^* - 1$, then for $t = \tau_n^* - 1$ we have

$$\Big|\frac{\Lambda_{\tau_n^*-1}}{\tau_n^*-1} - I_1\Big| \le h, \quad \forall h > 0, \tag{190}$$

which indicates that for all $h \in (0, I_1)$ we have

$$\tau_n^* \le \frac{\Lambda_{\tau_n^*-1}}{I_1 - h} + 1 \overset{(189)}{\le} \frac{\gamma^{\mathrm{U}}}{I_1 - h} + 1. \tag{191}$$

Hence, from (191) for all $h \in (0, I_1)$ we have

$$\tau_n^* = \tau_n^* \cdot \mathbb{1}_{\{\tau_n^* > R_1(h,\psi^{\tau_n^*})+1\}} + \underbrace{\tau_n^* \cdot \mathbb{1}_{\{\tau_n^* \le R_1(h,\psi^{\tau_n^*})+1\}}}_{\le R_1(h,\psi^{\tau_n^*})+1} \tag{192}$$

$$\overset{(191)}{\le} \left[\frac{\gamma^{\mathrm{U}}}{I_1 - h} + 1\right] \cdot \mathbb{1}_{\{\tau_n^* > R_1(h,\psi^{\tau_n^*})+1\}} + R_1(h,\psi^{\tau_n^*}) + 1 \tag{193}$$

$$\le 2 + \frac{\gamma^{\mathrm{U}}}{I_1 - h} + R_1(h,\psi^{\tau_n^*}). \tag{194}$$

Subsequently,

$$\tau_n^* \le 2 + \inf_{h \in (0, I_1)} \frac{\gamma^{\mathrm{U}}}{I_1 - h} + R_1(h,\psi^{\tau_n^*}) \tag{195}$$

$$\le 2 + \frac{\gamma^{\mathrm{U}}}{I_1} + R_1(h,\psi^{\tau_n^*}). \tag{196}$$

Since

$$\mathbb{E}_1\{T_1(h,\psi^\infty)\} < +\infty, \quad \forall h > 0, \tag{197}$$

by recalling that $\gamma^{\mathrm{U}} = n\alpha$, from (188) and (195)-(196) we obtain

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n^*\}}{n} \le \frac{\alpha}{I_1}. \tag{198}$$

Similarly, by also considering

$$\Lambda_{\tau_n^*-1} > \gamma^{\mathrm{L}} \tag{199}$$

and following the same line of argument we obtain

$$\lim_{n\to\infty} \frac{\mathbb{E}_0\{\tau_n^*\}}{n} \le \frac{\beta}{I_0}, \tag{200}$$

which concludes the proof.

Appendix E
Proof of Lemma 2

We start by showing that there exist positive constants B and c such that for all $t \in V$

$$\mathbb{P}_1(\hat{\tau}_n \ge t) \le B\mathrm{e}^{-ct}. \tag{201}$$

For this purpose, note that

$$\mathbb{P}_1(\hat{\tau}_n \ge t) = \sum_{u=t}^{n} \mathbb{P}_1(\hat{\tau}_n = u) \tag{202}$$

$$= \sum_{u=t}^{n} \mathbb{P}_1\Big(\delta_{\mathrm{ML}}(u-1) = \mathsf{H}_0, \quad \delta_{\mathrm{ML}}(u) = \cdots = \delta_{\mathrm{ML}}(n) = \mathsf{H}_1\Big) \tag{203}$$

$$\le \sum_{u=t}^{n} \mathbb{P}_1(\delta_{\mathrm{ML}}(u-1) = \mathsf{H}_0) \tag{204}$$

$$\overset{(15)}{=} \sum_{u=t-1}^{n-1} \mathbb{P}_1(\Lambda_u < 0). \tag{205}$$

Next, we find an upper bound on $\mathbb{P}_1(\Lambda_u < 0)$. For this purpose, note that for any $s \in \mathbb{R}$ we have

$$\mathbb{P}_1(\Lambda_t < 0) \cdot \mathbb{E}_1\big\{\exp\{s\Lambda_t\} \mid \mathbb{1}_{\{\Lambda_t<0\}}\big\} \tag{206}$$

$$= \mathbb{E}_1\big\{\exp\{s\Lambda_t\}\mathbb{1}_{\{\Lambda_t<0\}}\big\} \tag{207}$$

$$\le \mathbb{E}_1\big\{\exp\{s\Lambda_t\}\big\}. \tag{208}$$

Furthermore, for any $s < 0$ we have

$$\mathbb{E}_1\big\{\exp\{s\Lambda_t\} \mid \mathbb{1}_{\{\Lambda_t<0\}}\big\} \ge 1. \tag{209}$$

By combining (206)–(209) we find that for any $s < 0$

$$\mathbb{P}_1(\Lambda_t < 0) \le \mathbb{E}_1\big\{\exp\{s\Lambda_t\}\big\}. \tag{210}$$

The right hand side of (210) can be expanded by using the towering property of expectation as follows:

$$\mathbb{E}_1\Big\{\exp\{s\Lambda_t\}\Big\} \tag{211}$$

$$\overset{(9)}{=} \mathbb{E}_1\Big\{\exp\{s\Lambda_{t-1}\} \cdot \mathbb{E}_1\Big\{\Big[\frac{f_1(Y_t;\psi(t)|\mathcal{F}_{t-1})}{f_0(Y_t;\psi(t)|\mathcal{F}_{t-1})}\Big]^s \;\Big|\; \mathcal{F}_{t-1}\Big\}\Big\}. \tag{212}$$

Now, consider the inner expectation and define

$$\xi_t(s) \triangleq \mathbb{E}_1\Big\{\Big[\frac{f_1(Y_t;\psi(t)|\mathcal{F}_{t-1})}{f_0(Y_t;\psi(t)|\mathcal{F}_{t-1})}\Big]^s \;\Big|\; \mathcal{F}_{t-1}\Big\}. \tag{213}$$

It can be ready verified that $\xi_t(s)$ is convex in s and satisfies

$$\xi_t(-1) = \xi_t(0) = 1. \tag{214}$$

$\xi_t(s)$ can have two possible behaviors in the range $s \in (-1, 0)$:

Case 1: $\xi_t(s) = 1,\ \forall s \in (-1, 0)$. This occurs only when the likelihood ratio inside the expectation is equal to 1, i.e., the sample taken at time t has the same likelihood values under both hypotheses. This event has measure zero. As a result, the probability of this case occurring is 0.

Case 2: $\xi_t(s) < 1,\ \forall s \in (-1, 0)$. It means that in this case there exists a constant $c > 0$ such that for some $s^* \in (-1, 0)$ and $\forall t \le \tau_n^*$

$$\xi_t(s^*) \le \mathrm{e}^{-c} < 1. \tag{215}$$

By successively applying the towering property as in (211), and accounting for Case 1 we obtain

$$\mathbb{P}_1(\Lambda_t < 0) \overset{(210)}{\le} \mathbb{E}_1\Big\{\exp\{s^*\Lambda_t\}\Big\} \le \mathrm{e}^{-ct}. \tag{216}$$

Next, by combining (205) and (216) we obtain

$$\mathbb{P}_1(\hat{\tau}_n \geq t) \leq \sum_{u=t-1}^{n-1} \mathrm{e}^{-cu} \tag{217}$$

$$\leq \sum_{u=t-1}^{\infty} \mathrm{e}^{-cu} \tag{218}$$

$$= \frac{\mathrm{e}^{c}}{1-\mathrm{e}^{-c}} \mathrm{e}^{-ct} \tag{219}$$

$$= b \cdot \mathrm{e}^{-ct}, \tag{220}$$

where we have defined $b \triangleq \frac{1}{1-\mathrm{e}^{-c}}$. By using this result, it can be ready verified that $\mathbb{E}_1\{\hat{\tau}_n\}$ is finite. Specifically,

$$\mathbb{E}_1\{\hat{\tau}_n\} = \sum_{t=1}^{\infty} \mathbb{P}(\hat{\tau}_n \geq t) \leq \sum_{t=0}^{\infty} b\mathrm{e}^{-ct} = \frac{b}{1-\mathrm{e}^{-c}}, \tag{221}$$

which shows that $\mathbb{E}_1\{\hat{\tau}_n\}$ is asymptotically upper bounded by a constant. The proof for $\mathbb{E}_0\{\hat{\tau}_n\}$ being bounded by a constant follows a similar line of argument.

Appendix F
Proof of Lemma 3

We prove the lemma by contradiction. Specifically, we show that if $\lim_{t\to\infty} \frac{1}{t}|\mathcal{H}_t|$ is bounded away from zero, then for the sequence $W^t \triangleq \{\psi^*(1), \ldots, \psi^*(t)\}$ we have $I_\ell(W^\infty) > I_\ell(U^\infty)$, contradicting the definition of U^∞ in (56). For this purpose, for $t > \hat{\tau}_n$ consider the expansion

$$\mathbb{E}_\ell\left[\frac{1}{t}\ln\frac{f_\ell(Y^t;W^t)}{f_{1-\ell}(Y^t;W^t)}\right]$$

$$= \frac{1}{t}\sum_{s=1}^{t}\mathbb{E}_\ell\left[\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right] \tag{222}$$

$$= \frac{1}{t}\mathbb{E}_\ell\left[\sum_{s=1}^{\hat{\tau}_n}\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right] + \frac{1}{t}\mathbb{E}_\ell\left[\sum_{s\notin\mathcal{H}_t}\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right] + \frac{1}{t}\mathbb{E}_\ell\left[\sum_{s\in\mathcal{H}_t}\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right]. \tag{223}$$

Corresponding to the three summands in (223) we show the following three properties:

$$\frac{1}{t}\sum_{s=1}^{\hat{\tau}_n}\mathbb{E}_\ell\left[\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right] \geq 0, \tag{224}$$

$$\frac{1}{t}\sum_{s\notin\mathcal{H}_t}\mathbb{E}_\ell\left[\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right] \geq \frac{t-|\mathcal{H}_t|}{t}\cdot I_\ell(U^\infty), \tag{225}$$

$$\frac{1}{t}\sum_{s\in\mathcal{H}_t}\mathbb{E}_\ell\left[\ln\frac{f_\ell(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi^*(s)|\mathcal{F}_{s-1})}\right] \geq \frac{|\mathcal{H}_t|}{t}[I_\ell(U^\infty)+\varepsilon], \tag{226}$$

where $\varepsilon \in \mathbb{R}^+$ is a positive constant. These three inequalities in conjunction with (223) establish that

$$\mathbb{E}_\ell\left[\frac{1}{t}\ln\frac{f_\ell(Y^t;W^t)}{f_{1-\ell}(Y^t;W^t)}\right] \geq I_\ell(U^\infty) + \frac{|\mathcal{H}_t|}{t}\cdot\varepsilon. \tag{227}$$

Hence, when $\lim_{t\to\infty}\frac{1}{t}|\mathcal{H}_t|$ is bounded away from zero we have

$$I_\ell(W^\infty) = \lim_{t\to\infty}\mathbb{E}_\ell\left[\frac{1}{t}\ln\frac{f_\ell(Y^t;W^t)}{f_{1-\ell}(Y^t;W^t)}\right] > I_\ell(U^\infty), \tag{228}$$

which contradicts the definition of U^∞. By noting that (224) follows the non-negativity of the KL divergence, next we prove the inequalities in (225) and (226).

We complete the proof in four steps. In the first step, we show that for any set $A \subseteq U^\infty \cap \mathcal{S}(\mathbb{N})$, the normalized log-likelihood ratio of the random variables in the set A converges to $I_\ell(U^\infty)$. In the second step, we leverage this property to establish that the information measure of each node $j \in U^\infty \cap \varphi^t$ is greater than or equal to $I_\ell(U^\infty)$. In the third step, we show that the contribution of any node in $\mathcal{H}(t)$ to the normalized log-likelihood ratio is greater than $I_\ell(U^\infty)$. Finally, in the fourth step we show that if $\lim_{t\to\infty}\frac{|\mathcal{H}(t)|}{t} > 0$, then for the set $\hat{U} \triangleq U^\infty \cup \mathcal{H}(t)$ we have $I_\ell(\hat{U}) > I_\ell(U^\infty)$.

Step 1: We prove that for any set $A \subseteq U^\infty \cap \mathcal{S}(\mathbb{N})$, the normalized log-likelihood ratio of any infinite subsequence of the set A converges to $I_\ell(U^\infty)$. From the definition of the set U^∞ and complete convergence in (33), for any sequence of nodes $\psi^\infty \subseteq U^\infty$ and $\forall h > 0$, we have

$$\sum_{t=1}^{\infty}\mathbb{P}_\ell\Big(\big|\mathsf{nLLR}_\ell(Y^t;\psi^t) - I_\ell(U^\infty)\big| > h\Big) < \infty. \tag{229}$$

Since (229) holds for any infinite subsequence of U^∞, it holds for any subsequence $\hat{\psi}^\infty \subseteq A \subseteq U^\infty$ as well. Hence,

$$\sum_{t=1}^{\infty}\mathbb{P}_\ell\Big(\big|\mathsf{nLLR}_\ell(Y^t;\hat{\psi}^t) - I_\ell(U^\infty)\big| > h\Big) < \infty, \tag{230}$$

which proves the complete convergence of $\mathsf{nLLR}_\ell(Y^t;\hat{\psi}^t)$ to $I_\ell(U^\infty)$ for any set $A \subseteq U^\infty \cap \mathcal{S}(\mathbb{N})$.

Step 2: By leveraging the property established in Step 1, we show that at any time $t \geq \hat{\tau}_n$, if the ML decision at time $t-1$ is H_ℓ, i.e., $\delta_{\mathrm{ML}}(t-1) = \mathsf{H}_\ell$, then for any node $j \in U^\infty$ we have

$$\max_{\mathcal{S}\in\mathcal{R}_t^j}\mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}\right\} \geq I_\ell(U^\infty). \tag{231}$$

We first show this property for $t = \hat{\tau}_n$. For this purpose, expand the left hand side of (231) as in (232)–(238), shown at the bottom of the next page, where (233) is due to shrinking the feasible set of the maximization problem, expressions in (234) and (236) are due to the properties of conditional pdfs, (235) and (237) follow Lemma 2, which states that $\hat{\tau}_n$ is upper bounded by a constant, while we have $\mathcal{S} \in \mathcal{S}(\mathbb{N})$ dictating that $|\mathcal{S}| = \infty$, and (238) follows from the definition of U^∞ and applying Step 1. From (233)–(238) we observe that for any node $j \in U^\infty$ we have

$$\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat{\tau}_n-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat{\tau}_n-1})}\right\} \geq I_\ell(U^\infty). \tag{239}$$

Furthermore, by noting that we are restricting $\mathcal{S}$ to belong to $\mathcal{S}(\mathbb{N})$, the inequality will hold with equality, i.e.

$$\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\ \mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat\tau_n-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat\tau_n-1})}\right\}=I_\ell(U^\infty). \tag{240}$$

By following a similar line of argument, we generalize this property to any $t>\hat\tau_n$. By defining $\mathcal{V}\stackrel{\triangle}{=}\{\psi(s)\ :\ \hat\tau_n\le s\le t\}$,[6] for the left hand side of (231) we have what follows in (241)–(244), shown at the bottom of the next page, where (242) follows as in (234), the expansion of the conditional pdf gives (243), and (244) holds due to $\hat\tau_n$ being upper bounded by a constant. Finally, if $|\mathcal{V}|$ is finite, the second term in (244) is 0 because the conditional KL divergence is finite and $|\mathcal{S}|=\infty$ and (244) is lower bounded by $I_\ell(U^\infty)$ as shown in (235)–(238). On the other hand, if $|\mathcal{V}|\to\infty$, from (244) we have (245)–(247), shown at the bottom of the next page, where (246) is due to replacing the second term in (245) from (240), and (247) is due to the definition of U^∞ and Step 1. Hence, the proof for (231) is concluded.

Step 3: In Step 2, we proved that at any time $t\ge\hat\tau_n$ there exists a node from U^∞ for which the information measure is greater than or equal to $I_\ell(U^\infty)$. Next, we prove that the contribution of any node in $\mathcal{H}(t)$ to the normalized log-likelihood ratio is greater than $I_\ell(U^\infty)$. To this end, we note that for any $s\in\mathcal{H}(t)$ there exists some $\mathcal{S}\in\mathcal{R}_s^{\psi(s)}$ such that

$$\frac{1}{|\mathcal{S}|}\mathbb{E}_\ell\left\{\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{s-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{s-1})}\right\}>I_\ell(U^\infty), \tag{249}$$

because otherwise, according to Step 2, one node from the set U^∞ should have been selected. By defining $\bar{\mathcal{S}}\stackrel{\triangle}{=}\mathcal{S}\setminus\{\psi(s)\}$ and expanding the joint pdfs we obtain

$$\mathbb{E}_\ell\left\{\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{s-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{s-1})}\right\} \tag{250}$$

$$=\mathbb{E}_\ell\left\{\ln\frac{f_\ell(Y_s;\psi(s)\,|\,\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi(s)\,|\,\mathcal{F}_{s-1})}\right\} \tag{251}$$

$$+\mathbb{E}_\ell\left\{\ln\frac{f_\ell(X_{\bar{\mathcal{S}}};\bar{\mathcal{S}}\,|\,\mathcal{F}_{s-1},Y_s)}{f_{1-\ell}(X_{\bar{\mathcal{S}}};\bar{\mathcal{S}}\,|\,\mathcal{F}_{s-1},Y_s)}\right\}. \tag{252}$$

If the first term on the right-hand side of (250) is not greater than $I_\ell(U^\infty)$, the from (249) we conclude that the second term in (250) must be greater than $|\bar{\mathcal{S}}|\cdot I_\ell(U^\infty)$. Thus, for any node in $\bar{\mathcal{S}}$ we have

$$\frac{1}{|\bar{\mathcal{S}}|}\mathbb{E}_\ell\left\{\ln\frac{f_\ell(X_{\bar{\mathcal{S}}};\bar{\mathcal{S}}\,|\,\mathcal{F}_{s-1},Y_s)}{f_{1-\ell}(X_{\bar{\mathcal{S}}};\bar{\mathcal{S}}\,|\,\mathcal{F}_{s-1},Y_s)}\right\}>I_\ell(U^\infty). \tag{253}$$

Hence, one node from $\bar{\mathcal{S}}$ is selected at time $s+1$. Therefore, by sequentially applying this principle, by construction

$$\forall s\in\mathcal{H}(t),\exists\bar{\mathcal{S}}\subseteq\psi^t\ : \tag{254}$$

$$\frac{1}{|\bar{\mathcal{S}}|}\mathbb{E}_\ell\left\{\ln\frac{f_\ell(X_{\bar{\mathcal{S}}};\bar{\mathcal{S}}\,|\,\mathcal{F}_{s-1})}{f_{1-\ell}(X_{\bar{\mathcal{S}}};\bar{\mathcal{S}}\,|\,\mathcal{F}_{s-1})}\right\}>I_\ell(U^\infty). \tag{255}$$

Step 4: Finally, we prove the if $\lim_{t\to\infty}\frac{|\mathcal{H}(t)|}{t}=0$ does not hold, it contradicts the definition of U^∞. For this purpose, suppose that we have

$$\lim_{t\to\infty}\frac{|\mathcal{H}(t)|}{t}\ge c>0, \tag{256}$$

for some constant $c>0$. Then, we can expand the log-likelihood ratio of the samples up to time $t\in\mathbb{N}$ according

[6] In principle, set $\mathcal{V}$ should be indexed by t. However, for clarity in notations, we drop index t.

$$\max_{\mathcal{S}\in\mathcal{R}_t^j}\ \mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat\tau_n-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat\tau_n-1})}\right\} \tag{232}$$

$$\ge\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\ \mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat\tau_n-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{\hat\tau_n-1})}\right\} \tag{233}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\left[\frac{|\mathcal{S}|+\hat\tau_n-1}{|\mathcal{S}|}\,\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\mathcal{S}\cup\psi^{\hat\tau_n-1}};\mathcal{S}\cup\psi^{\hat\tau_n-1})\}-\underbrace{\frac{\hat\tau_n-1}{|\mathcal{S}|}\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\psi^{\hat\tau_n-1}};\psi^{\hat\tau_n-1})\}}_{=0\ \text{(Lemma 2)}}\right] \tag{234}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\frac{|\mathcal{S}|+\hat\tau_n-1}{|\mathcal{S}|}\,\mathbb{E}_\ell\left\{\mathsf{nLLR}(X_{\mathcal{S}\cup\psi^{\hat\tau_n-1}};\mathcal{S}\cup\psi^{\hat\tau_n-1})\right\} \tag{235}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\left[\mathbb{E}_\ell\{\mathsf{nLLR}(X_\mathcal{S};\mathcal{S})\}+\underbrace{\frac{1}{|\mathcal{S}|}\mathbb{E}_\ell\left\{\ln\frac{f_\ell(Y^{\hat\tau_n-1};\psi^{\hat\tau_n-1}|X_\mathcal{S};\mathcal{S})}{f_{1-\ell}(Y^{\hat\tau_n-1};\psi^{\hat\tau_n-1}|X_\mathcal{S};\mathcal{S})}\right\}}_{=0\ \text{(Lemma 2)}}\right] \tag{236}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\ \mathbb{E}_\ell\{\mathsf{nLLR}(X_\mathcal{S};\mathcal{S})\} \tag{237}$$

$$=I_\ell(U^\infty). \tag{238}$$

to

$$\frac{1}{t}\ln\frac{f_\ell(Y^t;\psi^t)}{f_{1-\ell}(Y_t;\psi^t)} \tag{257}$$

$$= \frac{1}{t}\sum_{s=1}^{t}\ln\frac{f_\ell(Y_s;\psi(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi(s)|\mathcal{F}_{s-1})} \tag{258}$$

$$= \frac{1}{t}\underbrace{\sum_{s\in\mathcal{H}(t)}\ln\frac{f_\ell(Y_s;\psi(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi(s)|\mathcal{F}_{s-1})}}_{\overset{(254)}{>}|\mathcal{H}(t)|\cdot I_\ell(U^\infty)} \tag{259}$$

$$+\frac{1}{t}\underbrace{\sum_{s\le t,s\notin\mathcal{H}(t)}\ln\frac{f_\ell(Y_s;\psi(s)|\mathcal{F}_{s-1})}{f_{1-\ell}(Y_s;\psi(s)|\mathcal{F}_{s-1})}}_{\overset{(231)}{\ge}(t-|\mathcal{H}(t)|)\cdot I_\ell(U^\infty)} \tag{260}$$

$$> \frac{|\mathcal{H}(t)|}{t}I_\ell(U^\infty)+\frac{t-|\mathcal{H}(t)|}{t}I_\ell(U^\infty) \tag{261}$$

$$> I_\ell(U^\infty). \tag{262}$$

This inequality holds for $t\to\infty$, which means that we have found a subset of nodes ψ^t for which the information measure $I_\ell(\psi^t)$ is greater than that of the set U^∞ under H_ℓ, contradicting the definition of the set U^∞. Hence,

$$\lim_{t\to\infty}\frac{|\mathcal{H}(t)|}{t}=0. \tag{263}$$

Appendix G
Proof of Lemma 4

Following the definition of $T_1(h,\psi^\infty)$ in (35) for heterogeneous networks, we provide a truncated counterpart for it defined as follows. For this purpose, $\forall h>0$ we denote the first n elements of U^∞ by U^n.

$$R_1(h,U^n)\triangleq\sup\left\{t\le n\,:\,\left|\frac{\Lambda_t}{t}-I_1^*\right|>h\right\}, \tag{264}$$

where clearly

$$\lim_{n\to\infty}R_1(h,U^n)=T_1(h,U^\infty). \tag{265}$$

Hence, by accounting for the first $\hat\tau_n$ samples, some of which may have been observed from nodes not included in the first $\hat\tau_n$ elements of U^∞, the last time that the normalized log-likelihood ratios $\frac{\Lambda_t}{t}$ leaves the interval $[I_1^*-h,I_1^*+h]$ will happen no later than $R_1(h,U^n)+\hat\tau_n+|\mathcal{H}(n)|$. In other words,

$$\forall t\ge R_1(h,U^n)+\hat\tau_n+|\mathcal{H}(n)|\ : \tag{266}$$

$$-h\le\frac{\Lambda_t-\Lambda_{\hat\tau_n}}{(t-\hat\tau_n)}-I_1^*\le h,\quad\forall h>0. \tag{267}$$

If $\tau_n^*>R_1(h,U^n)+\hat\tau_n+|\mathcal{H}(n)|$, then for all $h\in(0,I_1^*)$ we have

$$\tau_n^*-\hat\tau_n\le\frac{\gamma^{\mathrm{U}}-\Lambda_{\hat\tau_n}}{I_1^*-h}+1. \tag{268}$$

$$\max_{\mathcal{S}\in\mathcal{R}_t^j}\mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}\right\} \tag{241}$$

$$\ge\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}{f_{1-\ell}(X_\mathcal{S};\mathcal{S}\,|\,\mathcal{F}_{t-1})}\right\} \tag{242}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\Bigg[\frac{|\mathcal{S}|+|\mathcal{V}|+\hat\tau_n-1}{|\mathcal{S}|}\,\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\mathcal{S}\cup\psi^{t-1}};\mathcal{S}\cup\psi^{t-1})\}$$
$$-\mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{V};\mathcal{V}\,|\,\mathcal{F}_{\hat\tau_n-1})}{f_{1-\ell}(X_\mathcal{V};\mathcal{V}\,|\,\mathcal{F}_{\hat\tau_n-1})}\right\}-\underbrace{\frac{\hat\tau_n-1}{|\mathcal{S}|}}_{=0\ \text{(Lemma 2)}}\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\psi^{\hat\tau_n-1}};\psi^{\hat\tau_n-1})\}\Bigg] \tag{243}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\left[\frac{|\mathcal{S}|+|\mathcal{V}|+\hat\tau_n-1}{|\mathcal{S}|}\,\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\mathcal{S}\cup\psi^{t-1}};\mathcal{S}\cup\psi^{t-1})\}-\frac{1}{|\mathcal{S}|}\mathbb{E}_\ell\left\{\ln\frac{f_\ell(X_\mathcal{V};\mathcal{V}\,|\,\mathcal{F}_{\hat\tau_n-1})}{f_{1-\ell}(X_\mathcal{V};\mathcal{V}\,|\,\mathcal{F}_{\hat\tau_n-1})}\right\}\right], \tag{244}$$

$$\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\left[\frac{|\mathcal{S}|+|\mathcal{V}|+\hat\tau_n-1}{|\mathcal{S}|}\,\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\mathcal{S}\cup\psi^{t-1}};\mathcal{S}\cup\psi^{t-1})\}-\mathbb{E}_\ell\left\{\frac{1}{|\mathcal{S}|}\ln\frac{f_\ell(X_\mathcal{V};\mathcal{V}\,|\,\mathcal{F}_{\hat\tau_n-1})}{f_{1-\ell}(X_\mathcal{V};\mathcal{V}\,|\,\mathcal{F}_{\hat\tau_n-1})}\right\}\right] \tag{245}$$

$$=\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\left[\frac{|\mathcal{S}|+|\mathcal{V}|+\hat\tau_n-1}{|\mathcal{S}|}\,\mathbb{E}_\ell\{\mathsf{nLLR}(X_{\mathcal{S}\cup\psi^{t-1}};\mathcal{S}\cup\psi^{t-1})\}-\frac{|\mathcal{V}|}{|\mathcal{S}|}I_\ell(U^\infty)\right] \tag{246}$$

$$\ge\max_{\mathcal{S}\in\mathcal{R}_t^j\cap\,\mathcal{S}(\mathbb{N})}\left[\frac{|\mathcal{S}|+|\mathcal{V}|+\hat\tau_n-1}{|\mathcal{S}|}I_\ell(U^\infty)-\frac{|\mathcal{V}|}{|\mathcal{S}|}I_\ell(U^\infty)\right] \tag{247}$$

$$=I_\ell(U^\infty), \tag{248}$$

Hence, for all $h \in (0, I_1^*)$ we have

$$\tau_n^* - \hat{\tau}_n = (\tau_n^* - \hat{\tau}_n) \cdot \mathbb{1}_{\{\tau_n^* > R_1(h,U^n)+\hat{\tau}_n+|\mathcal{H}(n)|\}} \tag{269}$$

$$+ \underbrace{(\tau_n^* - \hat{\tau}_n) \cdot \mathbb{1}_{\{\tau_n^* \le R_1(h,U^n)+\hat{\tau}_n+|\mathcal{H}(n)|\}}}_{\le R_1(h,U^n)+|\mathcal{H}(n)|} \tag{270}$$

$$\overset{(268)}{\le} \left[\frac{\gamma^{\mathrm{U}} - \Lambda_{\hat{\tau}_n}}{I_1^* - h} + 1\right] \cdot \mathbb{1}_{\{\tau_n^* > R_1(h,U^n)+\hat{\tau}_n+|\mathcal{H}(n)|\}} \tag{271}$$

$$+ R_1(h, U^n) + |\mathcal{H}(n)| \tag{272}$$

$$\le \frac{\gamma^{\mathrm{U}} - \Lambda_{\hat{\tau}_n}}{I_1^* - h} + R_1(h, U^n) + |\mathcal{H}(n)| + 1. \tag{273}$$

Hence,

$$\tau_n^* - \hat{\tau}_n \le 1 + \inf_{h \in (0, I_1^*)} \frac{\gamma^{\mathrm{U}} - \Lambda_{\hat{\tau}_n}}{I_1^* - h} + R_1(h, U^n) + |\mathcal{H}(n)| \tag{274}$$

$$= 1 + \frac{\gamma^{\mathrm{U}} - \Lambda_{\hat{\tau}_n}}{I_1^*} + R_1(h, U^n) + |\mathcal{H}(n)|. \tag{275}$$

Since the convergence of the nLLR is complete, we can conclude the proof of (54) by combining (11) and (274)–(275) to obtain

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n^* - \hat{\tau}_n\}}{n} \le \frac{\alpha}{I_1^*} - \lim_{n\to\infty} \frac{\mathbb{E}_1\{\Lambda_{\hat{\tau}_n}\}}{nI_1^*} \tag{276}$$

$$+ \lim_{n\to\infty} \frac{\mathbb{E}_1[R_1(h, U^n)]}{n} \tag{277}$$

$$+ \lim_{n\to\infty} \frac{\mathbb{E}_1\{|\mathcal{H}(n)|\}}{n} \tag{278}$$

$$\le \frac{\alpha}{I_1^*} + \lim_{n\to\infty} \frac{\mathbb{E}_1[T_1(h, U^\infty)]}{n} \tag{279}$$

$$+ \lim_{n\to\infty} \frac{\mathbb{E}_1\{|\mathcal{H}(n)|\}}{n} \tag{280}$$

$$= \frac{\alpha}{I_1^*}, \tag{281}$$

where (276) holds since $\mathbb{E}_1\{\Lambda_{\hat{\tau}_n}\}$ is a KL divergence term and it is non-negative, and (279) holds due to the dominated convergence theorem and the fact that by definition $R_1(h, U^n) \le T_1(h, U^\infty), \forall n$, and (281) holds by leveraging Lemma 3 and noting that $\mathbb{E}_1\{T_1(h, U^\infty)\}$ is finite.

Appendix H
Proof of Theorem 5

The error exponents of the NP test are studied in [87], where it is shown that when $\mathsf{P}_{\mathrm{NP}}^0$ is fixed, which is equivalent to an error exponent of 0, the error exponent of P_n^1 is the convergence limit of $\mathsf{nLLR}_0(Y^n; \psi^n)$ as n grows under the assumption that $\{Y_1, \ldots, Y_n\}$ are drawn from distribution f_0. This is equivalent to the definition of I_0. Hence, for the NP test we have $E_{\mathrm{NP}}^1 = I_0$ and $E_{\mathrm{NP}}^0 = 0$. For the sequential sampling setting, based on the analysis of the average delay in Theorems 2 and 3 we have

$$\text{and} \quad \lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n^*\}}{n} = \frac{\alpha}{I_1}. \tag{282}$$

For the error exponent of P_n^0 yielded by Algorithm 1 we have

$$E_n^0 = -\lim_{n\to\infty} \frac{1}{r_1} \ln \mathsf{P}_n^0(r_1) \tag{283}$$

$$\ge -\lim_{n\to\infty} -\frac{n\alpha}{r_1} \tag{284}$$

$$= \lim_{n\to\infty} \frac{n}{\mathbb{E}_1\{\tau_n^*\}} \cdot \alpha \tag{285}$$

$$\overset{(282)}{=} I_1, \tag{286}$$

where (284) follows from Algorithm 1 generating (α, β)-accurate decisions. Next, we define $\Delta \triangleq E_n^0 - I_1 \ge 0$, based on which we have

$$-\ln \mathsf{P}_n^0(r_1) \overset{(283)}{=} r_1 E_n^0 + o(n) = r_1\Delta + r_1 I_1 + o(n). \tag{287}$$

On the other hand, we observe that in the proof of Theorem 2, P_n^0 has been replaced by its upper bound. By keeping P_n^0 throughout the proof it can be readily shown that

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n^*\}}{n} \ge \frac{|\ln \mathsf{P}_n^0(r_1)|}{nI_1}. \tag{288}$$

By combining (287) and (288) we obtain

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n^*\}}{n} \ge \lim_{n\to\infty} \frac{|r_1\Delta + r_1 I_1 + o(n)|}{nI_1} \tag{289}$$

$$= \lim_{n\to\infty} \frac{r_1}{n} \cdot \frac{\Delta + I_1}{I_1} \tag{290}$$

$$\overset{(282)}{=} \frac{\alpha}{I_1} \cdot \frac{\Delta + I_1}{I_1}. \tag{291}$$

On the other hand, from Theorem 3 we have

$$\lim_{n\to\infty} \frac{\mathbb{E}_1\{\tau_n^*\}}{n} \le \frac{\alpha}{I_1}. \tag{292}$$

By comparing (291) and (292) and noting that $\Delta \ge 0$, in the asymptote of large n we should have $\Delta = 0$, and consequently, $E_n^0 = I_1$. The error exponent of P_n^1 can be obtained by following the same line of argument.

Appendix I
Proof of Theorem 6

Without loss of generality assume that at time $t-1$ we have $\delta_{\mathrm{ML}}(t-1) = \mathsf{H}_\ell$. By recalling the definition of $\mathcal{R}_t^i$ given in (19), corresponding to any unobserved node $i \in \varphi^t$ at time t we define $\bar{\mathcal{S}}_t^i \in \mathcal{R}_t^i$ as the *smallest* set of nodes that maximizes the normalized information measure assigned to node $i \in \varphi^t$ at time t, i.e.,

$$\bar{\mathcal{S}}_t^i \triangleq \arg\max_{\mathcal{S} \in \mathcal{R}_t^i} \frac{M_\ell^i(t, \mathcal{S})}{|\mathcal{S}|}. \tag{293}$$

Also, we define

$$u \triangleq \arg\max_i \frac{M_\ell^i(t, \bar{\mathcal{S}}_t^i)}{|\bar{\mathcal{S}}_t^i|}, \tag{294}$$

as the index of the node that exhibits the largest normalized information measure,[7] selected by the selection rule specified in (24). Hence, the optimal sampling path is the set $\bar{\mathcal{S}}_t^u$.

[7]For convenience in notation, we suppressed the dependence of u on t, ℓ, and the past samples.

In order to prove the theorem, we show that the maximum normalized information measure achieved by the set $\bar{\mathcal{S}}_t^u$ is equal to the normalized information measure achieved by only the members of $\bar{\mathcal{S}}_t^i \in \mathcal{R}_t^i$ that are neighbors of u. In other words, by defining

$$\mathcal{T}_t^u \triangleq \bar{\mathcal{S}}_t^u \cap \mathcal{L}_t^u, \tag{295}$$

we show that

$$\frac{M_\ell^u(t, \bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} = \frac{M_\ell^u(t, \mathcal{T}_t^u)}{|\mathcal{T}_t^u|}. \tag{296}$$

We prove this identity by removing the nodes not neighboring u in four steps, and in each step showing that removing those nodes does not penalize $\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}$.

1) Removing any node in $\bar{\mathcal{S}}_t^u$ that belongs to a subgraph of $\mathcal{G}$ different from the subgraph that contains u, does not decrease $\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}$.
2) Furthermore, removing any node of $\bar{\mathcal{S}}_t^u$ whose path to u contains a node that has been observed earlier, does not decrease $\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}$.
3) Moreover, removing any node of $\bar{\mathcal{S}}_t^u$ whose path to u contains an unobserved node that does not belong to $\bar{\mathcal{S}}_t^u$, does not decrease $\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}$.
4) Finally, removing any remaining node that is not a neighbor of u does not decrease $\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}$.

Step 1: First we show that removing the nodes from all subgraphs of $\mathcal{G}$ other than the one that containing node u, does not increase the information measure of node u. For this purpose, we partition $\bar{S}_t^u$ according to

$$\bar{S}_t^u = A \cup \bar{A}, \quad \text{and} \quad A \cap \bar{A} = \phi, \tag{297}$$

where $A \subseteq \bar{S}_t^u$ is the set of nodes that belong to the same subgraph as u, and $\bar{A} \triangleq \bar{S}_t^u \setminus A$. We expand the information measure of u as follows:

$$\frac{M_\ell^u(t, \bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} = \frac{D_{\mathrm{KL}}\big(f_\ell(X_{\bar{A}}|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{A}}|\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|} \tag{298}$$

$$+ \frac{D_{\mathrm{KL}}\big(f_\ell(X_A|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_A|\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|}. \tag{299}$$

We note that A is non-empty since $u \in A$. We show that if $\bar{A}$ is not empty, removing it does not decrease the information measure of node u. Suppose otherwise, i.e., $\bar{A}$ is non-empty and

$$\frac{D_{\mathrm{KL}}\big(f_\ell(X_A|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_A|\mathcal{F}_{t-1})\big)}{|A|} < \frac{M_\ell^u(t, \bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}. \tag{300}$$

Then, in order for (298)–(299) to hold, we must have

$$\frac{D_{\mathrm{KL}}\big(f_\ell(X_{\bar{A}}|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{A}}|\mathcal{F}_{t-1})\big)}{|\bar{A}|} > \frac{M_\ell^u(t, \bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}. \tag{301}$$

Denote one of the members of $\bar{A}$ by v. Then, by noting that $\bar{A} \subseteq \mathcal{S}_t^v$ and invoking the definition of u, we have

$$\frac{D_{\mathrm{KL}}\big(f_\ell(X_{\bar{A}}|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{A}}|\mathcal{F}_{t-1})\big)}{|\bar{A}|} \tag{302}$$

$$\overset{(293)}{\leq} \max_{\mathcal{S} \in \mathcal{R}_t^v} \frac{M_\ell^v(t, \mathcal{S})}{|\mathcal{S}|} \tag{303}$$

$$\overset{(294)}{\leq} \frac{M_\ell^u(t, \bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}, \tag{304}$$

which contradicts (300). Hence, we remove all the nodes that do not belong to the subgraph of $\mathcal{G}$ that contains u, and assume that the optimal set $\bar{\mathcal{S}}_t^u$ is free of such nodes. In the next steps, we focus only on the nodes that belong to the same subgraph that u lies in.

Step 2: Next, we show that further removing the nodes whose path to u contains a node that has been observed earlier, does not increase the information measure of u. For this purpose, we partition $\bar{\mathcal{S}}_t^u$ according to

$$\bar{S}_t^u = B \cup \bar{B}, \quad \text{and} \quad B \cap \bar{B} = \phi, \tag{305}$$

where $B \subseteq \bar{S}_t^u$ is the set of nodes whose paths to u includes an observed node, i.e. an element of ψ_n^{t-1}. According to the global Markov property we have

$$B \perp\!\!\!\perp \bar{B} \;\big|\; \mathcal{F}_{t-1}, \tag{306}$$

Hence, we have the decomposition

$$\frac{M_\ell^u(t, \bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} = \frac{D_{\mathrm{KL}}\big(f_\ell(X_B|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_B|\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|} \tag{307}$$

$$+ \frac{D_{\mathrm{KL}}\big(f_\ell(X_{\bar{B}}|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{B}}|\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|}. \tag{308}$$

We can follow the exact same line of argument as in Step 1, to prove that removing the nodes in B does not decrease the information measure of node u, and consequently, the selected node.

Step 3: In the next step, we show that further removing any node of $\mathcal{S}_t^u$ whose path to u contains an unobserved node that does not belong to $\mathcal{S}_t^u$ can be also removed without penalizing the desired information measure. For this purpose, we partition the set $\bar{\mathcal{S}}_t^u$ according to

$$\bar{S}_t^u = C \cup \bar{C}, \quad \text{and} \quad C \cap \bar{C} = \phi, \tag{309}$$

where $\bar{C}$ is the set of nodes whose paths to u contains at least one node that does not belong to $\bar{\mathcal{S}}_t^u$. Let us also define the set C_j as a subset of $\bar{C}$ whose paths to u contains the unobserved node $j \notin \mathcal{S}_t^u$. Since the graph is acyclic, the sets $\{C_j\}$ are disjoint and partition $\bar{C}$, i.e.

$$\bar{C} = \bigcup_{j \in J} C_j, \quad \text{and} \quad C_j \cap C^{j'} = \phi, \ \forall j, j' \in J, \tag{310}$$

where we have defined J as the smallest set that separates C and $\bar{C}$. Then, we expand the information measure of u

as follows in (311)–(314), shown at the bottom of the page, where we have defined

$$\mathcal{M}_j \triangleq \mathcal{N}_j \cap C. \tag{315}$$

We prove this step by contradiction. Suppose that

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} > \frac{D_{\mathrm{KL}}\big(f_\ell(X_C|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_C|\mathcal{F}_{t-1})\big)}{|C|}. \tag{316}$$

Hence, for (311)–(314) to hold, we should have (317), shown at the bottom of the next page, indicating that there exists at least one $j \in J$ such that (318), shown at the bottom of the next page, holds. Next, by defining $\bar{C}_j \triangleq C_j \cup \{j\}$, we consider the following two different expansions for

$$D_{\mathrm{KL}}\big(f_\ell(X_{\bar{C}_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{C}_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big). \tag{319}$$

Specifically, on one hand we have

$$\begin{aligned} &D_{\mathrm{KL}}\big(f_\ell(X_{\bar{C}_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{C}_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big) && (320)\\ &\quad = D_{\mathrm{KL}}\big(f_\ell(X_j|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_j|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big) && (321)\\ &\quad + D_{\mathrm{KL}}\big(f_\ell(X_{C_j}|X_j,\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_j,\mathcal{F}_{t-1})\big), && (322)\end{aligned}$$

and on the other hand we have (323) and (325), shown at the bottom of the next page. Since the KL divergence is a convex function in both of its arguments and $f_\ell(X_j|X_{\mathcal{M}_j},\mathcal{F}_{t-1})$ is the average of $f_\ell(X_j|X_{\mathcal{M}_j},X_{C_j},\mathcal{F}_{t-1})$, by applying Jensen's inequality we obtain (326), shown at the bottom of the next page. By combining (320)–(326) we get

$$\begin{aligned} &D_{\mathrm{KL}}\big(f_\ell(X_{C_j}|X_j,\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_j,\mathcal{F}_{t-1})\big)\\ &\quad \geq D_{\mathrm{KL}}\big(f_\ell(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big), \end{aligned} \tag{327}$$

which in conjunction with (318) yields (328), shown at the bottom of the next page. This identity, however, contradicts the optimality of u, that is u is the node with the largest information measure.

Step 4: The first three steps, collectively, establish that based on the definition of $\bar{\mathcal{S}}_t^u$ (being the smallest set that maximizes the information measure), the graph formed by the set of nodes in $\bar{\mathcal{S}}_t^u$ is connected and is not separated by any subset of nodes in $V \setminus \bar{\mathcal{S}}_t^u$. This indicates that so far we have shown that $\bar{\mathcal{S}}_t^u$ should contain only neighbors of u or other nodes that are connected to u via a neighbor of u. In the final stage we show cannot contain any node other than the neighbors of u. By contradiction, suppose that $\bar{\mathcal{S}}_t^u$ contains at least one node that is not a neighbor of u. We denote this node by k. By defining

$$\mathcal{S}_t^u \triangleq \bar{\mathcal{S}}_t^u \setminus \{k\}, \tag{329}$$

we have (330), shown at the bottom of the next page, where we have defined

$$\mathcal{M}_k \triangleq \mathcal{N}_k \cap \mathcal{S}_t^u. \tag{331}$$

Since $\bar{\mathcal{S}}_t^u$ maximizes the normalized information content of u, we have

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} > \frac{M_\ell^u(t,\mathcal{S}_t^u)}{|\mathcal{S}_t^u|}, \tag{332}$$

and, consequently, in order for (330) to hold we should have

$$\begin{aligned} &D_{\mathrm{KL}}\big(f_\ell(X_k|X_{\mathcal{M}_k},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_k|X_{\mathcal{M}_k},\mathcal{F}_{t-1})\big)\\ &\qquad > \frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|}. \end{aligned} \tag{333}$$

On the other hand, we have

$$\begin{aligned} &D_{\mathrm{KL}}\big(f_\ell(X_k|X_{\mathcal{M}_k},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_k|X_{\mathcal{M}_k},\mathcal{F}_{t-1})\big)\\ &\qquad < \frac{M_\ell^k(t,\bar{\mathcal{S}}_t^k)}{|\bar{\mathcal{S}}_t^k|}, \end{aligned} \tag{334}$$

which combined with (333) indicates

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} < \frac{M_\ell^k(t,\bar{\mathcal{S}}_t^k)}{|\bar{\mathcal{S}}_t^k|}. \tag{335}$$

This contradicts the optimality of u, and as a result $\bar{\mathcal{S}}_t^u$ cannot contain any node that is not a neighbor of u. This completes the proof.

APPENDIX J
PROOF OF THEOREM 8

For a GMRF with an underlying line dependency graph, when $\Sigma_{ij} = \sigma$ among the neighboring nodes, we have a homogeneous networks in which

$$I_0 = \ln(1-\sigma^2) + \frac{2\sigma^2}{1-\sigma^2}, \tag{336}$$

$$\text{and} \qquad I_1 = \ln\frac{1}{1-\sigma^2}. \tag{337}$$

$$\begin{aligned} \frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} &= \frac{D_{\mathrm{KL}}\big(f_\ell(X_C|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_C|\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|} && (311)\\ &\quad + \sum_{j\in J} \frac{D_{\mathrm{KL}}\big(f_\ell(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|} && (312)\\ &\leq \max\Bigg\{ \frac{D_{\mathrm{KL}}\big(f_\ell(X_C|\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_C|\mathcal{F}_{t-1})\big)}{|C|}, && (313)\\ &\qquad \max_{j\in J} \frac{D_{\mathrm{KL}}\big(f_\ell(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big)}{|C_j|} \Bigg\}, && (314)\end{aligned}$$

By applying these identities to sets A and B in which $\sigma > a$ and $\sigma < b$, respectively, and noting that I_0 and I_1 are monotonically increasing functions of $|\sigma|$ we have

$$\frac{I_0(A)}{I_0(B)} \tag{338}$$
$$\geq \frac{\ln(1-a^2)+\frac{2a^2}{1-a^2}}{\ln(1-b^2)+\frac{2b^2}{1-b^2}} \tag{339}$$
$$= \frac{-a^2-\frac{a^4}{2}-\frac{a^6}{3}-o(a^6)+2a^2\big(1+a^2+a^4+o(a^4)\big)}{-b^2-\frac{b^4}{2}-\frac{b^6}{3}-o(b^6)+2b^2\big(1+b^2+b^4+o(b^4)\big)} \tag{340}$$
$$= \frac{a^2+\frac{3}{2}a^4+\frac{5}{6}a^6+o(a^6)}{b^2+\frac{3}{2}b^4+\frac{5}{6}b^6+o(b^6)} \tag{341}$$
$$\geq \frac{a^2}{b^2}, \tag{342}$$

where the last inequality holds since $a > b$. Similarly, for the expected delays under H_1 we have

$$\frac{I_1(A)}{I_1(B)} \geq \frac{-\ln(1-a^2)}{-\ln(1-b^2)} \tag{343}$$
$$= \frac{a^2+\frac{a^4}{2}+\frac{a^6}{3}o(a^6)}{b^2+\frac{b^4}{2}+\frac{b^6}{3}+o(b^6)} \tag{344}$$
$$= \frac{a^2(1+\frac{1}{2}a^2+\frac{1}{3}a^4+o(a^4)}{b^2(1+\frac{1}{2}b^2+\frac{1}{3}b^4+o(b^4)} \tag{345}$$
$$\geq \frac{a^2}{b^2}. \tag{346}$$

When $|A| = o(n)$, Chernoff's rule starts the sampling process from set B with probability 1 and since the graph is connected stays in set B until it exhaust all its nodes. By invoking the results of Theorem 3, we can conclude that the expected delay of Chernoff's rule under H_ℓ is inversely proportional to $I_\ell(B)$. Furthermore, from Corollary 2 and Theorem 4 the expected delay of our strategy under H_ℓ is inversely proportional to $I_\ell(A)$, which concludes the proof

Appendix K
Proof of Theorem 9

We define $\tau_d \triangleq \tau_{\rm c} - \tau_n^*$. The optimal sampling strategy starts by directly sampling from set A. For Chernoff's rule, however, there is a chance that it starts sampling from B before entering A. We define $\tau_{\rm c}^A$ and $\tau_{\rm c}^B$ as the number of samples that Chernoff's rule spends on sets A and B, respectively. We show that

$$\mathbb{E}_\ell\{\tau_{\rm c}^A\} \geq \mathbb{E}_\ell\{\tau_n^*\}, \quad \text{and} \quad \mathbb{E}_\ell\{\tau_{\rm c}^B\} = \Theta\left(\frac{n}{p}\right), \tag{347}$$

which indicates the desires result, i.e.,

$$0 \leq \mathbb{E}_\ell\{\tau_d\} = \mathbb{E}_\ell\{\tau_{\rm c}^A\} + \mathbb{E}_\ell\{\tau_{\rm c}^B\} - \mathbb{E}_\ell\{\tau_n^*\} \ \geq \Theta\left(\frac{n}{p}\right). \tag{348}$$

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} < \max_{j\in J} \frac{D_{\rm KL}\big(f_\ell(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big)}{|C_j|}, \tag{317}$$

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} < \frac{D_{\rm KL}\big(f_\ell(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big)}{|C_j|}. \tag{318}$$

$$D_{\rm KL}\big(f_\ell(X_{\bar{C}_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{\bar{C}_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big) \tag{323}$$
$$= D_{\rm KL}\big(f_\ell(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big) \tag{324}$$
$$+ D_{\rm KL}\big(f_\ell(X_j|X_{\mathcal{M}_j},X_{C_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_j|X_{\mathcal{M}_j},X_{C_j},\mathcal{F}_{t-1})\big). \tag{325}$$

$$D_{\rm KL}\big(f_\ell(X_j|X_{\mathcal{M}_j},X_{C_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_j|X_{\mathcal{M}_j},X_{C_j},\mathcal{F}_{t-1})\big) \geq D_{\rm KL}\big(f_\ell(X_j|X_{\mathcal{M}_j},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_j|X_{\mathcal{M}_j},\mathcal{F}_{t-1})\big). \tag{326}$$

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} < \frac{D_{\rm KL}\big(f_\ell(X_{C_j}|X_j,\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_{C_j}|X_j,\mathcal{F}_{t-1})\big)}{|C_j|} \leq \frac{M_\ell^u(t,C_j)}{|C_j|}. \tag{328}$$

$$\frac{M_\ell^u(t,\bar{\mathcal{S}}_t^u)}{|\bar{\mathcal{S}}_t^u|} = \frac{M_\ell^u(t,\mathcal{S}_t^u)}{|\bar{\mathcal{S}}_t^u|} + \frac{D_{\rm KL}\big(f_\ell(X_k|X_{\mathcal{M}_k},\mathcal{F}_{t-1}) \parallel f_{1-\ell}(X_k|X_{\mathcal{M}_k},\mathcal{F}_{t-1})\big)}{|\bar{\mathcal{S}}_t^u|}, \tag{330}$$

The first identity in (347) follows the optimality of τ_n^*. Specifically, the optimal rule starts by sampling from A and stays inside A until the stopping time τ_n^*. On the other hand, Chernoff's rule might start from sampling B, but once it enters A it remains there until it takes τ_{c}^A samples. By noting the optimality of τ_n^*, we immediately have the first identity in (347). In order to establish the second identity in (347), we provide lower and upper bounds on the asymptotic value of $\mathbb{E}_\ell\{\tau_{\mathrm{c}}^B\}$. By definition, any sampling rule can take at most $(n-p)$ samples from set B. Hence, we obtain an upper bound as follows:

$$\mathbb{E}_\ell\{\tau_{\mathrm{c}}^B\} = \sum_{k=0}^{n-p} k \cdot \mathbb{P}_\ell(\tau_{\mathrm{c}}^B = k) \tag{349}$$

$$= \sum_{k=0}^{n-p} k \cdot \frac{\binom{n-p}{k}}{\binom{n}{k}} \cdot \frac{p}{n-k} \tag{350}$$

$$= \sum_{k=1}^{n-p} k \cdot \frac{p}{n} \cdot \frac{(n-p)!}{(n-p-k)!} \cdot \frac{(n-k-1)!}{(n-1)!} \tag{351}$$

$$= \sum_{k=1}^{n-p} k \cdot \frac{p}{n} \prod_{i=0}^{k-1} \underbrace{\frac{n-p-i}{n-1-i}}_{\leq \frac{n-p}{n-1}} \tag{352}$$

$$\leq \frac{p}{n} \sum_{k=1}^{n-p} k \cdot \left(1 - \frac{p-1}{n-1}\right)^k. \tag{353}$$

Hence, by noting that $p = o(n)$ we obtain

$$\lim_{n\to\infty} \frac{\mathbb{E}_\ell\{\tau_{\mathrm{c}}^B\}}{\frac{n}{p}} \leq \lim_{n\to\infty} \left(\frac{p}{n}\right)^2 \sum_{k=1}^{n-p} k \cdot \left(1 - \frac{p-1}{n-1}\right)^k \tag{354}$$

$$= \lim_{n\to\infty} \left(\frac{p}{n}\right)^2 \left(\frac{n-1}{p-1}\right)^2 = 1. \tag{355}$$

For the lower bound, from (352) we have

$$\mathbb{E}\{\tau_d\} = \sum_{k=1}^{n-p} k \cdot \frac{p}{n} \prod_{i=0}^{k-1} \frac{n-p-i}{n-1-i} \tag{356}$$

$$\geq \sum_{k=1}^{\lfloor \frac{n-p}{2} \rfloor} k \cdot \frac{p}{n} \prod_{i=0}^{k-1} \frac{n-p-i}{n-1-i} \tag{357}$$

$$\geq \sum_{k=1}^{\lfloor \frac{n-p}{2} \rfloor} k \cdot \frac{p}{n} \prod_{i=0}^{k-1} \frac{n-p-\lfloor \frac{n-p}{2} \rfloor}{n-1-\lfloor \frac{n-p}{2} \rfloor} \tag{358}$$

$$\geq \sum_{k=1}^{\lfloor \frac{n-p}{2} \rfloor} k \cdot \frac{p}{n} \prod_{i=0}^{k-1} \frac{\frac{n-p}{2}}{\frac{n+p}{2}} \tag{359}$$

$$= \frac{p}{n} \sum_{k=1}^{\lfloor \frac{n-p}{2} \rfloor} k \left(\frac{n-p}{n+p}\right)^k \tag{360}$$

$$= \frac{p}{n} \sum_{k=1}^{\lfloor \frac{n-p}{2} \rfloor} k \left(1 - \frac{2p}{n+p}\right)^k. \tag{361}$$

Hence, by noting that $p = o(n)$ we obtain

$$\lim_{n\to\infty} \frac{\mathbb{E}_\ell\{\tau_{\mathrm{c}}^B\}}{\frac{n}{p}} \geq \lim_{n\to\infty} \left(\frac{p}{n}\right)^2 \sum_{k=1}^{\lfloor \frac{n-p}{2} \rfloor} k \left(1 - \frac{2p}{n+p}\right)^k \tag{362}$$

$$= \lim_{n\to\infty} \left(\frac{p}{n}\right)^2 \left(\frac{n+p}{2p}\right)^2 \tag{363}$$

$$= \frac{1}{4}. \tag{364}$$

Hence, from (354) and (364) we have

$$\mathbb{E}_\ell\{\tau_{\mathrm{c}}^B\} = \Theta\left(\frac{n}{p}\right), \tag{365}$$

which completes the proof.

Appendix L
Derivations of $D_\ell^i(t)$ and $\mathscr{J}_\ell(\{i\}, \psi^{t-1})$

By leveraging (97) we have

$$D_0^i(t) = -\mathbb{E}_0\left\{ \sum_{j\in\mathcal{N}_i^t} \mathsf{LLR}(i,j) \;\middle|\; \mathcal{F}_{t-1} \right\}, \tag{366}$$

where by replacing $\mathsf{LLR}(i,j)$ from (96) and noting that Chernoff's rule first observes the neighbors of already-observed nodes due to its information measure structure, we have the derivation in (367)–(370), shown at the bottom of the page.

Derivation of $D_1^i(t)$ follows the same line of argument, as shown in (371)–(374), at the top of the next page.

$$D_0^i(t) = -\frac{1}{2} \sum_{j\in\mathcal{N}_i^t} \mathbb{E}_0\left\{ \ln \frac{1}{1-\Sigma_{ij}^2} - \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\left(X_i^2 + X_j^2\right) + \frac{2\Sigma_{ij}}{1-\Sigma_{ij}^2} X_i X_j \;\middle|\; \mathcal{F}_{t-1} \right\} \tag{367}$$

$$= -\frac{1}{2} \sum_{j\in\mathcal{N}_i^t} \left[\ln \frac{1}{1-\Sigma_{ij}^2} - \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\Big(\underbrace{\mathbb{E}_0\{X_i^2 \mid \mathcal{F}_{t-1}\}}_{=\mathbb{E}_0\{X_i^2\}=1} + X_j^2\Big) + \frac{2\Sigma_{ij}}{1-\Sigma_{ij}^2} X_j \underbrace{\mathbb{E}_0\{X_i \mid \mathcal{F}_{t-1}\}}_{=\mathbb{E}_0\{X_i\}=0} \right] \tag{368}$$

$$= -\frac{1}{2} \sum_{j\in\mathcal{N}_i^t} \left[\ln \frac{1}{1-\Sigma_{ij}^2} - \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\left(X_j^2 + 1\right) \right] \tag{369}$$

$$= \frac{1}{2} \sum_{j\in\mathcal{N}_i^t} \left[\ln(1-\Sigma_{ij}^2) + \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\left(X_j^2 + 1\right) \right]. \tag{370}$$

$$D_1^i(t) = \frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\mathbb{E}_1\left\{\ln\frac{1}{1-\Sigma_{ij}^2} - \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\left(X_i^2+X_j^2\right) + \frac{2\Sigma_{ij}}{1-\Sigma_{ij}^2}X_iX_j \;\middle|\; \mathcal{F}_{t-1}\right\} \tag{371}$$

$$= \frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\left[\ln\frac{1}{1-\Sigma_{ij}^2} - \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\Big(\underbrace{\mathbb{E}_1\{X_i^2\mid\mathcal{F}_{t-1}\}}_{=\mathbb{E}_1\{X_i^2|X_j\}=1+\Sigma_{ij}^2(X_j^2-1)} + X_j^2\Big) + \frac{2\Sigma_{ij}}{1-\Sigma_{ij}^2}X_j\ \underbrace{\mathbb{E}_1\{X_i\mid\mathcal{F}_{t-1}\}}_{=\mathbb{E}_1\{X_i|X_j\}=\Sigma_{ij}X_j}\right] \tag{372}$$

$$= \frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\left[\ln\frac{1}{1-\Sigma_{ij}^2} - \frac{\Sigma_{ij}^2}{1-\Sigma_{ij}^2}\left(1+\Sigma_{ij}^2(X_j^2-1)+X_j^2\right) + \frac{2\Sigma_{ij}^2}{1-\Sigma_{ij}^2}X_j^2\right] \tag{373}$$

$$= \frac{1}{2}\sum_{j\in\mathcal{N}_i^t}\left[\ln\frac{1}{1-\Sigma_{ij}^2} + \Sigma_{ij}^2\left(X_j^2-1\right)\right]. \tag{374}$$

In order to derive $\mathscr{J}_0(\{i\},\psi^{t-1})$, besides the terms in $D_0^i(t)$, we must account for the nodes that have been observed prior to sampling node i and are neighbors in $\mathcal{G}(V_{t-1},E_{t-1})$ but non-neighbors in $\mathcal{G}(V_t,E_t)$, similar to what occurs in the toy example in Fig. 2 when transitioning from $t=2$ to $t=3$, i.e., some edges in $\mathcal{G}(V_t,E_t)$ for $t=2$ are removed at a later time $t=3$. Therefore, the terms associated with those edges that are removed at time t by sampling node i should be subtracted from the information measure, i.e.,

$$\mathscr{J}_0(\{i\},\psi^{t-1}) = D_0^i(t) - \sum_{(j,k)\in\Delta_t^i} -\mathsf{LLR}(j,k) \tag{375}$$

$$= D_0^i(t) + \sum_{(j,k)\in\Delta_t^i}\mathsf{LLR}(j,k), \tag{376}$$

where, according to the definition of Δ_t^i in (100), the second term removes the edges (j,k) that have been removed from the graph at time t. We can find $\mathscr{J}_1(\{i\},\psi^{t-1})$ similarly.

REFERENCES

[1] S. G. Brush, "History of the Lenz-Ising model," *Rev. Mod. Phys.*, vol. 39, no. 4, pp. 883–893, Oct. 1967.

[2] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques-Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

[3] G. R. Cross and A. K. Jain, "Markov random field texture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-5, no. 1, pp. 25–39, Jan. 1983.

[4] D. Gleich, "Markov random field models for non-quadratic regularization of complex SAR images," *IEEE J. Sel. Topics Appl. Earth Observat., Remote Sens.*, vol. 5, no. 3, pp. 952–961, Jun. 2012.

[5] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, Jul. 2008.

[6] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Statist.*, vol. 38, no. 3, pp. 1287–1319, Jun. 2010.

[7] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, Jun. 2010, pp. 1373–1377.

[8] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, Jul. 2012.

[9] T. Cai and W. Liu, "Adaptive thresholding for sparse covariance matrix estimation," *J. Amer. Statist. Assoc.*, vol. 106, no. 494, pp. 672–684, 2011.

[10] T. T. Cai, W. Liu, and X. Luo, "A constrained ℓ_1 minimization approach to sparse precision matrix estimation," *J. Amer. Stat. Assoc.*, vol. 106, no. 494, pp. 594–607, 2011.

[11] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2293–2337, Aug. 2012.

[12] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes," *J. Roy. Statist. Soc. B, Stat. Methodol.*, vol. 76, no. 2, pp. 373–397, Mar. 2014.

[13] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 129–145, Mar. 1996.

[14] S. Tong and D. Koller, "Active learning for parameter estimation in Bayesian networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2001, pp. 647–653.

[15] S. Tong and D. Koller, "Active learning for structure in Bayesian networks," in *Proc. Int. Joint Conf. Artif. Intell.*, Seattle, WA, USA, no. 1, Aug. 2001, pp. 863–869.

[16] Y.-B. He and Z. Geng, "Active learning of causal networks with intervention experiments and optimal designs," *J. Mach. Learn. Res.*, vol. 9, pp. 2523–2547, Nov. 2008.

[17] D. Vats, R. Nowak, and R. Baraniuk, "Active learning for undirected graphical model selection," in *Proc. Int. Conf. Artif. Intell. Statist.*, Reykjavík, Iceland, Apr. 2014, pp. 958–967.

[18] G. Dasarathy, A. Singh, M.-F. Balcan, and J. H. Park, "Active learning algorithms for graphical model selection," in *Proc. Int. Conf. Artif. Intell. Statist.*, Cadiz, Spain, May 2016, pp. 1356–1364.

[19] D. M. Chickering, "Learning Bayesian networks is NP-complete," in *Learning From Data*. New York, NY, USA: Springer, pp. 121–130, 1996.

[20] C. J. Ku and T. L. Fine, "A Bayesian independence test for small datasets," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 4026–4031, Oct. 2006.

[21] A. Anandkumar, L. Tong, and A. Swami, "Detection of Gauss–Markov random fields with nearest-neighbor dependency," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 816–827, Feb. 2009.

[22] V. Solo and A. Pasha, "Testing for independence between a point process and an analog signal," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 3762–3765.

[23] E. Arias-Castro, S. Bubeck, and G. Lugosi, "Detection of correlations," *Ann. Statist.*, vol. 40, no. 1, pp. 412–435, Feb. 2012.

[24] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of geometric structures over networks," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5034–5046, Oct. 2017.

[25] E. Arias-Castro, S. Bubeck, and G. Lugosi, "Detecting positive correlations in a multivariate sample," *Bernoulli*, vol. 21, no. 1, pp. 209–241, Feb. 2015.

[26] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *Ann. Statist.*, vol. 41, no. 4, pp. 1780–1815, 2013.

[27] A. Hero and B. Rajaratnam, "Hub discovery in partial correlation graphs," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 6064–6078, Sep. 2012.

[28] Y. Xia and L. Li, "Hypothesis testing of matrix graph model with application to brain connectivity analysis," *Biometrics*, vol. 73, no. 3, pp. 780–791, Sep. 2017.

[29] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, 1945.

[30] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, U.K.: Cambridge Univ. Press, Dec. 2009.

[31] V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli, "Multihypothesis sequential probability ratio tests—Part I: Asymptotic optimality," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2448–2461, Nov. 1999.

[32] A. G. Tartakovsky, "Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case," *Stat. Inference Stochastic Process.*, vol. 1, no. 3, pp. 265–295, Oct. 1998.

[33] A. O. Hero and D. Cochran, "Sensor management: Past, present, and future," *IEEE Sensors J.*, vol. 11, no. 12, pp. 3064–3075, Dec. 2011.

[34] G. Hollinger, B. Englot, F. Hover, U. Mitra, and G. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity," *Int. J. Robot. Res.*, vol. 32, no. 1, pp. 3–18, Nov. 2013.

[35] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Müller, *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL, USA: CRC Press, 2011.

[36] P. Shenoy and A. J. Yu, "Rational decision-making in inhibitory control," *Frontiers Hum. Neurosci.*, vol. 5, pp. 224–236, May 2011.

[37] R. D. Nowak, "The geometry of generalized binary search," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7893–7906, Dec. 2011.

[38] M. V. Burnashev, "Data transmission over a discrete channel with feedback. Random transmission time," *Problemy Inf. Inform.*, vol. 12, no. 4, pp. 10–30, 1976.

[39] J. Heydari, A. Tajer, and H. V. Poor, "Quickest detection of Gauss–Markov random fields," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Sep. 2015, pp. 808–814.

[40] J. Heydari, A. Tajer, and H. V. Poor, "Quickest detection of Markov networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1341–1345.

[41] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Ann. Statist.*, vol. 29, no. 2, pp. 295–327, 2001.

[42] A. Soshnikov, "A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices," *J. Statist. Phys.*, vol. 108, no. 5, pp. 1033–1056, 2002.

[43] S. Péché, "Universality results for the largest eigenvalues of some sample covariance matrix ensembles," *Probab. Theory Rel. Fields*, vol. 143, nos. 3–4, pp. 481–516, 2009.

[44] O. Ledoit and M. Wolf, "Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size," *Ann. Statist.*, vol. 30, no. 4, pp. 1081–1102, 2002.

[45] T. Jiang, "The asymptotic distributions of the largest entries of sample correlation matrices," *Ann. Appl. Probab.*, vol. 14, no. 2, pp. 865–880, 2004.

[46] T. T. Cai and T. Jiang, "Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices," *Ann. Statist.*, vol. 39, no. 3, pp. 1496–1525, 2011.

[47] B. Shao and W. Zhou, "Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices," *Ann. Probab.*, vol. 42, no. 2, pp. 623–648, 2014.

[48] N. Sugiura and H. Nagao, "Unbiasedness of some test criteria for the equality of one or two covariance matrices," *Ann. Math. Statist.*, vol. 39, no. 5, pp. 1686–1692, Oct. 1968.

[49] S. D. Gupta and N. Giri, "Properties of tests concerning covariance matrices of normal distributions," *Ann. Statist.*, vol. 1, no. 6, pp. 1222–1224, Nov. 1973.

[50] M. D. Perlman, "Unbiasedness of the likelihood ratio tests for equality of several covariance matrices and equality of several multivariate normal populations," *Ann. Statist.*, vol. 8, no. 2, pp. 247–263, Mar. 1980.

[51] A. K. Gupta and J. Tang, "Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models," *Biometrika*, vol. 71, no. 3, pp. 555–559, 1984.

[52] Y. Qiu and S. X. Chen, "Test for bandedness of high-dimensional covariance matrices and bandwidth estimation," *Ann. Statist.*, vol. 40, no. 3, pp. 1285–1314, 2012.

[53] T. T. Cai, "Global testing and large-scale multiple testing for high-dimensional covariance structures," *Annu. Rev. Statist. Appl.*, vol. 4, no. 1, pp. 423–446, Mar. 2017.

[54] T. T. Cai and A. Zhang, "Inference for high-dimensional differential correlation matrices," *J. Multivariate Anal.*, vol. 143, pp. 107–126, Jan. 2016.

[55] T. Cai, W. Liu, and Y. Xia, "Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings," *J. Amer. Stat. Assoc.*, vol. 108, no. 501, pp. 265–277, Mar. 2013.

[56] Y. Xia, T. Cai, and T. T. Cai, "Testing differential networks with applications to the detection of gene-gene interactions," *Biometrika*, vol. 102, no. 2, pp. 247–266, Jun. 2015.

[57] H. Chernoff, "Sequential design of experiments," *Ann. Math. Statist.*, vol. 30, no. 3, pp. 755–770, 1959.

[58] S. Bessler, "Theory and applications of the sequential design of experiments, *k*-actions and infinitely many experiments—Part I theory," Dept. Statist., Stanford Univ., Stanford, CA, USA, Tech. Rep., Mar. 1960.

[59] A. E. Albert, "The sequential design of experiments for infinitely many states of nature," *Ann. Math. Statist.*, vol. 32, no. 3, pp. 774–799, Sep. 1961.

[60] G. E. P. Box and W. J. Hill, "Discrimination among mechanistic models," *Technometrics*, vol. 9, no. 1, pp. 57–71, Feb. 1967.

[61] W. J. Blot and D. A. Meeter, "Sequential experimental design procedures," *J. Amer. Stat. Assoc.*, vol. 68, no. 343, pp. 586–593, Sep. 1973.

[62] S. Nitinawarat, G. K. Atia, and V. V. Veeravalli, "Controlled sensing for multihypothesis testing," *IEEE Trans. Autom. Control*, vol. 58, no. 10, pp. 2451–2464, Oct. 2013.

[63] S. Nitinawarat and V. V. Veeravalli, "Controlled sensing for sequential multihypothesis testing with controlled Markovian observations and non-uniform control cost," *Sequential Anal.*, vol. 34, no. 1, pp. 1–24, Feb. 2015.

[64] K. Cohen and Q. Zhao, "Active hypothesis testing for anomaly detection," *IEEE Trans. Inf. Theory*, vol. 61, no. 3, pp. 1432–1450, Mar. 2015.

[65] N. K. Vaidhiyan and R. Sundaresan, "Active search with a cost for switching actions," 2015, *arXiv:1505.02358*.

[66] R. M. Castro and E. Tánczos, "Adaptive sensing for estimation of structured sparse signals," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2060–2080, Apr. 2015.

[67] G. Atia and S. Aeron, "Controlled sensing for sequential estimation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Austin, TX, USA, Dec. 2013, pp. 125–128.

[68] D. S. Zois, M. Levorato, and U. Mitra, "Active classification for POMDPs: A Kalman-like state estimator," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6209–6224, Dec. 2014.

[69] J. G. Ligo, G. K. Atia, and V. V. Veeravalli, "A controlled sensing approach to graph classification," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6468–6480, Dec. 2014.

[70] G. Schwarz, "Asymptotic shapes of Bayes sequential testing regions," *Ann. Math. Statist.*, vol. 33, no. 1, pp. 224–236, Mar. 1962.

[71] J. Kiefer and J. Sacks, "Asymptotically optimum sequential inference and design," *Ann. Math. Statist.*, vol. 34, no. 3, pp. 705–750, Sep. 1963.

[72] S. P. Lalley and G. Lorden, "A control problem arising in the sequential design of experiments," *Ann. Probab.*, vol. 14, no. 1, pp. 136–172, Jan. 1986.

[73] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *Ann. Statist.*, vol. 41, no. 6, pp. 2703–2738, 2013.

[74] M. Naghshvar and T. Javidi, "Sequentiality and adaptivity gains in active hypothesis testing," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 768–782, Oct. 2013.

[75] J. Wang, "Bayes-optimal sequential multi-class diagnosis," 2015, *arXiv:1506.08915*.

[76] K. Cohen, Q. Zhao, and A. Swami, "Optimal index policies for anomaly localization in resource-constrained cyber systems," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4224–4236, Aug. 2014.

[77] K. Cohen and Q. Zhao, "Asymptotically optimal anomaly detection via sequential testing," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2929–2941, Jun. 2015.

[78] K. P. Murphy, "Active learning of causal Bayes net structure," Tech. Rep., 2001.

[79] D. Siegmund, *Sequential Analysis: Tests Confidence Intervals*. New York, NY, USA: Springer-Verlag, 1985.

[80] P. L. Hsu and H. Robbins, "Complete convergence and the law of large numbers," *Proc. Nat. Acad. Sci. USA*, vol. 33, no. 2, pp. 25–31, Feb. 1947.

[81] A. F. Karr, *Probability*. New York, NY, USA: Springer, 1993.

[82] A. G. Tartakovsky, "On asymptotic optimality in sequential changepoint detection: Non-iid case," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3433–3450, Jun. 2017.

[83] T. L. Lai, "Asymptotic optimality of invariant sequential probability ratio tests," *Ann. Statist.*, vol. 9, no. 2, pp. 318–333, Mar. 1981.

[84] M. Granovetter, "The strength of weak ties," *Amer. J. Sociol.*, vol. 78, no. 6, pp. 1360–1380, 1973.

[85] L. Spencer and R. Pahl, *Rethinking Friendship: Hidden Solidarities Today*. Princeton, NJ, USA: Princeton Univ. Press, 2006.
[86] R. Dunbar, *How Many Friends Does One Person Need? Dunbar's Number and Other Evolutionary Quirks*. Fabriano, Italy: Faber, 2010.
[87] P.-N. Chen, "General formulas for the Neyman–Pearson type-II error exponent subject to fixed and exponential type-I error bounds," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 316–323, Jan. 1996.

Ali Tajer (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology in 2002 and 2004, respectively, and the M.A. degree in statistics and the Ph.D. degree in electrical engineering from Columbia University. From 2007 to 2010, he was with Columbia University. From 2010 to 2012, he was with Princeton University as a Postdoctoral Research Associate. He is currently an Associate Professor in electrical, computer, and systems engineering at Rensselaer Polytechnic Institute. His recent publications include an edited book entitled *Advanced Data Analytics for Power Systems* (Cambridge University Press, 2021). His research interests include mathematical statistics, statistical signal processing, and network information theory, with applications in wireless communications, and power grids. He has received an NSF CAREER Award in 2016 and AFRL Faculty Fellowship in 2019. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. In the past, he has served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, a Guest Editor for the *IEEE Signal Processing Magazine*, an Editor for the IEEE TRANSACTIONS ON SMART GRID, and the Guest Editor-in-Chief for the Special Issue on Theory of Complex Systems with Applications to Smart Grid Operations of the IEEE TRANSACTIONS ON SMART GRID.

Javad Heydari (Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology and the M.Sc. degree in mathematics and the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, in 2018. From 2010 to 2012, he was with the Advanced Communication Research Institute, Sharif University of Technology, where he worked on the applications of signal processing in communication systems. He is currently a Senior AI Research Scientist at LG Electronics, Santa Clara, CA, USA, where he is focusing on providing machine learning-based solutions to product-driven problems and conducting fundamental research in the broader area of machine learning.

H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from Princeton University in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana–Champaign. Since 1990, he has been on the faculty at Princeton, where he is currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean of Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other institutions, most recently at Berkeley and Cambridge. His research interests include the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems, and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications* (Cambridge University Press).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and also a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.