Quickest Search for Local Structures in Random Graphs

Javad Heydari, *Student Member, IEEE*, and Ali Tajer, *Senior Member, IEEE*

***Abstract*—A network of agents that form a random graph is considered. Each agent represents an information source that generates a sequence of random variables (RVs). The RVs generated by an unknown subset of nodes are correlated according to a known kernel, while the remaining nodes generate independent and identically distributed random variables. To identify and localize the desired unknown subset of correlated nodes, this paper formalizes and delineates a quickest search process, which is the strategy that minimizes the average number of measurements. Despite its widespread applications, the problem of identifying subgraphs with such desired correlation structures is often investigated under the *fixed sample-size* settings, in which the data acquisition process and the inferential mechanisms are decoupled. Motivated by the significant advantages of sequential methods for agile inference, this paper analyzes this problem under a *fully sequential* setting. Specifically, it offers a framework that unifies the intertwined processes of information gathering and decision making, and through a constructive proof, it provides an optimal sequential data-gathering process as well as the attendant decision rules for the quickest search of interest.**

***Index Terms*—Correlation, dynamic decisions, local structure, quickest search, random graph.**

I. Introduction

THE proliferation of networked data and the need for extracting refined information from these networks necessitate detecting small structures in large networks and analyzing their local properties. Specifically, while being a constituent module in a very large network, each information source may also belong to some sub-networks. For instance, each subscriber in a social network may be a member of some local communities, or a group of co-located sensors dispersed in a wide surveillance system generate locally correlated data in the presence of abnormal activities. In general, the data generated by such subnetworks bears certain correlation structures, which are induced by the underlying coupling or interaction among their constituents. When such subnetworks exist, their structures can be exploited to extract more precise information from the network. Hence, detecting and localizing such subnetworks, which can be equivalently abstracted as detecting and localizing *local* correlation structures in random graphs, are of paramount significance.

The problem of detecting a *local* correlation structure in given networked data arises in many application domains, such as wireless sensor networks [1], distributed attack detection in computer networks [2], anomaly detection with correlated anomalous sources [3] and [4], health monitoring [5], social sciences [6], and anomaly detection [7]. In order to formalize local correlation, in this paper we consider a large graph in which each node generates a sequence of random variables. The stream of random variables generated by a specific *subset* of nodes form a correlation structure according to a known kernel, while all the remaining nodes generate independent random variables. In such a model, the nodes generating independent random variables are isolated nodes, while the ones that belong to the local structure form a connected graph in which the edges of the graph capture the dependency structure of the generated data, i.e., two nodes have one edge connecting them if their generated random variables are dependent given the rest of the random variables. It is assumed that the correlation structure and its size are known, and the objective is to identify a subset of the correlated nodes. The recent studies in detecting and localizing local correlation structures that are relevant to the scope of this paper can be, broadly, categorized into two general directions.

In one direction, the data-gathering and decision-making processes are decoupled, and a decision about the locality of the local correlation structure is formed after the data is collected in its entirety and based on a pre-specified strategy [8]–[19]. Specifically, the work in [8] considers detecting a local Markov random field in a large graph, which is motivated by localizing textured objects in images. The studies in [9] and [10] focus on a graph of Gaussian random variables such that it is populated by independent random variables and the locally correlated random variables appear only sparsely. These studies aim at identifying such sparse locally correlated Gaussian random variables under different sampling constraints. The work in [11] applies principal component analysis to localize the coupled nodes, and [12] models the network as a graph and applies the partial correlation screening algorithms to identify strongly correlated nodes. The studies in [13]–[17] investigate detection and localization of geometric structures in lattice graphs such as one dimensional grids, while [18] and [19] study the existence of local clusters in graph structures. These approaches, irrespective of their discrepancies, are designed as *universally* efficient ones, the cost of which is that they are not data-adaptive and lack efficiency

Manuscript received July 12, 2017; accepted July 19, 2017. Date of publication July 24, 2017; date of current version August 14, 2017. This work was supported in part by the U.S. National Science Foundation under Grant ECCS-1455228 and the CAREER Award ECCS-1554482. This paper was presented in part at the 2016 Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, Sep. 27–30, 2017. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Michael Rabbat. *(Corresponding author: Ali Tajer.)*

The authors are with the Electrical, Computer, and Systems Engineering Department, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: heydas@rpi.edu; tajera@rpi.edu).

Digital Object Identifier 10.1109/TSIPN.2017.2731125

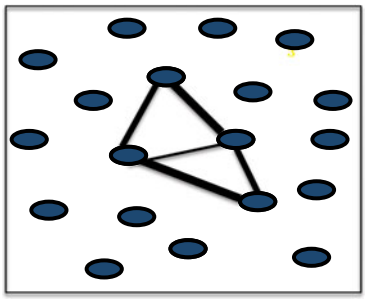

Fig. 1. A graph with a local correlation structure.

in adapting the data-gathering process to the actual structure of the data, especially when the graph (data dimension) grows.

Driven by the advantages of sequential sampling [20] and controlled sensing [21]–[26], in this paper we formulate the problem of searching for local correlation structures in a *fully* sequential setting, in which the data samples are collected in a sequence. The ultimate objective is to minimize the average number of samples required in order to identify the correlation structure of interest with target reliability. We characterize a sampling and decision-making strategy and establish its optimality properties, that are valid for all arbitrary (but well-behaved) statistical distributions. The problems of detecting or localizing correlation structures via data-adaptive sampling, albeit under different settings and objectives, are also studied in [27]–[31]. Specifically, the problem of data-adaptive sampling for deciding whether a graph of Gaussian random variables in the finite-horizon setting contains sparse locally correlated elements is investigated in [27] and [28]. The objective of [27] and [28] is to form a binary decision about whether such structures exist, and they do not localize the structures, which is in contrast to the objective of this paper. A similar correlation detection problem in the infinite-horizon setting is studied in [29]. The work in [30] deploys an adaptive sensing strategy based on correlation estimators to find the most mutually correlated nodes in a Gaussian graph. Finally, forming a binary decision about whether the entire graph exhibits a correlation structure is analyzed in [31].

II. Problem Formulation

A. Data Collection Model

Consider a graph, as depicted in Fig. 1, consisting of n nodes indexed by $\mathcal{N} \triangleq \{1,\dots,n\}$. Each node $i \in \mathcal{N}$ represents one information source that generates a sequence of random variables denoted by $\mathcal{X}^i \triangleq \{X_j^i \ : \ j \in \mathbb{N}\}$. It is assumed that there exists a subset of nodes denoted by $\mathcal{M} \subseteq \mathcal{N}$ that generate correlated random variables according to a specified kernel, while the remaining nodes in the graph generate independent and identically distributed (i.i.d.) random variables.

The objective is to identify set $\mathcal{M}$ with the fewest number of measurements, on average, while controlling the quality of the decision. For this purpose, we focus on a fully sequential sampling model in which the measurements are collected sequentially and one at-a-time up to a stochastic stopping time, at which point the data-gathering process is terminated and a decision about the set $\mathcal{M}$ is made. We denote the stochastic stopping time of the process by τ. In order to capture the index of the node from which we make a measurement at time $t \in \{1,\dots,\tau\}$, we define function $\psi : \{1,\dots,\tau\} \to \mathcal{N}$. Hence, the ordered sequence of the nodes from which we have made measurements up to time $t \in \{1,\dots,\tau\}$ is given by

$$\psi^t \triangleq \big[\psi(1),\dots,\psi(t)\big]\,. \tag{1}$$

The sampling strategy is data-adaptive and the information from all the past measurements is leveraged to guide the collection of future measurements. To characterize this process, we define Y_t as the measurement made at time t. Accordingly, for any given time $t \in \{1,\dots,\tau\}$ we define the ordered set of the measurements collected by

$$Y^t \triangleq [Y_1,\dots,Y_t]\,. \tag{2}$$

Hence, the information accumulated sequentially by observing the sequence of measurements Y^t from the sequence of nodes ψ^t can be abstracted by the filtration $\{\mathcal{F}_t : t = 1,2,\dots\}$, where $\mathcal{F}_t \triangleq \sigma(Y^t,\psi^t)$. Based on the information accumulated sequentially, i.e. $\mathcal{F}_t$, at time t the sampling procedure dynamically takes one of the following actions.

- A_1) *Exploration:* due to lack of sufficient information, no confident decision can be formed about the location of correlation structure and one more measurement is made; or,
- A_2) *Detection:* the sampling process terminates, and based on the collected information a reliable decision about the location of set $\mathcal{M}$ is made.

B. Data Model

For the convenience in notations we assume that the random variables generated by all nodes have well-behaved probability density functions (pdfs). We define f_0 as the marginal pdf of the i.i.d. random variables generated by the nodes in $\mathcal{N}\backslash\mathcal{M}$, i.e., the pdf of any random variable in $\{X_j^i : j \in \mathbb{N} \ \text{ and } \ i \notin \mathcal{M}\}$. For the correlated measurements generated by the nodes in $\mathcal{M}$ we adopt a spatio-temporal correlation model. In order to further specify this model, at each time t we partition the set of observed data Y^t into two subsets based on the membership of their corresponding nodes to the set $\mathcal{M}$. To this end, we first define two ordered sets of time instants $\mathcal{H}_t^0$ and $\mathcal{H}_t^1$ such that they partition the set $\{1,\dots,t\}$. The set $\mathcal{H}_t^0 \subseteq \{1,\dots,t\}$ is the sequence of time instants at which the measurement is taken from one node in set $\mathcal{N} \setminus \mathcal{M}$, and $\mathcal{H}_t^1 = \{1,\dots,t\}\backslash\mathcal{H}_t^0$ is the sequence of the time instants at which one node from set $\mathcal{M}$ is observed, i.e.,

$$\mathcal{H}_t^0 \triangleq \{\ s \in \{1,\dots,t\} \ : \ \psi(s) \notin \mathcal{M}\ \}, \tag{3}$$

$$\text{and} \quad \mathcal{H}_t^1 \triangleq \{\ s \in \{1,\dots,t\} \ : \ \psi(s) \in \mathcal{M}\ \}\,. \tag{4}$$

Based on these definitions we denote the joint pdf of the ordered sequence of random variables Y^t by f^t. Similarly, we define the joint pdf of the ordered sequence of random variables $\{Y^s : s \in \mathcal{H}_t^1\}$ by f_1^t. Based on the definitions of f_0 and f_1^t we have the following decomposition for f^t.

$$f^t(Y^t) = f_1^t\Big(\{Y_s \ : \ s \in \mathcal{H}_t^1\}\Big) \cdot \prod_{s\in\mathcal{H}_t^0} f_0(Y_s). \tag{5}$$

We assume that all joint distributions f_0, f^t and f_1^t are known.

C. Problem Formulation

The objective is to identify a subset of $\mathcal{M}$ with $p \in \{1, \ldots, |\mathcal{M}|\}$ nodes. Setting $p = 1$ represents the extreme case in which we identify only one element of $\mathcal{M}$, while setting $p = |\mathcal{M}|$ represents the other extreme in which we aim to identify $\mathcal{M}$ in its entirety. By defining $\mathcal{C}$ as the set of all sub-graphs with p nodes, we define

$$\delta : Y^\tau \times \psi^\tau \to \mathcal{C}, \tag{6}$$

as the decision rule exerted at the end of the sampling process for determining the indices of the p nodes of interest. By recalling the definitions of the stopping time τ and the node selection sequence ψ^τ, the tuple $\Phi \triangleq (\tau, \delta, \psi^\tau)$ uniquely specifies the intertwined information-gathering and decision-making processes. The adaptive strategy Φ is a measurable function with respect to the filtration $\mathcal{F}_t$.

In order to find the optimal choice of the sampling and decision making rules Φ, we define two figures of merit that capture the quality and the agility of the process. For a given strategy Φ, the agility of the process can be captured by the *average delay* in reaching a decision, which is given by

$$\mathsf{AD}(\Phi) \triangleq \mathbb{E}\{\tau\}\,. \tag{7}$$

The quality of the final decision, on the other hand, is captured by the frequency of erroneous decisions in localizing set $\mathcal{M}$, which for a given strategy Φ is defined as

$$\mathsf{DE}(\Phi) \triangleq \sup_{\mathcal{M}} \mathbb{P}(\delta \not\subseteq \mathcal{M})\,. \tag{8}$$

It is noteworthy that there exists an inherent tension between $\mathsf{AD}(\Phi)$ and $\mathsf{DE}(\Phi)$, as improving the quality of the decision, on average, requires collecting more measurements, which in turn penalizes the delay. Hence, designing an optimal sampling and decision-making strategy involves striking a balance between these two measures. For this purpose, we control the quality of the final decision and find the quickest sampling process among all possible adaptive strategies. Therefore, the optimal strategy is the solution to

$$\begin{aligned} \inf_{\Phi} \quad & \mathsf{AD}(\Phi) \\ \text{s.t.} \quad & \mathsf{DE}(\Phi) \leq \alpha \end{aligned} \tag{9}$$

where $\alpha \in (0, 1)$ controls the reliability of the final decision. We denote the set of all feasible solutions to (9) by $\Gamma(\alpha)$, i.e.,

$$\Gamma(\alpha) \triangleq \{\Phi \,:\, \mathsf{DE}(\Phi) \leq \alpha\}\,. \tag{10}$$

Finally, we define the following two information measures that are instrumental to characterizing the sampling process as well as the performance guarantees on the average delay of the process. Specifically, for any sequence of nodes ψ^t and their corresponding sequence of measurements Y^t, we assume that the following two sequences converge completely as $t \to \infty$:

$$\frac{1}{t} \log \frac{\prod_{s=1}^t f_0(Y_s)}{f_1^t(Y^t)} \to I_0, \quad \text{if } \psi^t \cap \mathcal{M} = \emptyset, \tag{11}$$

$$\text{and } \frac{1}{t} \log \frac{f_1^t(Y^t)}{\prod_{s=1}^t f_0(Y_s)} \to I_1, \quad \text{if } \psi^t \subseteq \mathcal{M}, \tag{12}$$

where $\emptyset$ is the empty set. The following two conditions, established in [32], are sufficient to guarantee complete convergence of (11) and (12). Specifically, when $\psi^t \cap \mathcal{M} = \emptyset$, if for any $\epsilon > 0$

$$\int_0^\infty \mathbb{P}\left(\left| \sup_{0 \leq u \leq t} \left[\log \frac{\prod_{s=1}^u f_0(Y_s)}{f_1^u(Y^u)} - u I_0 \right] \right| > \epsilon u \right) \mathrm{dt} < \infty$$

holds, then (11) converges completely. Similarly, when $\psi^t \subseteq \mathcal{M}$, if for any $\epsilon > 0$

$$\int_0^\infty \mathbb{P}\left(\left| \sup_{0 \leq u \leq t} \left[\log \frac{f_1^u(Y^u)}{\prod_{s=1}^u f_0(Y_s)} - u I_1 \right] \right| > \epsilon u \right) \mathrm{dt} < \infty$$

holds, the limit in (12) converges completely. Note that when the random variables generated by the nodes in set $\mathcal{M}$ are also independent but their distribution is different from f_0, the results reduces to the Stein's Lemma [33, Theorem 12.8.1] and the limits in (11) and (12) converge to the Kullback-Leibler divergence between the distributions.

D. Some Application Domains

1) Surveillance: The problem formalized in this paper arises in many application domains. For example consider a surveillance system which consists of a large array of sensors (e.g., cameras). For the area with no unusual activity, the sensors collect some static data that can be considered as background noise and assumed to be independent (or the measurements can be whitened via pre-processing). Hence, the data collected by each sensor monitoring the area can be modeled as an independent random variable with pdf f_0. Wherever there exists an unusual activity in a certain subregion, the observations made by the sensor in that specific sub-region will be following a different statistical distribution, and furthermore, they will be correlated since they observe a common phenomenon. This correlation structure among the measurements in that locality is captured by f_1. In this setting, f_0 can be, for instance, a standard Gaussian distribution, i.e., $f_0 = N(0, 1)$, and f_1 can refer to the joint pdf of a zero-mean multivariate Gaussian vector with covariance matrix $\mathbf{\Sigma}$, i.e., $f_1 = N(\mathbf{0}, \mathbf{\Sigma})$.

2) Fault Detection in Power Grids: Searching for local structures can also be applied to fault detection in power grids. A power grid can be modeled as a graph in which each bus represents a node and the transmission lines are the edges. Measurement devices are installed at each bus to collect voltage, current, and power measurements and send it to the central monitoring unit. It is shown that [34] these measurements bear a correlation structure that can be modeled by a random graph with the same connectivity structure as the physical network. In [35] it is shown that the measurements can be modeled as Gaussian random variables where the covariance matrix (correlation structure) depends on the topology of the graph. When a fault

event happens in one of the transmission lines or system buses, the local correlation structure changes. In order to prevent a cascade of failures and mitigate the cost of recovery, we should localize the faulty equipments in the quickest fashion. When there exist multiple faults in the grid, it may be of interest to localize a subset of faulty equipments as the reliability measures taken for multiple contingency enables the grid to operate normally under some contingencies. The optimal strategy for this case can be designed by selecting an appropriate parameter p.

In these applications data become available in real time and one by one. For instance, in a surveillance system or monitoring of a power system new data becomes available during the monitoring process and we want to localize the unusual activities such as intrusion in a surveillance system or fault in a power grid as quickly as possible. Therefore, we collect measurements as they become available and stop when a confident decision can be formed.

III. Quickest Search Algorithm

A. Definitions

We provide a constructive proof for solving the quickest search problem formalized in (9). For this purpose we specify two graph settings based on the relative values of I_0, I_1, and n, $|\mathcal{M}|$, and p, and then provide two algorithms each satisfying the optimality properties in one of these two setting. Intuitively, these tend to quickly identify p nodes that are the most likely candidates for being member of the set $\mathcal{M}$. Further measurements are taken afterwards (up to the stopping time) to further refine the set of the candidates by possibly replacing them with the ones emerging as the more likely candidates. For this purpose, to each node $i \in \mathcal{N}$ we assign a time series of weights denoted by $\mathcal{W}^i \triangleq \{w_t^i : t \in \{0, \dots, \tau\}\}$ such that w_t^i captures the posterior likelihood of $i \in \mathcal{M}$ at time t. Next, we discuss the rule for updating the elements of $\mathcal{W}^i$ over time. All these time series are initiated by setting $w_0^i = 0$ for all $i \in \mathcal{N}$. At time $t = 1$, we update the weight for node $\psi(1)$, i.e., the node observed at time $t = 1$ according to

$$w_1^i \triangleq \log \frac{f_1^1(Y_1)}{f_0(Y_1)}, \qquad \text{for } i = \psi(1)\,. \tag{13}$$

This process continues at time $t \in \{2, \dots, \tau\}$ by updating the weights for *only* the nodes that are observed up to time t, i.e., nodes represented by $\psi^t = [\psi(1), \dots, \psi(t)]$. Updating these weights is facilitated by defining the following log-likelihood ratio measures. At time t and corresponding to any arbitrary subset of measured nodes $B \subseteq \psi^t$, we define

$$L_t(B) \triangleq \log \frac{f_1^t\big(\{Y_s \ : \ \psi(s) \in B\}\big)}{\prod_{\psi(s)\in B} f_0(Y_s)} \tag{14}$$

as the log-likelihood ratio of measurements from nodes in set B up to time t. We also set $L_t(B) = 0$ when $B \cap \psi^t = \emptyset$. Based on this definition, at time t and corresponding to all the nodes observed up to time t, i.e., $\forall i \in \psi^t$, we update the weight w_t^i according to

$$w_t^i = \max_{\{B : B \subseteq \psi^t \,,\, i \notin B\}} \left[L_t(B \cup \{i\}) - L_t(B)\right]. \tag{15}$$

B. Dynamic Node Selection

The sequential sampling process terminates when a sub-graph of size at least p is deemed to be a subset of $\mathcal{M}$ with high accuracy. Prior to the stopping time, the process dynamically selects a node for sampling at each time. At time $t = 1$ we have no side information about any node and all the nodes are equally likely to belong to set $\mathcal{M}$. Hence, the selection rule picks one node from set $\mathcal{N}$ according to a uniform distribution. Based on the first measurement Y_1 taken from node $\psi(1)$, we update the weight of the observed node according to (13). The updates continue over time according to the rule specified in (15). Given the sequence of observed nodes ψ^t and their corresponding measurements Y^t, we define

$$\Lambda(t) \triangleq \max_{\{B \subseteq \psi^t : |B| = p\}} L_t(B), \tag{16}$$

as the generalized log-likelihood ratio of the measurements from the p most likely members of set $\mathcal{M}$ at time $t \in \mathbb{N}$. When $|\psi^t| < p$, which indicates that we have observed less than p nodes, we set $\Lambda(t) = 0$. We also define the set $\mathcal{P}(t)$ to track the set associated with the maximizer of (16), i.e.,

$$\mathcal{P}(t) \triangleq \{B \ : \ L_t(B) = \Lambda(t)\}\,. \tag{17}$$

In case there exist multiple choices for B, we select one randomly. Clearly we have $|\mathcal{P}(t)| = p$. We also define the node $d(t)$ as the node that is the $(|\mathcal{M}| + 1)$-th most likely candidate node to be a member of $\mathcal{M}$ based on the nodes most updated weights, i.e., there exist exactly $|\mathcal{M}|$ other nodes in the graph with larger weights. Based on this, we define $\Delta_i(t)$ as the difference between the log-likelihood ratio of nodes i and $d(t)$, i.e.,

$$\Delta_i(t) \triangleq w_t^i - w_t^{d(t)}\,. \tag{18}$$

The optimal stopping time and decision rules take different forms based on the relative values of I_0 and I_1 defined in (11) and (12), and the relative size of the correlation structure and the set of interest $q \triangleq \frac{n-m}{p}$, where we have set $m = |\mathcal{M}|$.

In order to localize the correlation structure, we can follow two different approaches; one is to identify the nodes that belong to the local structure and sample them to reinforce the belief; the other one is to reinforce the belief about the nodes that are less likely to belong to the local structure. The size of the set of nodes that do not belong to the local structure is q times the size of the desired set of p nodes in the local structure. Hence, for the second approach to be optimal we require the amount of the information provided by the independent nodes I_0 to be q times the information provided by the measurements from the nodes that belong to the local structure I_1. When $qI_1 = I_0$ both approaches have the same asymptotic performance.

1) Case 1: $qI_1 > I_0$: In this setting, the quickest approach indicates that we need to dynamically, and at any time t, compute the log-likelihood ratio defined in (14), and then maximize it to find $\Lambda(t)$ according to (16). Then the nodes contained in the maximizer of $\Lambda(t)$, denoted by $\mathcal{P}(t)$, are considered as candidates to be observed at time $(t + 1)$, among which one element is selected randomly. Hence, we set $\psi(t + 1)$ randomly and

according to the following distribution:

$$\mathbb{P}\big(\psi(t+1)=i\big)=\begin{cases}\frac{1}{p} & i\in\mathcal{P}(t)\\ 0 & i\notin\mathcal{P}(t)\end{cases}. \tag{19}$$

Based on this randomized selection rule, the measurement at time $(t+1)$ is collected from one of the p nodes that are considered to be the p most likely candidates to be members of $\mathcal{M}$.

2) Case 2: $qI_1 < I_0$: The procedure in this setting follows the same spirit as in the previous case, albeit with some distinctions. These changes are induced by the relative values of the information measures I_0 and I_1, and enforce a different optimal approach to finding the correlation structure of interest. Specifically, when $qI_1 < I_0$, identifying the correlation structure by directly identifying its members incurs more delay compared to an indirect approach in which the less likely candidates are identified and winnowed out. Based on this observation, we provide different decision rules for the optimal sampling process under this relative values of the information measures. Specifically, based on the information accumulated up to time t we set $\psi(t+1)$ according to

$$\psi(t+1)=d(t)\,. \tag{20}$$

This selection rule, intuitively, identifies the second most likely set of candidate nodes for being members of $\mathcal{M}$. This is in contrast to the behavior of node selection in previous setting, i.e., $qI_1 > I_0$, in which the selection rule dynamically identifies the most likely set of candidate nodes for being members of $\mathcal{M}$.

Remark 1: The underlying reason that (20) selects the *second* best set of candidate nodes is that it aims to balance a tradeoff. On the one hand it aims to winnow out some nodes that are not strong candidates for being members of $\mathcal{M}$, and on the other hand, it tends to increase the information that can be used later in the process, which prevents the sampling process to select the least likely nodes. In other words, if the node selection identifies and removes the *least* likely nodes, even though they are better candidates for being removed, nevertheless the measurements collected from those nodes are entirely uninformative about the location of the correlation structure of interest, and the overall impact of such a removal strategy will be an increase in the delay of the process.

Remark 2: The node selection rules introduced in (19) and (20) capture some of the ideas underlying the Chernoff rule developed for controlled sensing [21], [23], [24], which tends to select the action with the best immediate return (gain) based on the past observations. In the context of the setting of this paper, Chernoff's rule at each time t identifies the best set of candidates as if the stopping time is $\tau = t$. Then based on this decision, it selects the next set of nodes such that the decision made is reinforced most strongly.

C. Stopping Time and Decision Rule

The dynamic selection of the nodes terminates when the accumulated observations suffice to form a reliable final decision. Building on the likelihood ratio $\Delta_i(t)$ defined in (18), we adopt a likelihood ratio test as the stopping rule, and show that it achieves asymptotic optimality. By defining

$$\gamma \triangleq \log\frac{p(r-1)}{\alpha}, \tag{21}$$

where $r \triangleq \lceil \frac{n}{m} \rceil$, the sampling process terminates when the value of $\Delta_i(t)$ exceeds γ for p nodes. Hence, the stopping time is given by

$$\tau \triangleq \inf\{\, t \,:\, \Delta_i(t)\geq\gamma,\ \text{for } i\in\mathcal{P}(t)\,\}\,. \tag{22}$$

This implies that there exists a subset of observed nodes $\mathcal{P}(t)$ corresponding to which w_t^i is sufficiently larger than the rest of the nodes in $\mathcal{N}$ and we are confident enough the they belong to set $\mathcal{M}$. Also, after the stopping, the process identifies a set of p nodes as strong candidates for being members of $\mathcal{M}$, i.e.,

$$\delta = \mathcal{P}(\tau)\,. \tag{23}$$

The quickest search algorithm for isolating the local structures in random graph is summarized in Algorithm 1.

Algorithm 1: quickest search for local structures in random graphs.

Initialize $\gamma = \log\frac{p(r-1)}{\alpha}$, $t=0$,
$\Delta_i(0)=0$, and $w_0^i=0$ for $i\in\mathcal{N}$
While $|\{i\,:\,\Delta_i\geq\gamma\}|<p$ do
  $t\leftarrow t+1$
  $\mathcal{P}(t)\leftarrow p$ nodes with the largest w_t^i
  If $qI_1\geq I_0$
    $j\leftarrow$ Select one member of $\mathcal{R}$ randomly
  Else
    $j\leftarrow$ Select node with the $(m+1)$-th largest w_t^i
  End if
  $Y_t\leftarrow X_t^j$
  Update w_t^i for every node according to (15)
  Update $\Delta_i(t)$ according to (18)
End while
Return $\tau=t$ and $\delta=\mathcal{P}(t)$

IV. Performance Analysis

In this section we evaluate the performance of the proposed information-gathering and decision-making processes. For this purpose, we first characterize the error performance of the optimal sampling strategy in the following theorem.

Theorem 1: For the problem formulated in (9), in the asymptote of $\alpha\to 0$ for any sampling strategy $\Phi\in\Gamma(\alpha)$ we have

$$\mathsf{AD}(\Phi)\geq\min\left\{p\frac{|\log\alpha|}{I_1}, pq\frac{|\log\alpha|}{I_0}\right\}(1+o(1))\,. \tag{24}$$

Proof: See Appendix A. ■

We show that this performance can be achieved by the proposed adaptive sampling strategy. First, we characterize the decision error performance of the proposed sampling strategy in the following theorem.

Theorem 2: For any adaptive sampling strategy with the stopping time and the final decision rule given in (22) and (23), respectively, we have

$$\mathbb{P}(\delta \not\subseteq \mathcal{M}) \leq \alpha\,. \tag{25}$$

Proof: See Appendix B. ■

Next, we show that besides achieving the desired quality in the final decision, the proposed sequential strategy also exhibits optimality guarantees regarding the average delay in forming a decision. This asymptotic optimality is formalized in the following theorem.

Theorem 3: The proposed adaptive strategy with the dynamic node selection rule delineated by (19)-(20), and the stopping time and the final decision rule determined by (22) and (23), respectively, is optimal in the asymptote of $\alpha \to 0$, i.e.,

$$\mathbb{E}\{\tau\} = \mathsf{AD}(\Phi^*)(1+o(1)), \tag{26}$$

where Φ^* is the optimal sequential strategy.

Proof: See Appendix C. ■

According to this theorem, the sequential strategy proposed in this paper is optimal in the sense that it asymptotically achieves the minimum average delay for localizing set $\mathcal{M}$ with high accuracy.

It is noteworthy that while the node selection rules proposed in this paper are similar to the Chernoff rule in spirit, it requires no assumption on distributions to be distinguishable under every action, i.e., have positive KL divergence. The original Chernoff rule requires this assumption to achieve asymptotic optimality for a binary composite hypothesis testing [21], otherwise at certain time instants a randomization in sampling should be incorporated into the selection rule to ensure its asymptotic optimality [23].

V. Gain of Side Information

In many applications, it is often possible to obtain additional information about the membership of different nodes in a particular cluster. For instance some nodes might be exhibiting more similar stochastic behavior, whereas some might be known to be entirely independent. In this section we adopt a generic form side information, which essentially provides some prior information about which set of nodes have the potential of jointly forming a cluster and a correlation structure. Based on such a side information model, we assess the quality and the quickness of the process, where we show that significant gains can be achieved when the side information is exploited judiciously. We provide these gains in the context of large graphs, for which the nodes are heterogeneous enough to create a diverse set of statistical behavior, in which case assuming side information about the behavior of the nodes is more meaningful.

Specifically, we focus on the graphs in which n, m, and p simultaneously grow, but the ratio $\frac{n-m}{p}$ tends to a finite value, i.e., $\frac{n-m}{p} \xrightarrow{n,m,p\to\infty} q$. We assume that there exists some prior information about how the nodes are likely to form groups and be candidates for forming the correlation structure of interest. Specifically, we assume that the set of the nodes that are possible candidates to form set $\mathcal{M}$ together with node $i \in \mathcal{N}$ is a small subset of the graph, and for each node $i \in \mathcal{N}$ we denote the set of such nodes by $\mathcal{S}_i$. Without loss of generality, we assume that $|\mathcal{S}_i| \geq m$ for every $i \in \mathcal{N}$, since otherwise for any node ℓ with $|\mathcal{S}_\ell| < m$ we can conclude that $\ell \notin \mathcal{M}$ and discard such nodes and perform localization over the rest of the graph. Also, we assume that $|\mathcal{S}_i| < n-1$ for every $i \in \mathcal{N}$, since the setting $|\mathcal{S}_i| = n-1$ reduces to the setting that does not involve the aforementioned side information. Based on these definitions of $\{\mathcal{S}_i : i \in \mathcal{N}\}$, corresponding to any set $A \subseteq \mathcal{N}$ we define

$$\mathcal{S}_A = \left\{ i \;:\; i \in \bigcap_{j\in A} \mathcal{S}_j \right\}. \tag{27}$$

Set $\mathcal{S}_A$ is the set of all nodes that in conjunction with the nodes in A might form the local correlation structure of interest.

Defining sets $\{\mathcal{S}_i : i \in \mathcal{N}\}$ facilitates localizing the correlation structure more efficiently. For instance, if we measure node $i \in \mathcal{N}$ at $t=1$ and it turns out that node i belongs to the set $\mathcal{M}$ with high probability, then the likelihood of the nodes in set $\mathcal{S}_i$ being also members of $\mathcal{M}$ increases. Based on this observation, we note that when node i is observed and it is more (less) likely to belong to the set $\mathcal{M}$, i.e., the likelihood ratio of its measurements is greater (less) than 1, then the likelihood that the nodes in $\mathcal{S}_i$ are also members of the set $\mathcal{M}$ increases (decreases). Hence, after observing node i at time t we update the weights of all the nodes in $\mathcal{S}_i$ according to

$$\hat{w}_t^i = \max_{\{B:B\subseteq(\psi^t\cap\mathcal{S}_i)\,,\, i\notin B\}} \left[L_t(B\cup\{i\}) - L_t(B)\right]. \tag{28}$$

This weight updating rule in conjunction with the node selection rules provided in Section III-B determines the dynamic sampling strategy before the stopping time. In order to characterize the stopping time we define

$$\hat{\Delta}(t) \triangleq \Lambda(t) - \hat{w}_t^{d(t)}, \tag{29}$$

and set the stopping time

$$\hat{\tau} \triangleq \inf\{t \;:\; \hat{\Delta}(t) \geq \gamma\}\,. \tag{30}$$

The final decision is provided in Section III-C in (23).

By denoting the stopping time, final decision rule, and the node selection sequence of the quickest search strategy for identifying the local structure when side information is available by $\hat{\tau}$, $\hat{\delta}$, and $\hat{\psi}^\tau$, respectively, the tuple $\hat{\Phi} \triangleq (\hat{\tau}, \hat{\delta}, \hat{\psi}^\tau)$ uniquely specifies the processes. In order to analyze the performance of the proposed data collection and decision rules, we first characterize the error performance of the optimal sampling strategy in this setting in the following theorem.

Theorem 4: For the problem formulated in (9), when the graph grows such that $\frac{n-m}{p} \xrightarrow{n,m,p\to\infty} q$ and $p = o(|\log(\alpha)|)$, in the asymptote of $\alpha \to 0$ for any sampling strategy $\hat{\Phi} \in \Gamma(\alpha)$ we have

$$\mathsf{AD}(\hat{\Phi}) \geq \min\left\{ \frac{|\log\alpha|}{I_1}, q\frac{|\log\alpha|}{I_0} \right\}(1+o(1))\,. \tag{31}$$

Proof: See Appendix D. ■

By comparing the result of this theorem with that of Theorem 1 we observe that having side information in large graphs translated into a lower bound on $\mathsf{AD}(\hat{\Phi})$ that is smaller

by a factor p. Indeed comparing the lower bounds does not necessarily guarantee improvement. Nevertheless, in the sequel we show that the lower bound delineated by Theorem 4 is in fact achievable, which then establishes that we have p-fold improvement in the delay. In order to show that the lower bound provided by Theorem 4 can be achieved by the proposed strategy, in the next step we characterize the decision error performance of the proposed sampling strategy.

Theorem 5: The proposed sequential strategy with the final decision rule determined by (22) and (23) achieves

$$\mathbb{P}(\hat{\delta} \not\subseteq \mathcal{M}) \leq \alpha \,. \tag{32}$$

Proof: See Appendix E. ■

This theorem in conjunction with Theorem 4 establishes p-fold improvement in the delay of the sampling process when we have side information about the correlation structures. It is noteworthy that the condition required for achieving such gain is that p grows slowly enough, i.e., $p = o(|\log \alpha|)$. Next, we show that besides achieving the desired quality in the final decision, under some mild conditions on set $\mathcal{S}_i$ for every $i \in \mathcal{N}$, the proposed sequential strategy also exhibits optimality guarantees regarding the average delay in forming a decision. The asymptotic optimality of the designed sequential approach is stated in the following theorem.

Theorem 6: The proposed sequential strategy with the node selection rules given in (19)-(20), and the stopping time and the final decision rule determined by (22) and (23), is optimal in the asymptote of $\alpha \to 0$ and $n, m \to \infty$, i.e.,

$$\mathbb{E}\{\hat{\tau}\} = \mathsf{AD}(\hat{\Phi}^*)(1 + o(1)), \tag{33}$$

where $\hat{\Phi}^*$ is the optimal sequential strategy, provided that $|\mathcal{S}_A| = m + o(m)$ for any set A with $|A| \geq p$.

Proof: See Appendix F. ■

VI. Simulation Results

In this section, we evaluate the performance of the proposed sampling process via simulations. As a benchmark for comparing the performance of the proposed algorithm we use a fixed sampling method which can be described as follows. It is a non-sequential, non-adaptive sampling method in which the number of samples that will be collected from the graph and the nodes that will be sampled are pre-specified before the sampling. Also, the sampling budget is uniformly distributed among the nodes in the graph. For instance, if we want to collect kn samples from the graph and the graph has n nodes, each node will be sampled k times. After collecting the data, the final decision is based on a maximum generalized likelihood ratio given in (23). We also provide the analytical asymptotic lower bounds on average delay obtained in this paper for comparison. For the random variables generated in the graph, we consider a zero-mean Gaussian distribution with a covariance matrix given by

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } i \neq j \text{ and } \{i, j\} \subseteq \mathcal{M} \\ 0 & \text{otherwise} \end{cases}, \tag{34}$$

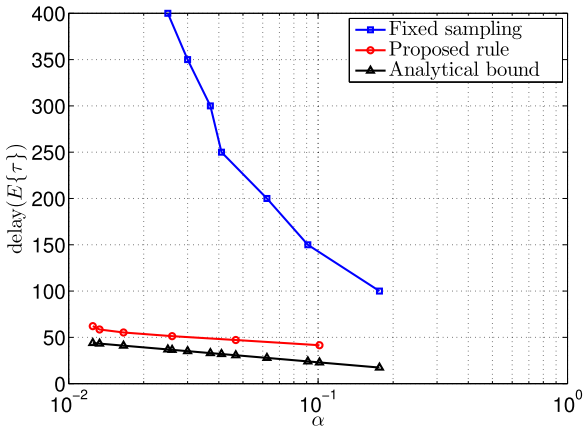

Fig. 2. Average delay versus decision quality for $\rho = 0.7$.

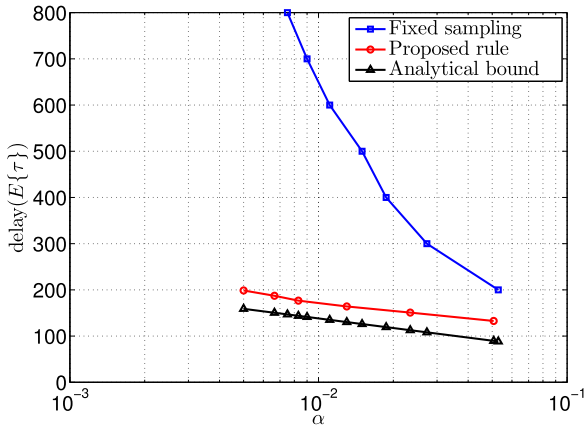

Fig. 3. Average delay versus decision quality for $\rho = 0.5$.

where ρ is a given constant determining the strength of the correlation among the nodes in $\mathcal{M}$.

In Fig. 2, we set $n = 400$, $m = 50$, $p = 10$, and $\rho = 0.7$, and compare the number of samples required by the proposed sampling process and the fixed sampling setting to achieve certain decision quality levels. It is observed that the proposed sampling process has a significant gain compared to the fixed sample-size setting. We also compare the results for weaker correlation models by setting $\rho = 0.5$ in Fig. 3. The relative performance of two methods is similar to Fig. 2, and we observe that for weaker correlation structures, i.e. smaller ρ, the average delay increases for the proposed sampling process, and the decision quality degrades for the fixed sampling strategy.

In order to compare the results with another data-adaptive approach, we consider the sequential thresholding strategy proposed in [27]. In the sequential thresholding method, the total number of samples is fixed, but the sampling budget is sequentially divided among the nodes. It starts by observing all the nodes and discard a fraction of them at each time. While it can observe multiple nodes at each time, we still can compare the total number of measurements for two methods. For this purpose, we set $n = 100$, $m = 20$, $p = 5$, and $\rho = 0.7$ and compare the average number of measurements versus the localization error probability in Fig. 4. It can be observed that our method out-

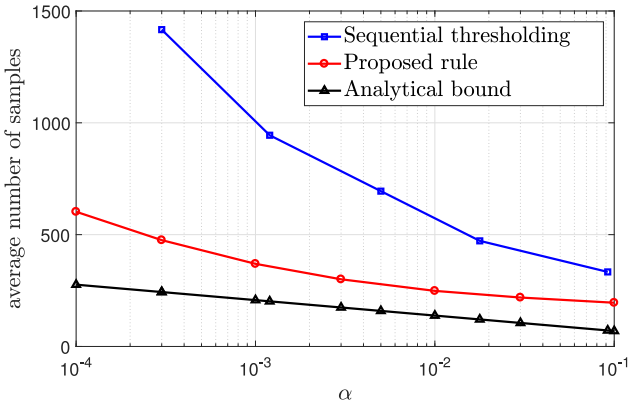

Fig. 4. Average number of measurements versus decision quality for $\rho = 0.7$.

performs sequential thresholding method uniformly. The main reason is that in the sequential thresholding method, at each time step all the past information is discarded and the attendant decision rules are only based on the current measurements.

VII. Conclusion

In this paper we have analyzed the problem of quickest search for identifying a local correlation structure in a random graph, in which a subset of nodes generate correlated data, and the rest generate independently of each other. The objective has been to co-design the information-gathering and decision-making processes for the purpose of localizing the cluster of the nodes that generate correlated data, with the fewest number of measurements and in the quickest fashion. A combination of sampling strategy and decision rules have been proposed, which exhibit asymptotic optimality in the sense that they minimize the expected number of measurements needed to make a decision while, in parallel, the rate of erroneous decisions is controlled.

Appendix A
Proof of Theorem 1

Throughout the proofs we assume that $r = \frac{n}{m} \in \mathbb{N}$. We denote the probability measure and the expectation with respect to pdf f_0 by $\mathbb{P}_0$ and $\mathbb{E}_0$, respectively. Also, we denote the probability measure and the expectation with respect to pdf f_1^t by $\mathbb{P}_1$ and $\mathbb{E}_1$, respectively. If we show that for all $0 < \eta < 1$

$$\lim_{\alpha\to 0} \inf_{\Phi} \mathbb{P}_0\left(\tau > \eta pq \frac{|\log\alpha|}{I_0}\right) = 1, \tag{35}$$

$$\text{and} \quad \lim_{\alpha\to 0} \inf_{\Phi} \mathbb{P}_1\left(\tau > \eta p \frac{|\log\alpha|}{I_1}\right) = 1, \tag{36}$$

then by applying Markov inequality, we obtain that for any $\eta < 1$

$$\mathbb{E}_0\left\{\frac{\tau}{\frac{pq|\log\alpha|}{I_0}}\right\} \geq \eta \cdot \mathbb{P}_0\left(\frac{\tau}{\frac{pq|\log\alpha|}{I_0}} > \eta\right) \overset{(35)}{\geq} \eta, \tag{37}$$

and

$$\mathbb{E}_1\left\{\frac{\tau}{\frac{p|\log\alpha|}{I_1}}\right\} \geq \eta \cdot \mathbb{P}_1\left(\frac{\tau}{\frac{p|\log\alpha|}{I_1}} > \eta\right) \overset{(36)}{\geq} \eta\,. \tag{38}$$

Since (37) and (38) hold for any $\eta < 1$ we should have

$$\mathbb{E}_0\left\{\frac{\tau}{\frac{pq|\log\alpha|}{I_0}}\right\} \geq 1,$$

$$\text{and} \quad \mathbb{E}_1\left\{\frac{\tau}{\frac{pq|\log\alpha|}{I_1}}\right\} \geq 1,$$

which give the lower bounds of Theorem 1. Now, we prove that when $qI_1 < I_0$, (35) holds and when $qI_1 > I_0$, (36) is true. Let divide the graph into q sub-graphs $\mathcal{N}_i$ for $i \in \{1,\ldots,r\}$ in a way that the set $\mathcal{M}$ is one of the sub-graphs, say $\mathcal{N}_1$. Also, we define the auxiliary decision rule $\tilde{\delta}$ as follows:

$$\tilde{\delta} = i, \quad \text{if } \delta \subseteq \mathcal{N}_i\,. \tag{39}$$

When $qI_1 > I_0$, we define the event

$$\mathcal{A}(i,L) \triangleq \{\tilde{\delta} = i\,,\, \tau \leq L\}\,. \tag{40}$$

Then, for any $L > 0$ and $F > 0$ we have

$$\begin{aligned}
\mathsf{DE}(\Phi) &= \mathbb{P}(\delta \not\subseteq \mathcal{M}) \\
&= \mathbb{P}(\delta \not\subseteq \mathcal{N}_1) \\
&\geq \mathbb{P}(\tilde{\delta} \neq 1) \\
&= \mathbb{E}\{\mathbb{1}_{(\tilde{\delta}\neq 1)}\} \\
&\geq \min\left\{\mathbb{E}\{\mathbb{1}_{(\tilde{\delta}\neq 1, \mathcal{P}(\tau)\subseteq\mathcal{M})}\}, \mathbb{E}\{\mathbb{1}_{(\tilde{\delta}\neq 1, \mathcal{P}(\tau)\not\subseteq\mathcal{M}))}\}\right\}.
\end{aligned} \tag{41}$$

We bound each term in (41) separately. For the first term and for any $G_0 < 0$ we have

$$\begin{aligned}
&\mathbb{E}\{\mathbb{1}_{(\tilde{\delta}\neq 1, \mathcal{P}(\tau)\subseteq\mathcal{M})}\} \\
&= \mathbb{E}_1\{\mathbb{1}_{(\tilde{\delta}\neq 1)}\} \\
&= \mathbb{E}_0\{\mathbb{1}_{(\tilde{\delta}\neq 1)} e^{\Lambda(\tau)}\} \\
&\geq \mathbb{E}_0\{\mathbb{1}_{(\mathcal{A}(j,L): j\neq 1, \Lambda(\tau) > G_0)} e^{\Lambda(\tau)}\} \\
&\geq e^{G_0}\,\mathbb{P}_0(\mathcal{A}(j,L): j \neq 1, \Lambda(\tau) > G_0) \\
&\geq e^{G_0}\,\mathbb{P}_0\Big(\mathcal{A}(j,L): j \neq 1, \sup_{t<L} \Lambda(t) > G_0\Big) \\
&\overset{(a)}{=} e^{G_0}\left(\mathbb{P}_0(\mathcal{A}(j,L): j \neq 1) - \mathbb{P}_0\big(\sup_{t<L} \Lambda(t) \leq G_0\big)\right) \\
&\overset{(b)}{=} e^{G_0}\left(\mathbb{P}_0(\tilde{\delta} \neq 1) - \mathbb{P}_0(\tau > L) - \mathbb{P}_0\big(\sup_{t<L} \Lambda(t) \leq G_0\big)\right),
\end{aligned} \tag{42}$$

where (a) and (b) hold due to the properties of set difference operation. Now we have

$$\begin{aligned}&\mathbb{P}_0(\tau > L)\\ &\quad\geq \mathbb{P}_0(\tilde{\delta}\neq 1) - e^{-G_0}\,\mathbb{P}(\tilde{\delta}\neq 1) - \mathbb{P}_0\big(\sup_{t<L}\Lambda(t)\leq G_0\big)\\ &\quad\geq 1-\alpha-e^{-G_0}\alpha - \mathbb{P}_0\big(\sup_{t<L}\Lambda(t)\leq G_0\big)\,. \end{aligned}\tag{43}$$

Since (43) holds for any $G_0<0$, we set $G_0=-cLI_0$ for some $c>1$. Then for any $1<K<L$ we obtain

$$\begin{aligned}&\mathbb{P}_0\Big(\sup_{t<L}\Lambda(t)\leq G_0\Big)\\ &=\mathbb{P}_0\Big(-\sup_{t<L}\Lambda(t)\geq -G_0\Big)\\ &=\mathbb{P}_0\Big(-\sup_{t<L}\Lambda(t)\geq cLI_0\Big)\\ &\leq \mathbb{P}_0\Big(-\sup_{t<K}\Lambda(t)-\sup_{K<t<L}\Lambda(t)\geq cLI_0\Big)\\ &\leq \mathbb{P}_0\left(-\sup_{t<K}\Lambda(t)-\sup_{K<t<L}\left(\frac{L}{t}\Lambda(t)\right)-LI_0\geq (c-1)LI_0\right)\\ &\leq \mathbb{P}_0\left(-\frac{1}{L}\sup_{t<K}\Lambda(t)-\sup_{K<t<L}\left(\frac{\Lambda(t)}{t}+I_0\right)\geq (c-1)I_0\right)\\ &\leq \mathbb{P}_0\left(-\frac{1}{L}\sup_{t<K}\Lambda(t)+\sup_{t>K}\Big|\frac{\Lambda(t)}{t}+I_0\Big|\geq (c-1)I_0\right)\,. \end{aligned}\tag{44}$$

According to (12), for any $\epsilon>0$ there exist a $\hat{K}(\epsilon)$ such that

$$\mathbb{P}_0\left(\Big|\frac{\Lambda(t)}{t}+I_0\Big|\leq\epsilon\right)=1,\quad \forall t>\hat{K}(\epsilon)\,.\tag{45}$$

Hence, we have

$$\begin{aligned}&\mathbb{P}_0\Big(-\sup_{t<L}\Lambda(t)\geq cLI_0\Big)\leq\\ &\qquad\mathbb{P}_0\left(-\frac{1}{L}\sup_{t<\hat{K}(\epsilon)}\Lambda(t)\geq (c-1)I_0-\epsilon\right)\,. \end{aligned}\tag{46}$$

Since, this is true for any $L>0$ and $c>1$, we assume the case that $L\to\infty$ and $c>1+\frac{\epsilon}{I_0}$. In this setting the right hand side of (46) approaches zero which indicates that for every $c>1$

$$\lim_{L\to\infty}\mathbb{P}_0\Big(-\sup_{t<L}\Lambda(t)\geq cLI_0\Big)=0\,.\tag{47}$$

Next, for any $0<\eta<\frac{1}{c}$, by defining

$$L_\alpha\triangleq \eta pq\frac{|\log\alpha|}{I_0},\tag{48}$$

and setting $L=L_\alpha$ we obtain

$$\begin{aligned}&\mathbb{P}_0\left(\tau>\eta p(q-1)\frac{|\log\alpha|}{I_0}\right)\geq\\ &\qquad 1-\alpha-\alpha^{1-\eta pqc}-\mathbb{P}_0\big(-\sup_{t<L_\alpha}\Lambda(t)\geq cI_0L_\alpha\big)\,. \end{aligned}\tag{49}$$

Now, by combining (47) and (49), and for the setting in which α approaches zero we obtain

$$\mathbb{P}_0\left(\tau>\eta pq\frac{|\log\alpha|}{I_0}\right)=1\,.\tag{50}$$

For the second term in (41), we can follow a similar procedure and obtain

$$\begin{aligned}&\mathbb{E}\{\mathbb{1}_{(\tilde{\delta}\neq 1,\mathcal{P}(\tau)\nsubseteq\mathcal{M})}\}\\ &\geq e^{-G_1}\Big(\mathbb{P}_1(\tilde{\delta}\neq 1)-\mathbb{P}_1(\tau>L)-\mathbb{P}_1\big(\sup_{t<L}\Lambda(t)\geq G_1\big)\Big), \end{aligned}\tag{51}$$

for some $G_1>0$. By following the same line of argument, we obtain

$$\mathbb{P}_1\left(\tau>\eta p\frac{|\log\alpha|}{I_1}\right)=1\,.\tag{52}$$

By defining

$$I^*=\max\left\{\frac{I_1}{p},\frac{I_0}{pq}\right\},\tag{53}$$

and combining (50) and (52), we obtain

$$\mathbb{P}\left(\tau>\eta\frac{|\log\alpha|}{I^*}\right)=1\,.\tag{54}$$

Since (54) holds regardless of the sampling procedure and stopping rule and only depends on the error performance of the strategy, it is valid for any strategy with the same decision quality, i.e.,

$$\lim_{\alpha\to 0}\inf_{\Phi}\ \mathbb{P}\left(\tau>\eta\frac{|\log\alpha|}{I^*}\right)=1,\tag{55}$$

and by applying the Chebyshev inequality the proof is concluded.

Appendix B
Proof of Theorem 2

This theorem is proved based on the union bound and the fact that when $\delta\nsubseteq\mathcal{M}$, *at least* one of the members of δ is not a member of $\mathcal{M}$. To analyze the decision error probability, we partition the graph into r sub-graphs $\{\mathcal{N}_j\}_{j=1}^r$ such that $\mathcal{N}_1=\mathcal{M}$. Note that each node $i\in\delta$ can belong to one of these subgraphs and when $i\in(\delta\setminus\mathcal{M})$ it is a member of one of $\mathcal{N}_j$, $j\in\{2,\ldots,r\}$. But from the stopping rule we know that the log-likelihood ratio of each node $i\in\delta$ is much larger than the rest of the graph. Hence, first by defining T_i as the sequence of time instants at which node $i\in\mathcal{N}$ has been observed, we provide the following lemma.

Lemma 1: For any node i that is identified as a member of set $\mathcal{M}$ at the stopping time τ, i.e., $i\in\delta$, and for any $j\in\{2,\ldots,r\}$ we have

$$\mathbb{P}\big(i\in\delta, i\in\mathcal{N}_j,\tau=t\big)\leq e^{-\gamma}\,\mathbb{P}(\tau=t)$$

Proof:

$$\begin{aligned}
&\mathbb{P}\big(i\in\delta, i\in\mathcal{N}_j, \tau=t\big)\\
&\quad\le \mathbb{P}\big(i\in\delta, \tau=t \mid i\in\mathcal{N}_j\big)\\
&\quad= \int_{(i\in\delta,\tau=t)} \Bigg(f_0(Y_{T_i}) f_1^t\big(\{Y_s \;:\; s\in\mathcal{H}_t^1\setminus\{i\}\}\big)\\
&\qquad\qquad \times \prod_{s\in\mathcal{H}_t^0\setminus T_i} f_0(Y_s)\Bigg)\,\mathrm{d}Y^t\\
&\quad= \int_{(\Delta_i(t)>\gamma)} \Bigg(f_0(Y_{T_i}) f_1^t\big(\{Y_s \;:\; s\in\mathcal{H}_t^1\setminus\{i\}\}\big)\\
&\qquad\qquad \times \prod_{s\in\mathcal{H}_t^0\setminus T_i} f_0(Y_s)\Bigg)\,\mathrm{d}Y^t\\
&\quad\overset{(a)}{\le} e^{-\gamma}\int_{(\Delta_i(t)>\gamma)} \Bigg(f_1^t(Y_{T_i}|Y^t\setminus Y_{T_i})\\
&\qquad\qquad \times f_1^t\big(\{Y_s \;:\; s\in\mathcal{H}_t^1\setminus\{i\}\}\big)\\
&\qquad\qquad \times \prod_{s\in\mathcal{H}_t^0\setminus T_i} f_0(Y_s)\Bigg)\,\mathrm{d}Y^t\\
&\quad\le e^{-\gamma}\mathbb{P}(\tau=t),
\end{aligned}$$

where (a) holds due to the definition of the stopping time. ■

Hence, for the probability of decision error we have

$$\begin{aligned}
\mathsf{DE}(\Phi) &= \mathbb{P}(\mathcal{P}(\tau)\nsubseteq\mathcal{M})\\
&= \sum_{t=1}^{\infty}\mathbb{P}(\mathcal{P}(t)\nsubseteq\mathcal{M},\tau=t)\\
&\le \sum_{t=1}^{\infty}\mathbb{P}(\cup_{i=1}^{p}\{i\}\nsubseteq\mathcal{M},\tau=t)\\
&\overset{(a)}{\le} \sum_{t=1}^{\infty}\sum_{i=1}^{p}\mathbb{P}(i\notin\mathcal{M},\tau=t)\\
&= \sum_{t=1}^{\infty}\sum_{i=1}^{p}\mathbb{P}\big(i\in\mathcal{N}_j \;:\; j\in\{2,\dots,r\},\tau=t\big)\\
&\overset{(b)}{\le} \sum_{t=1}^{\infty}\sum_{i=1}^{p}\sum_{j=2}^{r}\mathbb{P}\big(i\in\mathcal{N}_j,\tau=t\big)\\
&\overset{(c)}{\le} \sum_{t=1}^{\infty}\sum_{i=1}^{p}\sum_{j=2}^{r}e^{-\gamma}\mathbb{P}(\tau=t)\\
&= p(r-1)e^{-\gamma}\\
&= \alpha, \qquad (56)
\end{aligned}$$

where (a) and (b) are due to the union bound, and (c) holds because of Lemma 1.

Appendix C
Proof of Theorem 3

We prove this theorem in two steps. First we define τ_1 as the time instance after which the nodes with the largest likelihood ratios always belong to set $\mathcal{M}$, and show that τ_1 is exponentially bounded. Then, we prove that $\mathbb{E}\{\tau-\tau_1\}$ achieves the lower bound provided by the theorem. We show that for τ_1 we have

$$\mathbb{P}(\tau_1>t)\le K_1 e^{-c_1 t}, \qquad (57)$$

for some $K_1>0$ and $c_1>0$. We divide the graph into r subgraphs $\mathcal{N}_j$ for $j\in\{1,\dots,r\}$ such that $\mathcal{N}_1=\mathcal{M}$, and at time t denote the log-likelihood ratio of the measurements from subgraph $\mathcal{N}_j$ by $\bar{\Lambda}_j(t)$. Next we define an auxiliary random variable $\tilde{\tau}_1$ as the first time instant after which $\bar{\Lambda}_1(t)$ is always positive and $\bar{\Lambda}_j(t)$ for $j\in\{2,\dots,r\}$ are always negative. Since, it can be readily verified that $\tau_1\le\tilde{\tau}_1$, to prove (57) we show that

$$\mathbb{P}(\tilde{\tau}_1>t)\le K_1 e^{-c_1 t}\,.$$

For the left hand side we have

$$\begin{aligned}
\mathbb{P}(\tilde{\tau}_1>t) &\le \mathbb{P}(\bar{\Lambda}_1(t)<0) \qquad (58)\\
&= \mathbb{P}_1(\bar{\Lambda}_1(t)<0)\,. \qquad (59)
\end{aligned}$$

Hence, for any $s\le 0$ we have

$$\begin{aligned}
&\mathbb{E}_1\big\{e^{s\bar{\Lambda}_1(t)}\mathbb{1}_{(\bar{\Lambda}_1(t)<0)}\big\}\\
&\quad= \mathbb{P}_1(\bar{\Lambda}_1(t)<0)\cdot\mathbb{E}_1\big\{e^{s\bar{\Lambda}_1(t)}\mid\mathbb{1}_{(\bar{\Lambda}_1(t)<0)}\big\}\\
&\quad\le \mathbb{E}_1\big\{e^{s\bar{\Lambda}_1(t)}\big\}, \qquad (60)
\end{aligned}$$

and since $\mathbb{E}_1\big\{e^{s\bar{\Lambda}_1(t)}\mid\mathbb{1}_{(\bar{\Lambda}_1(t)<0)}\big\}\ge 1$ for any $s\le 0$, it yields

$$\mathbb{P}_1(\bar{\Lambda}_1(t)<0)\le\mathbb{E}_1\big\{e^{s\bar{\Lambda}_1(t)}\big\}. \qquad (61)$$

The right hand side of (61) can be rewritten by using the towering property of expectation as

$$\mathbb{E}_1\Big\{e^{s\bar{\Lambda}_1(t)}\Big\}=\mathbb{E}_1\left\{e^{s\bar{\Lambda}_1(t-1)}\mathbb{E}_1\left\{\left(\frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)}\right)^s\right\}\right\}. \qquad (62)$$

Now consider the inner expectation and define

$$\xi_t(s)\triangleq\mathbb{E}_1\left\{\left(\frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)}\right)^s\right\}\,. \qquad (63)$$

It is a convex function of s and is equal to 1 for $s=0,-1$. Two cases are possible for $\xi_t(s)$ when $-1<s<0$. In the first case, it is constant and $\xi_t(s)=1,\ \forall s\in(-1,0)$, which occurs only if the likelihood ratio inside the expectation is equal to 1, i.e., the measurement taken at time t has the same distribution under both f_0 and f_1^t, which cannot happen since we assumed that Y_t is taken from set $\mathcal{M}$. In the second case, $\xi_t(s)<1,\ \forall s\in(-1,0)$. It means that there exists a constant $c_1>0$ such that for some $0<s^*<1$ we have

$$\xi_t(s^*)\le e^{-c_1}<1\,. \qquad (64)$$

By successive application of this approach, we obtain

$$\mathbb{P}_1(\bar{\Lambda}_1(t)<0)\le\mathbb{E}_1\Big\{e^{s^*\bar{\Lambda}_1(t)}\Big\}\le K_1 e^{-c_1 t}\,. \qquad (65)$$

Next, we prove the asymptotic optimality property of the proposed adaptive strategy. To this end, at time t denote the log-likelihood ratio of the measurements from node i by $\tilde{\Lambda}_i(t)$. We show that when $qI_1 > I_0$, for any $i \in \delta$

$$\mathbb{P}_1(\tilde{\Lambda}_i(t) < p|\log\alpha|) \leq K_2 e^{-c_2 t}, \quad \text{for} \quad t \geq p\frac{|\log\alpha|}{I_1}(1+o(1)), \tag{66}$$

and when $qI_1 < I_0$, for any $j \notin \mathcal{M}$

$$\mathbb{P}_0(\tilde{\Lambda}_j(t) > pq\log\alpha) \leq K_3 e^{-c_3 t}, \quad \text{for} \quad t \geq pq\frac{|\log\alpha|}{I_0}(1+o(1))\,. \tag{67}$$

To prove (66), we denote the observation at time t by an optimal selection rule ψ_{opt} by $Y_{t,\,\text{opt}}$ and the sequence of observed nodes by Y^t_{opt}. Hence, we have

$$\begin{aligned}\tilde{\Lambda}_i(t) &= \left[\tilde{\Lambda}_i(t) - \mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\right\}\right] \\ &+ \left[\mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\right\} - \mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi_{\text{opt}}(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_{s,\text{opt}})}\right\}\right] \\ &+ \mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi_{\text{opt}}(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_{s,\text{opt}})}\right\}\end{aligned} \tag{68}$$

We consider the summands in (68) one by one. First, note that the last term is equal to $I_1 t$. For the first bracket we have

$$\tilde{\Lambda}_i(t) - \mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\right\} \tag{69}$$

$$= \tilde{\Lambda}_i(t-1) - \mathbb{E}_1\left\{\log\frac{f_1^{t-1}(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\right\} \tag{70}$$

$$+ \underbrace{\log\frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)} - \mathbb{E}_1\left\{\log\frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)}\right\}}_{\ell_t}. \tag{71}$$

Since $\mathbb{E}_1\{\ell_t\} = 0$, for any $\epsilon_1 > 0$ we have $\mathbb{E}_1\{\ell_t + \epsilon_1\} > 0$. Now consider the moment generating function of $\ell_t + \epsilon_1$ for any $s < 0$. Since for $s = 0$ the value of the moment generating function is 1 and we have

$$\left.\frac{\mathrm{d}\,\mathbb{E}_1\{e^{s(\ell_t+\epsilon_1)}\}}{\mathrm{d}s}\right|_{s=0} = \epsilon_1 > 0, \tag{72}$$

we can conclude that there exist $\hat{s} < 0$ and $\hat{c}_2 > 0$ such that

$$\mathbb{E}_1\{e^{\hat{s}(\ell_t+\epsilon_1)}\} \leq e^{-\hat{c}_2} < 1\,. \tag{73}$$

Successively applying this technique yields

$$\mathbb{E}_1\left\{e^{\hat{s}\left(\tilde{\Lambda}_i(t) - \mathbb{E}_1\{\log\frac{f_1^t(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\} + \epsilon_1 t\right)}\right\} \leq e^{-\hat{c}_2 t}, \tag{74}$$

Next by using the same line of thought as in (61) we obtain

$$\mathbb{P}_1\left(\tilde{\Lambda}_i(t) - \mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\right\} < -\epsilon_1 t\right) \leq e^{-\hat{c}_2 t}. \tag{75}$$

In order to work with the second bracket in (68), according to the definition of τ_1, for $t > \tau_1$ we have

$$\mathbb{E}_1\left\{\log\frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)}\right\} = \mathbb{E}_1\left\{\log\frac{f_1^t(Y_{t,\text{opt}}(t)|Y^{t-1}_{\text{opt}})}{f_0(Y_{t,\text{opt}})}\right\}. \tag{76}$$

Therefore, by defining

$$\zeta_t \triangleq \mathbb{E}_1\left\{\log\frac{f_1^t(\{Y_s\ :\ \psi(s)=i\})}{\prod_{s:\psi(s)=i} f_0(Y_s)}\right\} - \mathbb{E}_1\left\{\log\frac{f_1^t(Y^t_{\text{opt}})}{\prod_{s:\psi(s)=i} f_0(Y_{s,\text{opt}})}\right\}, \tag{77}$$

we can conclude that for some $K_4 > 0$, we have $|\zeta_t| \leq \tau_1 K_4$, and since τ_1 is exponentially bounded according to (57), for some $\hat{c}_3 > 0$ we have

$$\mathbb{P}_1(\zeta_t > \epsilon_2 t) \leq e^{-\hat{c}_3 t} < 1\,. \tag{78}$$

Finally, by defining $\epsilon_3 \triangleq \epsilon_1 + \epsilon_2$ and combining (68), (75), and (78) we obtain

$$\mathbb{P}_1\left(\tilde{\Lambda}_i(t) < (I_1 - \epsilon_3)t\right) \leq K_2 e^{-c_2 t}, \tag{79}$$

which concludes the proof for (66). The bound in (67) can be proved by following the same line of argument.

Appendix D
Proof of Theorem 4

To prove this theorem, we consider an auxiliary setting in which some extra information about the location of the nodes in set $\mathcal{M}$ is given. Therefore, the average delay of our setting will be lower bounded by the average delay of this auxiliary setting. In the new setting, we assume that the graph is divided into r known sub-graphs $\mathcal{N}_i$ for $i \in \{1,\ldots,r\}$. It is known that $\mathcal{M} = \mathcal{N}_{i^*}$ for some $i^* \in \{1,\ldots,r\}$. Therefore, the RVs generated from the nodes in one of the sub-graphs are drawn from the joint distribution f_1^t, and for the other $r-1$ sub-graphs the RVs have marginal pdf f_0. This problem is studied in [25] and it is shown that the lower bound on the average delay is the one provided in this theorem in (31).

Appendix E
Proof of Theorem 5

In Theorem 2, we proved the feasibility of the sampling process for any values of m and n. Hence, the proposed

stopping time and decision rule guarantee the feasibility when $n, m \to \infty$.

APPENDIX F
PROOF OF THEOREM 6

The idea for proving this theorem is the same as Theorem 3. By defining τ_1 as the time instance after which the nodes with the largest likelihood ratios always belong to set $\mathcal{M}$, we can follow the same line of argument to show that

$$\mathbb{P}(\tau_1 > t) \leq K_1 e^{-c_1 t}, \tag{80}$$

for some $K_1 > 0$ and $c_1 > 0$. To show the asymptotic optimality, we only need to show that by changing the stopping rule for large graphs, the factor p in the average delay disappears. To this end, we divide the graph into r sub-graphs denoted by $\mathcal{N}_i$ for $i \in \{1, \ldots, r\}$ in a way that $\mathcal{M} = \mathcal{N}_1$, and at time t denote the log-likelihood ratio of the measurements from sub-graph $\mathcal{N}_i$ by $\bar{\Lambda}_i(t)$. We show that when $qI_1 \geq I_0$,

$$\mathbb{P}_1(\bar{\Lambda}_1(t) < |\log \alpha|) \leq K_2 e^{-c_2 t}, \quad \text{for} \quad t \geq \frac{|\log \alpha|}{I_1}(1 + o(1)), \tag{81}$$

and when $qI_1 < I_0$,

$$\mathbb{P}_0(\bar{\Lambda}_j(t) > \log \alpha) \leq K_3 e^{-c_3 t}, \quad \text{for} \quad t \geq q\frac{|\log \alpha|}{I_0}(1 + o(1)) \, . \tag{82}$$

To prove (81), we denote the node observed at time t by an optimal selection rule by $Y_{t,\,\text{opt}}$ and the sequence of observed nodes by Y^t_{opt}. Hence, we have

$$\begin{aligned}\bar{\Lambda}_1(t) = \bar{\Lambda}_1(t) &- \mathbb{E}_1\left\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s) \in \mathcal{N}_1\})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_s)}\right\} \\ &+ \mathbb{E}_1\left\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s) \in \mathcal{N}_1\})}{\prod_{s=1}^t f_0(Y_s)}\right\} \\ &- \mathbb{E}_1\left\{\log \frac{f_1^t(Y^t_{\text{opt}})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_{s,\text{opt}})}\right\} \\ &+ \mathbb{E}_1\left\{\log \frac{f_1^t(Y^t_{\text{opt}})}{\prod_{s=1}^t f_0(Y_{s,\text{opt}})}\right\}\end{aligned} \tag{83}$$

We consider the summands in (83) one by one. First, note that the last term is equal to $I_1 t$. For the first bracket we have

$$\bar{\Lambda}_1(t) - \mathbb{E}_1\left\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s) \in \mathcal{N}_1\})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_s)}\right\} \tag{84}$$

$$= \bar{\Lambda}(t-1) - \mathbb{E}_1\left\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s) \in \mathcal{N}_1\})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_s)}\right\} \tag{85}$$

$$+ \underbrace{\log \frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)} - \mathbb{E}_1\left\{\log \frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)}\right\}}_{\ell_t} \, . \tag{86}$$

Since $\mathbb{E}_1\{\ell_t\} = 0$ for any $\epsilon_1 > 0$ we have $\mathbb{E}_1\{\ell_t + \epsilon_1\} > 0$. Now consider the moment generating function of $\ell_t + \epsilon_1$ for any $s < 0$. Since for $s = 0$ the value of the moment generating function is 1 and we have

$$\left.\frac{\mathrm{d}\, \mathbb{E}_1\{e^{s(\ell_t+\epsilon_1)}\}}{\mathrm{d}s}\right|_{s=0} = \epsilon_1 > 0, \tag{87}$$

we can conclude that there exists some $\hat{s} < 0$ and $\hat{c}_2 > 0$ such that

$$\mathbb{E}_1\{e^{\hat{s}(\ell_t+\epsilon_1)}\} \leq e^{-\hat{c}_2} < 1 \, . \tag{88}$$

Successively applying this technique yields

$$\mathbb{E}_1\left\{e^{\hat{s}\left(\bar{\Lambda}_1(t) - \mathbb{E}_1\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s)\in\mathcal{N}_1\})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_s)}\} + \epsilon_1 t\right)}\right\} \leq e^{-\hat{c}_2 t}, \tag{89}$$

Next by using the same line of thought as in (61) we obtain

$$\mathbb{P}_1\left(\bar{\Lambda}_1(t) - \mathbb{E}_1\left\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s) \in \mathcal{N}_1\})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_s)}\right\} < -\epsilon_1 t\right) \leq e^{-\hat{c}_2 t} \, . \tag{90}$$

In order to work with the second bracket, according to the definition of τ_1, for $t > \tau_1$ we have

$$\mathbb{E}_1\left\{\log \frac{f_1^t(Y_t|Y^{t-1})}{f_0(Y_t)}\right\} = \mathbb{E}_1\left\{\log \frac{f_1^t(Y_{t,\text{opt}}(t)|Y^{t-1}_{\text{opt}})}{f_0(Y_{t,\text{opt}})}\right\} \, . \tag{91}$$

Therefore, by defining

$$\begin{aligned}\zeta_t \triangleq \mathbb{E}_1&\left\{\log \frac{f_1^t(\{Y_s \, : \, \psi(s) \in \mathcal{N}_1\})}{\prod_{s:\psi(s)\in\mathcal{N}_1} f_0(Y_s)}\right\} \\ &- \mathbb{E}_1\left\{\log \frac{f_1^t(Y^t_{\text{opt}})}{\prod_{s=1}^t f_0(Y_{s,\text{opt}})}\right\},\end{aligned} \tag{92}$$

we can conclude that for some $K_4 > 0$

$$|\zeta_t| \leq \tau_1 K_4, \tag{93}$$

and since τ_1 is exponentially bounded according to (80), for some $\hat{c}_3 > 0$ we have

$$\mathbb{P}_1(\zeta_t > \epsilon_2 t) \leq e^{-\hat{c}_3 t} < 1 \, . \tag{94}$$

Finally, by defining $\epsilon_3 \triangleq \epsilon_1 + \epsilon_2$ and combining (83), (90), and (94) we obtain

$$\mathbb{P}_1\left(\bar{\Lambda}_1(t) < (I_1 - \epsilon_3)t\right) \leq K_2 e^{-c_2 t}, \tag{95}$$

which concludes the proof for (81). The bound in (82) can be proved by following the same line of argument.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, Mar. 2002.

[2] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDNs and web sites," in *Proc. 11th Int. Conf. World Wide Web*, New York, NY, USA, May 2002, pp. 293–304.
[3] S. Shekhar, C.-T. Lu, and P. Zhang, "Detecting graph-based spatial outliers: Algorithms and applications (a summary of results)," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.*, New York, NY, USA, Aug. 2001, pp. 371–376.
[4] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 5, pp. 631–645, May 2007.
[5] J. Lin, E. Keogh, A. Fu, and H. Van Herle, "Approximations to magic: Finding unusual medical time series," in *Proc. IEEE Symp. Comput.-Based Med. Syst.*, Dublin, Ireland, Jun. 2005, pp. 329–334.
[6] K. Steinhaeuser and N. Chawla, " Community detection in a large real-world social network," in *Social Computing, Behavioral Modeling, and Prediction*, H. Liu, J. Salerno, and M. Young, Eds. New York, NY, USA: Springer, 2008, pp. 168–175.
[7] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying sequence detection in large data sets: A data-driven approach," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 44–56, Sep. 2014.
[8] 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.
[9] E. Arias-Castro, S. Bubeck, and G. Lugosi, "Detection of correlations," *Annals Statist.*, vol. 40, no. 1, pp. 412–435, Feb. 2012.
[10] 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.
[11] Q. Berthet and P. Rigollet, "Optimal detection of sparse principal components in high dimension," *Annals Stat.*, vol. 41, no. 4, pp. 1780–1815, Aug. 2013.
[12] A. Hero and B. Rajaratnam, "Hub discovery in partial correlation graphs," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 6064–6078, Sep. 2012.
[13] E. Arias-Castro, D. L. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2402–2425, Jul. 2005.
[14] G. Walther, "Optimal and fast detection of spatial clusters with scan statistics," *Annals Stat.*, vol. 38, no. 2, pp. 1010–1033, Apr. 2010.
[15] M. P. Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman, "False discovery control for random fields," *J. Amer. Stat. Assoc.*, vol. 99, no. 468, pp. 1002–1014, Dec. 2004.
[16] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of an anomalous disk over a two-dimensional lattice network," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 2394–2398.
[17] 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.
[18] E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *Annals Stat.*, vol. 36, no. 4, pp. 1726–1757, Aug. 2008.
[19] E. Arias-Castro, E. J. Candès, and A. Durand, "Detection of an anomalous cluster in a network," *Annals Statist.*, vol. 39, no. 1, pp. 278–304, Feb. 2011.
[20] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, U.K.: Cambridge Univ. Press, Dec. 2009.
[21] H. Chernoff, "Sequential design of experiments," *Annals Math. Statist.*, vol. 30, no. 3, pp. 755–770, Sep. 1959.
[22] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *Annals Statist.*, vol. 41, no. 6, pp. 2703–2738, Dec. 2013.
[23] 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, May 2013.
[24] 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.
[25] K. Cohen and Q. Zhao, "Active hypothesis testing for anomaly detection," *IEEE Trans. Inf. Theory*, vol. 61, no. 3, pp. 1432–1450, Mar. 2015.
[26] 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.
[27] R. Castro, G. Lugosi, and P.-A. Savalle, "Detection of correlations with adaptive sensing," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7913–7927, Dec. 2014.
[28] 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.
[29] J. Geng, W. Xu, and L. Lai, "Detection of correlated components in multivariate Gaussian models," in *Proc. IEEE Inf. Theory Workshop*, Jeju Island, South Korea, Oct. 2015, pp. 224–228.
[30] C.-Y. Liu and S. Bubeck, "Most correlated arms identification," *arXiv:1404.5903v1*, Apr. 2014.
[31] J. Heydari, A. Tajer, and H. V. Poor, "Quickest detection of Gauss–Markov random fields," in *Proc. 53rd Annu. Allerton Conf. Commun. Control, Comput.*, Monticello, IL, USA, Sep. 2015, pp. 808–814.
[32] A. Tartakovsky, "Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case," *Stat. Inference Stoch. Process.*, vol. 1, no. 3, pp. 265–295, Oct. 1998.
[33] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[34] H. Sedghi and E. Jonckheere, "On the conditional mutual information in the Gaussian–Markov structured grids," in *Information and Control in Networks*. Springer International Publishing, 2014, pp. 277–297.
[35] M. He and J. Zhang, "A dependency graph approach for fault detection and localization towards secure smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 2, pp. 342–351, Jun. 2011.

Javad Heydari received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2010 and 2012, respectively. He is currently working toward the Ph.D. degree in the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA. During 2010–2012, he was with the Advanced Communication Research Institute, Sharif University of Technology, where he worked on the applications of signal processing in communication systems. His research interests include the general areas of statistical signal processing, wireless communications, and statistical information theory.

Ali Tajer (S'05–M'10–SM'15) received the M.A degree in statistics and Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA, during 2007–2010. He is currently an Assistant Professor of electrical, computer, and systems engineering in Rensselaer Polytechnic Institute, Troy, NY, USA. During 2010–2012, he was a Postdoctoral Research Associate at Princeton University. His research interests include mathematical statistics and network information theory, with applications in wireless communications and power grids. He is an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, an Editor of the IEEE TRANSACTIONS ON SMART GRID. He was the Guest Editor-in-Chief of the IEEE TRANSACTIONS ON SMART GRID Special Issue on Theory of Complex Systems with Applications to Smart Grid Operations. He received the NSF CAREER Award in 2016.