Anomaly Search Over Discrete Composite Hypotheses in Hierarchical Statistical Models

Tomer Gafni, Benjamin Wolff, Guy Revach, Nir Shlezinger, *Member, IEEE*, and Kobi Cohen, *Senior Member, IEEE*

***Abstract*—Detection of anomalies among a large number of processes is a fundamental task that has been studied in multiple research areas, with diverse applications spanning from spectrum access to cyber-security. Anomalous events are characterized by deviations in data distributions, and thus can be inferred from noisy observations based on statistical methods. In some scenarios, one can often obtain noisy observations aggregated from a chosen subset of processes. Such hierarchical search can further minimize the sample complexity while retaining accuracy. An anomaly search strategy should thus be designed based on multiple requirements, such as maximizing the detection accuracy; efficiency, be efficient in terms of sample complexity; and be able to cope with statistical models that are known only up to some missing parameters (i.e., composite hypotheses). In this paper, we consider anomaly detection with observations taken from a chosen subset of processes that conforms to a predetermined tree structure with partially known statistical model. We propose Hierarchical Dynamic Search (HDS), a sequential search strategy that uses two variations of the Generalized Log Likelihood Ratio (GLLR) statistic, and can be used for detection of multiple anomalies. HDS is shown to be order-optimal in terms of the size of the search space, and asymptotically optimal in terms of detection accuracy. An explicit upper bound on the error probability is established for the finite sample regime. In addition to extensive experiments on synthetic datasets, experiments have been conducted on the DARPA intrusion detection dataset, showing that HDS is superior to existing methods.**

***Index Terms*—Active hypothesis testing, anomaly detection, composite hypotheses testing, sequential design of experiments.**

I. Introduction

THE task of detecting anomalies in data streams arises in a wide variety of applications. These applications include dynamic spectrum access and sensing in wireless communication [2]; detecting attacks and intrusions in computer networks [3]; and detecting anomalies in infrastructures that may indicate catastrophes [4]. Such tasks involve distinguishing anomalous processes from typical ones based on noisy observations.

The noisy nature of the observations implies that the typical behavior can be modeled by a normal or benign distribution, and the anomalous behavior is captured by an abnormal distribution. The goal of a *decision-maker* boils down to deciding whether to reject the null hypothesis and to declare a process as anomalous. Here we consider the task of detecting an anomalous process (or processes) out of a large set of data streams. This requires to sample (observe) each process at least once, and preferably more for better detection accuracy due to uncertainty. Hence, a *decision-maker* should design an *efficient* sampling policy, that for a given detection accuracy minimizes the number of samples needed to reach a decision, or alternatively, given a sampling budget maximizes the detection accuracy.

The class of problems involving a sequential design of experiments for active binary hypothesis testing problem was pioneered by Chernoff [5]. Chernoff proposed a randomized strategy and showed that it is asymptotically optimal as the error probability approaches zero. However, the Chernoff test results in a linear sample complexity in the size of the search space. When the number of processes (data streams) is very large, as is often the case in practice, it is likely to be inefficient to sample each process multiple times. Therefore, sampling strategies with a sub-linear sampling complexity are desirable. The need for efficiency requires to exploit a certain structure in the data, which may lead to a significant performance gain. A common structure that can be utilized for this end is the ability to access the data in a hierarchical fashion.

The hierarchical structure model represents settings where a massive number of data streams can be observed at different levels of granularity. Such modeling faithfully captures the operation of various applications. In finance, transactions can be aggregated at different temporal and geographic scales [6]. In visual monitoring applications, the ability to zoom-in or zoom-out is equivalent to the aggregation of pixels, and can lead to faster detection of anomalies (targets, interesting events) by giving suspicious pixels more attention than others [7]. In internet traffic monitoring, there is a need for detecting heavy hitters, i.e., a small number of flows that accounts for most of the total traffic, and thus representing the measurements as a

Manuscript received 9 August 2022; revised 22 December 2022; accepted 30 January 2023. Date of publication 3 February 2023; date of current version 16 February 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yao Xie. This work was supported in part by the Israel Science Foundation under Grant 2640/20, and in part by the Israel National Cyber Directorate via the Cyber Security Research Center at Ben- Gurion University of the Negev. A short version of this paper that introduces the algorithm, and preliminary simulation results was presented at the IEEE International Symposium on Information Theory (ISIT) 2022 [DOI: 10.1109/ISIT50566.2022.9834631]. *(Tomer Gafni and Benjamin Wolff contributed equally to this work.) (Corresponding author: Tomer Gafni.)*

Benjamin Wolff and Guy Revach are with the Institute for Signal and Information Processing (ISI), D-ITET, ETH, 8092 Zürich, Switzerland (e-mail: bewolff@student.ethz.ch; grevach@ethz.ch).

Tomer Gafni, Nir Shlezinger, and Kobi Cohen are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: gafnito@post.bgu.ac.il; nirshl@bgu.ac.il; kobi.cohen10@gmail.com).

Digital Object Identifier 10.1109/TSP.2023.3242074

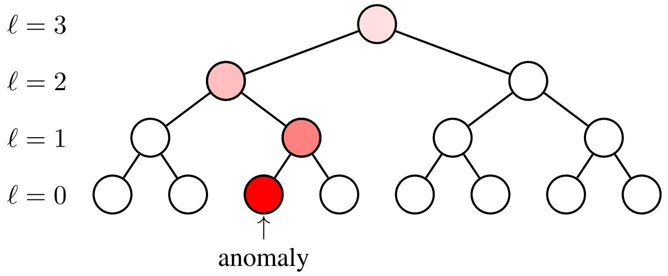


Fig. 1. A binary tree observation model with $M = 8$ processes, $\log_2 M = 3$ levels, and a single anomaly. The anomaly is measurable at the red nodes.

tree structure, where each node represents an aggregated flow can lead to efficient detection [8]. Other applications include direction of arrival estimation [9] and system control [10].

In light of the aforementioned potential gains of the hierarchical structure, here, we consider the problem of detecting anomalous processes (targets), for which there is uncertainty in the distribution of observations. We assume that in each time step, a decision-maker can observe a chosen subset of processes that conforms to a predetermined tree structure, and get access to aggregated observations that are drawn from a general distribution that depends on a chosen subset of processes as schematically depicted in Fig. 1. The uncertainty in the anomalous distribution yields a composite hypothesis case, where measurements drawn when observing a subset of processes follow a common distribution parameterized by an unknown vector when containing the target. The objective is to design a sampling policy (a search strategy), that minimizes a Bayesian risk that accounts for sample complexity and detection accuracy, by selecting which subset to observe, and when to terminate the search and make a decision, in an adaptive way.

Dynamic search strategies were proposed for various forms of anomaly detection problems. In [11], the Information-Directed Random Walk (IRW) algorithm was proposed, for cases where the statistical model is fully known. IRW was shown to be asymptotically optimal in terms of detection accuracy and order optimal with respect to the number of processes. When the anomalous hypothesis is composite, the IRW policy serves as a benchmark for the performance one can achieve with partially known modeling. The recent studies [12], [13], [14] considered hierarchical search under unknown observation models. The search strategies in [12], [13] are based on a sample mean statistic, which fails to detect a general anomalous distribution with a mean close to the mean of the benign distribution. The work in [14] does not assume a structure on the abnormal distribution, and uses the Kolmogorov-Smirnov statistic, which fails to utilize the parametric information considered in our setting. This motivates the derivation of a dynamic search policy for data of hierarchical structure which can cope with partially known anomalous distributions and reliably detect based on statistics of a higher order than a sample mean.

In this work we consider for the first time the task of hierarchical anomaly detection over a general and known family of distributions with unknown parameters. Here, the measurements can take continuous values and the decision-maker is allowed to sample an aggregated subset of processes that conforms to a tree structure. To cope with this observation model in a dynamic search setting with possibly multiple anomalies of different types, we develop a dedicated sequential search strategy, dubbed HDS. HDS uses two carefully chosen statistics to harness the information on the null hypothesis and the structure of the hierarchical samples, allowing it to achieve asymptotically optimal performance. The proposed policy is shown to be asymptotically optimal with respect to the detection accuracy and order optimal with respect to the size of the search space.

Extensive numerical experiments on synthetic and real datasets support the theoretical results. Our numerical evaluation shows that HDS effectively captures changes in the traffic that are associated with network anomalies. HDS with active local tests for the high level nodes is also analyzed numerically and is shown to outperform the fixed sample-size local test and approach the performance bound of IRW. Our non-synthetic experiments numerically evaluate the performance of HDS in a cyber-security task using the DARPA intrusion detection dataset. We show that the proper modeling of the network traffic data in a hierarchical fashion combined with the application of HDS can successfully detect denial of service (DoS) attacks from a limited number of samples.

The rest of this paper is organized as follows: in Section II we present the system model and discuss its relationship with the existing literature. Section III designs the HDS policy and analyzes its performance. We numerically evaluate HDS in Section IV, and provide concluding remarks in Section V.

II. System Model and Preliminaries

In this section, we describe the statistical setting of our system model and discuss some of the relevant related literature on dynamic search policies.

A. Problem Formulation

Anomaly Detection: We consider the problem of detecting K anomalous (targets) processes (data streams) out of a large set of M processes, where K is assumed to be known. Here, the decision-maker should actively collect evidence (data, observations, samples), and identify the K anomalous processes. Since there is cost on gaining samples, this problem presents an inherent trade-off between the need to maximize the detection accuracy to the need to minimize the length of the exploration phase.

In particular, in each time step t, where $t \in \{1, 2, \ldots\}$, the decision-maker can access only one process and sample an observation y_t in an i.i.d. manner. The main challenge is to know when to stop exploring and to reach a decision. We denote the data collected in the time horizon τ and provide a decision based on $\mathbf{D}_\tau = \{y_t\}_{t=1}^{\tau}$. Given the collected evidence, the decision rule boils down to *simultaneous* testing of multiple binary hypotheses. Let $\mathcal{H}_m = 0$ denote the *null hypothesis*, i.e., the process m is not anomalous, then the decision-maker should decide whether to reject the null hypothesis, and declare process m as anomalous, i.e., $\mathcal{H}_m = 1$, or not. Assuming that at time t, process m was chosen to be sampled by the decision-maker, then its sampling distribution is given by

$$y_t \sim f(y\,|\,\boldsymbol{\theta}), \quad \begin{cases} \boldsymbol{\theta} = \boldsymbol{\theta}_0, & \mathcal{H}_m = 0 \\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_1, & \mathcal{H}_m = 1 \end{cases} \tag{1}$$

where $f(\cdot)$ is a known family of a parametric probability distributions. While $\boldsymbol{\theta}_0$ is a known parameter describing the distribution of the non-anomalous samples, for anomalous processes, the parameter is not assumed to be known, but only that it is restricted to belong to a known finite set $\boldsymbol{\Theta}_1$. Note that in this paper we consider the case where $\boldsymbol{\Theta}_1$ is equal for all K anomalous processes. Also, the setting is homogeneous, where the distribution of a benign process is the same for all $M-K$ benign processes, and the distribution of an abnormal process is the same for all K abnormal processes.

Hierarchical Sampling: To reach a decision, the decision-maker must actively sample information from the M processes. Generally speaking, if the complexity of an active sampling policy is linear, when the number of processes M scales up, such policy becomes inefficient and can be computationally infeasible. Therefore, to cope more efficiently with a large number of processes, and to reduce the sampling complexity, we consider the case of hierarchical data streams. Here, in addition to observing individual processes, the decision-maker can measure aggregated processes that conform to a binary tree structure. Sampling an internal node of the tree gives a blurry image of the processes beneath it, as depicted in Fig. 1. The key to utilizing the hierarchical structure of the sampling space to its full extent, is to determine the number of samples one should obtain at each level of the tree, and when to zoom in or out on the hierarchy.

To model hierarchical sampling, let the tuple (l, j) denote node j at level l of the tree, with $l = 0, \ldots, \log_2 M$ and $j = 1, \ldots, 2^{\log_2 M - l}$. The tree structure encodes the relationship between the nodes. The abnormal distribution of a target leads to an abnormal distribution in every ancestor of the target, i.e., every node on the shortest path from this target to the root. We denote by $\mathcal{H}_{(l,j)} = 0$ the hypothesis that node (l, j) is not anomalous, and $\mathcal{H}_{(l,j)} = 1$ denotes that it is anomalous.

The observations y_t of an internal node j on level ℓ of the tree follow a similar statistical model as in (1):

$$y_t \sim f_\ell(y\,|\,\boldsymbol{\theta}), \quad \begin{cases} \boldsymbol{\theta} = \boldsymbol{\theta}_0^{(\ell)}, & \mathcal{H}_{(l,j)} = 0 \\ \boldsymbol{\theta} \in \boldsymbol{\Theta}_1^{(\ell)}, & \mathcal{H}_{(l,j)} = 1 \end{cases} \tag{2}$$

where $f_\ell(\cdot)$, $\boldsymbol{\theta}_0^{(\ell)}$, and $\boldsymbol{\Theta}_1^{(\ell)}$ are the probability distribution, the known parameter of the non-anomalous distribution, and the set of the anomalous parameter, respectively, at level ℓ, and $f_0(\cdot) \equiv f(\cdot)$, $\boldsymbol{\theta}_0^{(0)} \equiv \boldsymbol{\theta}_0$, $\boldsymbol{\Theta}_1^{(0)} \equiv \boldsymbol{\Theta}_1$. Here, we assume that the observations at all levels are informative, as formulated in the following:

AS1 The Kullback Leibler (KL) divergence $\mathcal{D}_\ell(\cdot||\cdot)$ between $f_\ell(\cdot\,|\,x)$ and $f_\ell(\cdot\,|\,z)$ satisfies:

$$\mathcal{D}_\ell\left(\boldsymbol{\theta}_0^{(\ell)}||\boldsymbol{\theta}\right) \geq \Delta, \quad \mathcal{D}_\ell\left(\boldsymbol{\theta}||\boldsymbol{\theta}_0^{(\ell)}\right) \geq \Delta, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_1^{(\ell)}, \tag{3}$$

for some $\Delta > 0$ independent of M for all l, where

$$\mathcal{D}(x||z) \triangleq \mathbb{E}_{f(y\,|\,x)}\left[\log \frac{f(y\,|\,x)}{f(y\,|\,z)}\right]. \tag{4}$$

Note that *AS1* implies that the anomalous and non-anomalous distributions are distinguishable. We allow a general relation between $f_\ell(y\,|\,\boldsymbol{\theta})$ and $f_0(y\,|\,\boldsymbol{\theta})$, which often depends on the specific application. We next provide two examples for the internal observation model.

Example 1: Consider the problem of detecting a heavy hitter [12] among Poisson flows where the measurements are exponentially distributed packet inter-arrival times. For the leaf nodes, the benign processes have an exponential distribution with parameter λ_0, i.e., $\boldsymbol{\theta}_0 = \lambda_0$, and the anomalous process has an exponential distribution, where it is known that its parameters lie in the set $\boldsymbol{\Theta}_1 = \{\lambda_1, \lambda_2\}$, where $\lambda_2 > \lambda_1 > \lambda_0$. Moreover, in the case of heavy hitter detection where the measurements are packet counts of an aggregated flow as in this example, $f_\ell(y\,|\,\boldsymbol{\theta})$ is given by multi-fold convolutions of $f_0(y\,|\,\boldsymbol{\theta})$. For independent Poisson flows, $f_\ell(y\,|\,\boldsymbol{\theta})$ is also exponential with mean values given by the sum of the mean values of its children at the leaf level. That is, in this case, $f_\ell(y\,|\,\boldsymbol{\theta})$ is also exponential, with $\boldsymbol{\theta}_0^{(\ell)} = 2^\ell \lambda_0$ and $\boldsymbol{\Theta}_1^{(\ell)} = \{(2^\ell - 1)\lambda_0 + \lambda_1, (2^\ell - 1)\lambda_0 + \lambda_2\}$. As is the case in most of the practical applications, we expect that observations from each individual process are more informative than aggregated observations. More precisely, we expect $\mathcal{D}_\ell\left(\boldsymbol{\theta}_0^{(\ell)}||\boldsymbol{\theta}\right) \geq \mathcal{D}_{\ell-1}\left(\boldsymbol{\theta}_0^{(\ell-1)}||\boldsymbol{\theta}\right)$ and $\mathcal{D}_\ell\left(\boldsymbol{\theta}||\boldsymbol{\theta}_0^{(\ell)}\right) \geq \mathcal{D}_{\ell-1}\left(\boldsymbol{\theta}||\boldsymbol{\theta}_0^{(\ell-1)}\right)$ for all $\ell > 0$ (note that these inequalities indeed hold in the example above). However, the results in this work hold for the general case without these monotonicity assumptions.

Example 2: Consider the Bernoulli distribution, which is widely adopted in the literature of hypothesis testing and group testing and also arises in distributed detection of aggregating local binary decisions. Assume that $f_\ell(y\,|\,\boldsymbol{\theta})$ follows Bernoulli distributions with parameters μ_ℓ and $1 - \mu_\ell$ for the benign processes and the anomalous processes, respectively. The parameter μ_ℓ can be set according to the specific application, and the observation is not a summation over all its leaf nodes. A special case of this example is the size-independent Bernoulli noise studied in [15]. Applying our setting to this special case implies: $\mu_\ell = \mu'_\ell$ for all l, l'. We provide more concrete examples in our numerical experiments in Section IV.

Search Policy: An active search (i.e., sampling and decision) policy (strategy), denoted as π, is defined by the tuple (ϕ, τ, δ). Here, ϕ is a sampling selection rule, i.e., a mapping from the time t and the data collected to a node from which we need to sample next, namely

$$\phi : (t, \mathbf{D}_t) \mapsto (\ell, j); \tag{5}$$

a stopping rule τ, i.e., the time at which the decision-maker decides to end the search; and a decision rule δ

$$\delta : (\mathbf{D}_\tau) \mapsto \{0, 1\}^M, \tag{6}$$

which is a mapping from the evidence collected until stopping time τ, to a Boolean vector of size M, where $\delta_m = 1$ corresponds to the decision that process m is anomalous.

Aim: We aim to find a policy π^* out of the set of possible policies $\mathbf{\Pi}$ that minimizes the *Bayesian risk*, namely

$$\pi^* = \arg\min_{\pi \in \mathbf{\Pi}} \{\mathcal{R}(\pi)\}, \tag{7}$$

where

$$\mathcal{R}(\pi) \triangleq \mathcal{P}_{\text{Err}}(\pi) + c \cdot \mathcal{Q}(\pi). \tag{8}$$

The term $\mathrm{P}_{\text{Err}}(\pi)$ is the error probability, $\mathcal{Q}(\pi)$ is the sample complexity, and $c \in (0,1)$ is the sampling cost assigned for each observation. Specifically, let $\mathcal{H}$ be a set of all Boolean vectors of size M with exactly K entries equal to 1, such that $|\mathcal{H}| = \binom{M}{K}$. Let $H_b \in \mathcal{H}$ be any Boolean vector of size M with exactly K entries equal to 1, and $H \in \mathcal{H}$ be a Boolean vector of size M corresponding to the true hypothesis, where an mth entry equals to 1 implies that process m is anomalous. Then, the error probability given that $H = H_b$ is:

$$\mathcal{P}_{\text{Err}}(\pi | H = H_b) \triangleq \mathcal{P}(\delta(\mathbf{D}_\tau) \neq H_b | \pi, H = H_b), \tag{9}$$

and the error probability is averaged under given prior p_b for hypothesis $H = H_b$:

$$\mathcal{P}_{\text{Err}}(\pi) \triangleq \sum_{H_b \in \mathcal{H}} p_b \cdot \mathcal{P}_{\text{Err}}(\pi | H = H_b), \tag{10}$$

[1] where

$$p_b \triangleq \mathcal{P}(H = H_b). \tag{11}$$

Similarly, the sampling complexity is averaged under prior p_b for hypothesis $H = H_b$, and is given by:

$$\mathcal{Q}(\pi) \triangleq \mathbb{E}[\tau | \pi]. \tag{12}$$

Generally, the priors p_b are determined according to the specific application using offline measurements, as commonly done in the Bayesian approach in the statistics literature. For example, when detecting an anomaly in an image, some areas might be more likely to contain the anomaly (e.g., due to the geographical characteristics of the area). When detecting attacks in computer networks, for example, some devices might be more vulnerable to attacks compared to other devices.

B. Related Literature

Target search problems have been widely studied under various scenarios. Optimal policies for target search with a fixed sample size were derived in [16], [17], [18], [19] under restricted settings involving binary measurements and symmetry assumptions. Results under the sequential setting can be found in [20], [21], assuming single process observations. In this paper we address the optimality question under the asymptotic regime as the error probability approaches zero. Asymptotically optimal results for sequential anomaly detection in a linear (i.e., non-hierarchical) search under various setting can be found in [22], [23], [24], [25]. In this paper, however, we consider a composite hypothesis case with finitely many parameter values, which was not addressed in the above. Results under the composite hypothesis case with linear search can be found in [26], [27], [28], [29], [30]. Detecting anomalies or outlying sequences has also been studied under different formulations, assumptions, and objectives [31], [32], [33], [34], [35], [36]; see survey in [37]. These studies, in general, do not address the optimal scaling in the detection accuracy or the size of the search space.

As mentioned in Section I, the problem considered here also falls into the general class of sequential design of experiments pioneered by Chernoff in [5]. Compared with the classical sequential hypothesis testing pioneered by Wald [38] where the observation model under each hypothesis is fixed, active hypothesis testing has a control aspect that allows the decision-maker to choose different experiments (associated with different observation models) at each time. The work [15] developed a variation of Chernoff's randomized test that achieves the optimal logarithmic order of the sample complexity in the number of hypotheses under certain implicit assumptions on the KL divergence between the observation distributions under different hypotheses. These assumptions, however, do not always hold for general observation models as considered here.

In contrast to Chernoff's randomized policy, in this paper we propose an active *deterministic* strategy. The work [39] have showed that a simpler deterministic algorithm applies in this setting and obtained the same asymptotic performance as Chernoff's policy, with better performance in the finite sample regime under a linear search setting with known distributions. A modified algorithm has been developed in [40] for spectrum scanning with time constraint. This setting was extended in [41] to the composite case, which proposed an asymptotically optimal deterministic policy. The problem addressed in this work is fundamentally different, focusing on efficient exploitation of aggregated and potentially low-quality measurements to achieve an optimal sublinear order with the size of the search space.

Note that we consider the case where the densities of the non-anomalous and anomalous distributions are known, except for some unknown parameters in the anomalous distribution. In this case, the GLLR and ALLR tests have great properties in practice, with strong optimality characteristics (which lead to the asymptotic optimality properties of HDS). These tests rely on all the available statistical knowledge, which goes beyond first and second-order statistical moments, thus inherently using higher-order moments for detection. Consequently, the usage of higher-order moments is encapsulated in the formulation of the tests comprising the proposed HDS algorithm. In cases where a side knowledge is given about the distribution moments, one can propose a modification of the algorithm that takes this into account, e.g., moment-based methods in [12], [13], and may result with better performances in some special cases. However, such algorithms require additional assumptions on the distributions (e.g., large statistical distance which can be

[1] Note that the error probability is defined with respect to the event of detecting a wrong process as abnormal, and thus is averaged under given prior p_b over the hypotheses. Therefore, the setting and error analysis coincides with a multi-hypothesis testing, and it is more general than averaging over type I and type II errors under binary hypothesis testing. We will show later that the error probability is bounded by $O(c)$. When translating to binary hypothesis testing in the special case of two hypotheses, the power of test is thus of order $1 - O(c)$.

detected efficiently by moment-based detection methods) and are not robust when the assumptions are violated.

Another related problem considered in the literature deals with detecting the first disorder of a system involving multiple processes, refer to as change point detection [42], [43], [44]. In this problem, a change occurs at some unknown time in the distribution of a sequence of random vectors that are monitored online, and the goal is to detect this change as quickly as possible subject to a certain false alarm constraint. A cumulative sum (CUSUM) test was established under this setting [45]. In this paper, however, the goal is to detect the abnormal processes (and not a change point), where the process states are fixed during the detection process.

Tree-based search in data structures is a classical problem in computer science (see, for example, [46], [47]). It is mostly studied in a deterministic setting; i.e., the observations are deterministic when the target location is fixed. The problem studied in this work is a statistical inference problem, where the observations taken from the tree nodes follow general statistical distributions. This problem also has intrinsic connections with several problems studied in different application domains, that are particularly geared towards handling high dimensional data. We discuss here three representative paradigms most pertinent to this paper and emphasize the differences in our approach from these existing studies:

1) The first is group testing, where the objective is to identify the defective items in a large population by performing tests on subsets of items that reveal whether the tested group contains any defective items. Formulations of group testing can be mapped to our setting by mapping the individual items to the leaf nodes of a tree. The action of testing a node on the tree corresponds to a group test. Differ from our setting, most existing work on Boolean group testing assumes error-free test outcomes, or limited noise models (e.g., binary symmetric noise or one-sided noise [48], [49]). Moreover, most of the existing results on noisy group testing focus on non-adaptive open-loop strategies [50], [51], [52], and the issue of sample complexity in terms of the detection accuracy is absent in the basic formulation.
2) The second is compressed sensing (CS), and particularly the Boolean CS setting [48], where the objective is to recover a sparse binary signal with aggregated observations. That is, in CS we are given an N-dimensional sparse signal with support size K. Random projections of the sparse signal are obtained. The goal is to identify the support set while minimizing the number of projections. Similarly as in group testing, the individual signal components in Boolean CS can be mapped to the leaf node in a tree, and the action of testing a node on the tree corresponds to a test of aggregated observations. Our setting, for which the proposed HDS policy is designed, aims for an adaptive solution to solve the compressed sensing with little offline or online computation and low memory requirement. The policy works for the general noisy observation models.
3) Our setting also applies to a special setting of offline change point detection, arises in the fundamental problem of estimating a step function in [0,1] [53]. Suppose that we are given a stream with a statistical change, and our objective is to identify when the change occurred. In this case, the decision maker sequentially chooses sampling points within the given interval and observes a noisy version of their values. This problem can be cast as one studied in this work, by partitioning the interval into δ-length sub-intervals, which form the $M = 1/\delta$ leaf nodes, with the sub-interval containing the change being the target. Successively combining two adjacent sub-intervals leads to a binary tree with the root being the entire interval. The main body of work on adaptive sampling is based on a Bayesian approach with binary noise of a known model. Although several strategies (e.g., the Probabilistic Bisection Algorithm) have been extensively studied in the literature [54], [55], there is little known about the theoretical guarantees, especially when it comes to unknown noise models. HDS, derived in the sequel based on the problem formulated in Subsection II-A can be considered as a non-Bayesian approach to the adaptive sampling problem under general parametric noise models, and its theoretical guarantees apply in this problem.

III. Hierarchical Dynamic Search

In this section we present and analyze the proposed HDS active search strategy. We start by introducing the algorithm in the case of one anomaly (i.e., $K = 1$) in Subsection III-A, after which we analyze its performance in Subsection III-B. In Subsection III-C we extend HDS to multi-target setting, and we conlclude the section with a discussion in Subsection III-D.

A. Algorithm Design

We start by focusing on detecting a single target ($K = 1$).

Rationale: The anomaly is searched using a random walk on the process tree that starts at the root node. The individual steps of the walk are determined by local tests. On internal (i.e., high level) nodes, the outcome of the test can be moving to the left or right child, or returning to the parent node (where the parent of the root is itself). The internal test is constructed to create a bias in the walk toward the anomalous leaf. On a leaf node of index m, the possible outcomes are either terminating the search and declaring process m anomalous, or moving back to parent node. The leaf test is designed to terminate at the anomaly with sufficiently high probability.

In particular, HDS uses the fixed sample size GLLR statistic for the high level nodes test and the sequential Adaptive Log Likelihood Ratio (ALLR) statistic for the leaf nodes test. The ALLR statistic, introduced by Robbins and Siegmund [56], [57], builds upon the one-stage delayed estimator of the unknown parameter; i.e., the density of the n-th observation is estimated based on the previous $n-1$ observations, while the current observation is not included in this estimate. As opposed to the GLLR, the ALLR preserves the martingale properties. This allows one to choose thresholds in a way to control specified rates of error probability, and so to ensure the desired asymptotic properties. In the following, we specify the internal and leaf tests.

Internal Test: Suppose that the random walk arrives at a node on level $\ell > 0$. A fixed number $K_{\ell-1}$ of samples $y(i)$ is drawn from both children, and are used to compute the GLLRs

$$\tilde{S}^{(l-1)}_{\text{GLLR}}(K_{l-1}) \triangleq \sum_{i=1}^{K_{\ell-1}} \log \frac{f_{\ell-1}\left(y(i)\,|\hat{\boldsymbol{\theta}}_1^{(l-1)}\right)}{f_{\ell-1}\left(y(i)\,|\,\boldsymbol{\theta}_0^{(l-1)}\right)}, \tag{13}$$

where$\hat{\boldsymbol{\theta}}_1^{(l-1)}$ is the maximum likelihood estimate of the anomaly parameter, given by

$$\hat{\boldsymbol{\theta}}_1^{(l-1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1^{(l-1)}} \prod_{i=1}^{K_{\ell-1}} f_{\ell-1}(y(i)\,|\,\boldsymbol{\theta}). \tag{14}$$

The statistics (13) utilize the information on the benign distribution. If at least one of the children has a strictly positive GLLR, the random walk moves to the child with the greater GLLR. Otherwise, it moves to the parent. The sample size K_ℓ for $\ell = 0, \ldots, \log_2 M - 1$ is determined offline, such that the probability of moving in the direction of the anomaly is greater than $\frac{1}{2}$. Note that K_ℓ is finite under *AS1*.

Leaf Test: When the random walk visits a leaf node, we perform an ALLR test. Here, samples $y(i)$ are drawn sequentially from the process and the local ALLR

$$\tilde{S}_{\text{ALLR}}(n) = \sum_{i=1}^{n} \log \frac{f_0\left(y(i)\,|\hat{\boldsymbol{\theta}}_1^{(0)}(i-1)\right)}{f_0\left(y(i)\,|\,\boldsymbol{\theta}_0^{(0)}\right)}, \tag{15}$$

is continuously updated, where

$$\hat{\boldsymbol{\theta}}_1^{(0)}(i-1) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1^{(0)}} \prod_{j=1}^{i-1} f_0(y(j)\,|\,\boldsymbol{\theta}), \tag{16}$$

is the delayed maximum likelihood estimate of $\boldsymbol{\theta}_1^{(0)}$. To initialize the estimate $\hat{\boldsymbol{\theta}}_1^{(0)}(0)$, a fixed number $N_{\text{leaf}} \geq 0$ (which is independent of M, c) of samples is drawn from the leaf. In Appendix B we elaborate on how to set N_{leaf}. As opposed to the GLLR, $\tilde{S}_{\text{ALLR}}(n)$ is a viable likelihood ratio, so that the Wald likelihood ratio identity can still be applied to upper-bound the error probabilities of the sequential test [38].

At every time step $n > 0$, the ALLR (15) is examined: If $\tilde{S}_{\text{ALLR}}(n) > \log \frac{\log_2 M}{c}$,[2] the random walk terminates and the tested process is declared anomalous, while a negative ALLR results in returning to the parent node. Thus, the stopping time τ is defined as

$$\tau = \inf_{n\geq 1} \left\{ \tilde{S}_{\text{ALLR}}(n) > \log \frac{\log_2 M}{c} \right\}. \tag{17}$$

The choice of the threshold $\log \frac{\log_2 M}{c}$ is to ensure that the error probability is in the order of $O(c)$, which in turn ensures the asymptotic optimality in c, as we elaborate in the proof outline in the next section. The resulting search policy is summarized in Algorithm 1.

Algorithm 1: Single Target HDS.

```
   Input: Inspected node at level $\ell$
1  if $l > 0$ (internal node) then
2      Measure $K_{\ell-1}$ samples from each child node;
3      Compute GLLR for each child via (13);
4      if Both GLLRs are negative then
5          Invoke Algorithm 1 on parent node;
6      else
7          Invoke Algorithm 1 on child with larger GLLR;
8  else
9      Init $\boldsymbol{\theta}_1^{(0)}$ according to (50) and $n = 1$;
10     Draw $y(n)$ and compute ALLR (15);
11     if $\tilde{S}_{\text{ALLR}}(n) > \log \frac{\log_2 M}{c}$ then
12         Identify node as target and terminate;
13     else if $\tilde{S}_{\text{ALLR}}(n) < 0$ then
14         Invoke Algorithm 1 on parent node;
15     Increment $n$ and jump to step 9;
```

B. Performance Analysis

We next theoretically analyze the HDS policy, denoted π_{HDS}, for $K = 1$. In particular, we establish that π_{HDS} is asymptotically optimal in c, i.e.,

$$\lim_{c\to 0} \frac{\mathcal{R}(\pi_{\text{HDS}})}{\mathcal{R}^*} = 1, \tag{18}$$

and order optimal in M, namely,

$$\lim_{M\to\infty} \frac{\mathcal{R}(\pi_{\text{HDS}})}{\mathcal{R}^*} = O(1) \tag{19}$$

where $\mathcal{R}^*$ is a lower bound on the Bayesian risk, i.e.,

$$\mathcal{R}^* = \inf_{\pi} \mathcal{R}(\pi). \tag{20}$$

This is stated in the following theorem:

Theorem 1: When *AS1* holds and $\boldsymbol{\Theta}_1^{(\ell)}$ is finite for all $0 \leq \ell \leq \log_2 M - 1$, the Bayesian risk of π_{HDS} is bounded by

$$\mathcal{R}(\pi_{\text{HDS}}) \leq cB\log_2 M + \frac{c \log \frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)} || \boldsymbol{\theta}_0^{(0)}\right)} + O(c)\ , \tag{21}$$

where B is a constant independent of M and c.

Proof: The complete proof is given in Appendix B. Here, we only present the proof outline, which divides the trajectory of the random walk into two stages: *search* and *target test*.

In the *search stage* the random walk explores the high level nodes and is expected to eventually concentrate on the true anomaly. Based on this insight, we partition the tree $\mathcal{T}$ into a sequence of subtrees $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_{\log_2 M}$ (Fig. 2). Subtree $\mathcal{T}_{\log_2 M}$ is obtained by removing the halftree that contains the target from $\mathcal{T}$. Subtree $\mathcal{T}_\ell$ is iteratively obtained by removing the halftree that contains the target from $\mathcal{T}\backslash\mathcal{T}_{\ell+1}$. $\mathcal{T}_0$ consists of only the target node. We then define the last passage time τ_ℓ of the search phase from each subtree $\mathcal{T}_\ell$. An upper bound on the end of this first stage is found by proving that the expected last passage time to each of the halftrees that do not contain the

[2] We note that the design in this paper is based on asymptotic analysis (as the error approaches zero). Therefore, the implementation and analysis do not depend on the prior p_b.

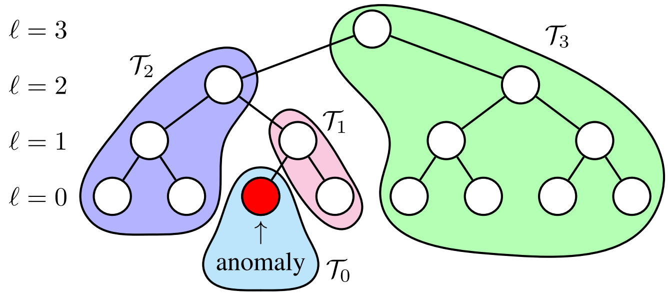


Fig. 2. An illustration of the subtrees $\mathcal{T}_0, \ldots, \mathcal{T}_{\log_2 M}$ used in the analysis of the HDS algorithm.

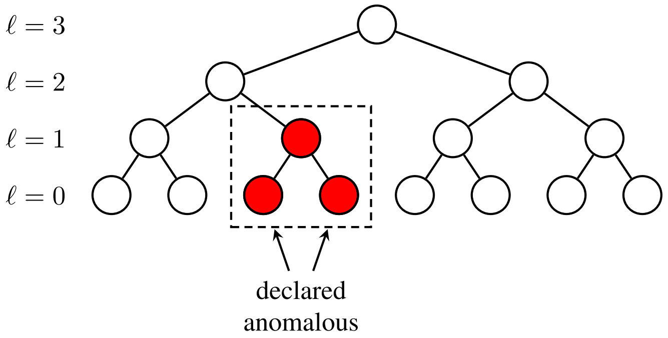


Fig. 3. Multi-target detection illustration. On the third run of the random walk, the nodes in the dashed box are no longer sampled from or visited.

target is bounded by a constant. Summing the upper bound on the last passage times yields the first term in (21).

The second stage is the *leaf target test*, which ends by declaring the target with expected time $\mathbb{E}[\tau_0]$. To bound $\mathbb{E}[\tau_0]$, we first define a random time τ_{ML} to be the smallest integer such that the estimator of the target leaf's parameter equals to $\boldsymbol{\theta}_1^{(0)}$ for all $n > \tau_{ML}$, and we show that $\mathbb{E}[\tau_{ML}]$ is bounded by a constant independent of c and M. We then bound $\mathbb{E}[\tau_0]$ using Wald's equation [38] and Lorden's inequality [58], which yields the second and third terms in (21). Finally, we show that the detection error is of order $O(c)$. By using the martingale properties of the ALLR statistic we prove that the false positive rate of the leaf test is bounded by $\frac{c}{\log_2 M}$. In addition, the expected number of times a benign leaf is tested is in the order of $\log_2 M$. The resulting error rate $\mathrm{P}_{\text{Err}}(\pi_{\text{HDS}})$ is therefore in the order of c (third term in (21)). ■

The optimality properties of the Bayesian risk of HDS in both c and M directly carry through to the sample complexity of HDS, as stated in the following corollary:

Corollary 1: The sample complexity of HDS is upper bounded by:

$$\mathcal{Q}(\pi_{\text{HDS}}) \le B \cdot \log_2 M + \frac{\log \frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\boldsymbol{\theta}_0^{(0)}\right)} + O(1). \tag{22}$$

The sample complexity of any policy π is bounded from below by

$$\mathcal{Q}(\pi) \ge \frac{\log_2 M}{I_{\max}} + \frac{\log \frac{1-c}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\boldsymbol{\theta}_0^{(0)}\right)} + O(1), \tag{23}$$

where $I_{\max}$ is the maximum mutual information between the true hypothesis and the observation under an optimal action.

Proof: The upper bound (22) follows directly from Theorem 1, while (23) is obtained using [15, Thm. 2]. ■

Corollary 1 indicates that HDS is asymptotically optimal in c and order optimal in M.

We point out that the leading constants in the bounds might be tightened for special cases (i.e., if the distributions $f_\ell(\cdot)$ were specified). Nevertheless, the analysis here focuses on establishing asymptotic/order optimality for general observation models. Hence, we resort to bounding techniques that are generally applicable and not restricted to specific distributions.

C. Multi-Target Detection

We next consider the detection of $K > 1$ anomalous processes. Our derivation and analysis is based on the following additional assumptions:

AS2 The number of anomalous processes K is a-priori known.

AS3 The search policy can remove a declared process from the tree, e.g., as in group testing the defective item is no longer tested in subsequent group tests.

AS4 The distinguishability assumption *AS1* is extended such that the distribution of a node that contains multiple anomalies is more similar to a node that contains a single anomaly, than to a benign node. To formulate mathematically, let $\boldsymbol{\Theta}_j^{(\ell)}$ be the set of parameters of a node that contains j anomalies. We require, that there is $\Delta > 0$ such that (3) holds and that for all levels $\ell = 1, \ldots, \log_2 M$, number of anomalies $j = 1, \ldots, \min(K, 2^\ell)$ and multi-anomaly parameter $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j^{(\ell)}$ it holds that

$$\exists \boldsymbol{\theta}_1^{(\ell)} \in \boldsymbol{\Theta}_1^{(\ell)} : \mathcal{D}_\ell\left(\boldsymbol{\theta}_j||\boldsymbol{\theta}_0^{(\ell)}\right) - \mathcal{D}_\ell\left(\boldsymbol{\theta}_j||\boldsymbol{\theta}_1^{(\ell)}\right) \ge \Delta. \tag{24}$$

This assumption holds in a wide variety of scenarios and ensures that there is a bounded number of samples K for the internal test, such that the random walk approaches the closest anomaly with a probability greater than 0.5.

Algorithm Design: Since K is known by *AS2*, HDS formulated in Algorithm 1 can be extended to locate the targets one-by-one. A process is declared anomalous by running the algorithm detailed in Subsection III-A. This operation is feasible by *AS3*. This means that subsequent random walks only visit nodes that contain undeclared processes (Fig. 3). As a result, we only have to sample from one of the children during some internal tests.

For the internal test, we still use the anomalous parameter set $\boldsymbol{\Theta}_1^{(\ell)}$ that represents the distribution for one anomaly within the node. This is justified by Assumption *AS4*. The resulting procedure is summarized as Algorithm 2.

Performance Analysis: The theoretical guarantees derived for a single target in Subsection III-B carry also to the multi-target

Algorithm 2: K Target HDS.

```
Input: Number of targets $K$
for $k = 1, \ldots, K$ do
    Identify $k$th target by invoking Algorithm 1 at level $l = 0$;
    Remove detected anomalous leaf node from tree;
end
```

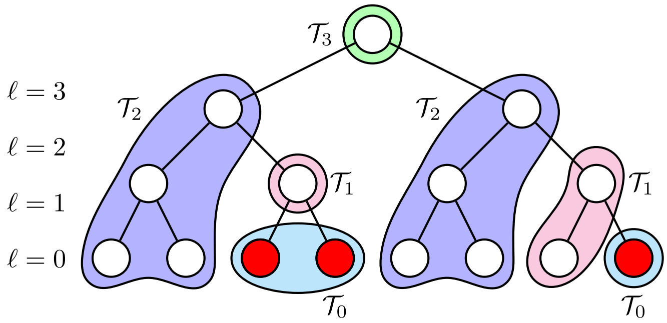


Fig. 4. Illustration of the tree partition $\mathcal{T}_0, \ldots, \mathcal{T}_{\log_2 M}$ used in the analysis of the HDS algorithm for multiple targets.

setting when *AS2–AS4* hold, in addition to *AS1*. This is stated in the following theorem:

Theorem 2: When *AS1–AS4* hold, and $\boldsymbol{\Theta}_1^{(\ell)}$ is finite for all $0 \le \ell \le \log_2 M - 1$, the Bayesian risk of π_{HDS} with K anomalous processes is bounded by:

$$\mathcal{R}(\pi_{\text{HDS}}) \le cKB\log_2 M + \frac{cK\log\frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\boldsymbol{\theta}_0^{(0)}\right)} + O(c)\ , \quad (25)$$

where B is a constant independent of M, c and K.

Proof: The complete proof is given in Appendix C. Here, we only present the proof outline, which extends on the rationale of the proof of Theorem 1: Again, we divide the tree $\mathcal{T}$ into a similar partition $\mathcal{T}_0, \ldots, \mathcal{T}_{\log_2 M}$, where the sets $\mathcal{T}_\ell$ are recursively obtained by removing the halftrees at level ℓ that contain at least one anomaly from $\mathcal{T} \setminus \mathcal{T}_{\ell+1}$ (Fig. 4). Roughly speaking, due to the assumption in (24), the internal test and the leaf test have a greater probability of moving toward the closest anomaly than away from it. This is explicitly shown in (67) in Appendix C, by breaking the likelihood of $\boldsymbol{\theta}_1$ with respect to $\boldsymbol{\theta}_0$ to the likelihood of testing $\boldsymbol{\theta}_j$ with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_j$ with respect to $\boldsymbol{\theta}_0$, and then using (24). This results in the same constant upper bound on the expected last passage times to the sets $\mathcal{T}_1, \ldots, \mathcal{T}_{\log_2 M}$ as in the single-target scenario, which implies that the first term in (25) is the first term of (21) scaled by the number of anomalies K. The leaf test is unaffected by the additional anomalies. Therefore, the sample complexity of a single random walk in the multi-target scenario has the same upper bound as in the single-target scenario resulting again in the sample complexity in the second and third terms being scaled by K. Finally, the upper bound on the probability of the declaring a benign process anomalous remains unaffected too. Applying the union bound over the K random walks yields an error rate in the order of c in the third term. ■

Similarly to risk guarantees, one can also bound the sample complexity of Algorithm 2, as stated in the following:

Corollary 2: The sample complexity of π_{HDS} for the detection of K anomalies under *AS1–AS4* is bounded via

$$\mathcal{Q}(\pi_{\text{HDS}}) \le KB\cdot\log_2 M + \frac{K\log\frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\boldsymbol{\theta}_0^{(0)}\right)} + O(1). \quad (26)$$

Proof: The upper bound (26) follows directly from Theorem 2. ■

Corollary 2 and the lower bound in (23) indicate that HDS is order optimal in M and has an asymptotic ratio of K when c approaches zero.

We note that we can modify the problem formulation and allow for partially correct predictions, i.e., to detect $1 \le \tilde{K} \le K$ anomalies. In this case, we run HDS in the same way, and stop the algorithm after detecting only $\tilde{K}$ anomalous processes. The locations of the remaining $K - \tilde{K}$ anomalous processes are declared arbitrarily. Following similar steps in the analysis, in this case we have $\mathcal{Q}_s \le \tilde{K}B\log_2 M$, $\mathcal{Q}_t \le \frac{\tilde{K}\log\frac{\log_2 M}{c}}{\mathcal{D}_0(\boldsymbol{\theta}_0^{(0)}||\boldsymbol{\theta}_1^{(0)})} + O(1)$, and the error (defined with respect to detecting only $\tilde{K}$ anomalous processes) is again of order $O(c)$. Therefore, the asymptotic optimality properties are preserved.

D. Discussion

The proposed HDS algorithm is designed to efficiently search in hierarchical data structures while coping with an unknown anomaly distribution. It can be viewed as an extension of the IRW method [11] to unknown anomaly parameters, while harnessing the existing knowledge regarding the distribution of the anomaly-free measurements. The uncertainty in the anomaly distribution makes both the algorithm design and the performance analysis much more involved. In contrast to existing hierarchical algorithms, HDS can incorporate general parameterized anomaly observation models, resulting in it being order optimal with respect to the search space size and asymptotically optimal in detection accuracy.

The derivation of HDS motivates the exploration of several extensions. First, HDS is derived for hierarchical data that can be represented as a binary tree, while anomaly search with adaptive granularity may take the form of an arbitrary tree. In such case, the path length from each leaf to the root may be different, and thus the distribution of each node does not depend solely on its level on the tree. We conjecture that with some modifications on the HDS algorithm, optimal performances can be also guaranteed in this case. Another interesting direction worth pursuing in the future is to extend our problem to online change point detection setting with multiple processes.

Furthermore, we design HDS for detecting leaf targets, while in some scenarios one may have to cope with hierarchical targets, i.e., where intermediate nodes can be anomalous. Additional extensions regarding the statistical model of the parameter space are as follows. First, we might consider a composite model for both benign and anomalous distributions. A second extension would be to consider an infinite (uncountable) parameter space, and finally, to consider the heterogeneous setting, where the distribution of a benign process might not be the same for all $M-K$ benign processes, and the distribution of an abnormal process might not be the same for all K abnormal processes. Whether asymptotic optimality can be achieved under these setting remains open. We leave the extension of HDS to these settings for future work.

IV. Numerical Evaluations

We next empirically compare HDS with the existing search strategies of Deterministic Search (DS) [41], IRW [11], and the Confidence Bounds based Random Walk (CBRW) algorithm [12]. The IRW algorithm has access to the true anomaly parameter $\boldsymbol{\theta}_1^{(\ell)}$, while the other algorithms only have access to $\boldsymbol{\Theta}_1^{(\ell)}$. IRW and HDS use fixed size internal tests that are not optimized for the specific simulation. Instead the sample sizes K_ℓ are chosen as small as possible such that the desired drift toward the target is ensured. The performance of IRW should therefore be a best-case scenario for HDS. IRW, DS, and HDS use $c = 10^{-2}$, and CBRW uses $p_0 = 0.2$ and $\epsilon = 10^{-2}$. The values are averaged over 10^6 Monte Carlo runs.[3]

Scenario 1: Exponential Distributions

We first simulate a scenario where the decision-maker observes the inter-occurrence time of Poisson point processes with benign rate $\lambda_0 = 1$ and anomalous rate $\lambda_1 = 10^3$. The rates at the internal nodes are equal to the sum of the rates of their children. The minimum rate that is considered anomalous is $\lambda_{1,\min} = \frac{\lambda_0+\lambda_1}{2}$ such that the anomaly parameter set is $\boldsymbol{\Theta}_1^{(0)} = [\lambda_{1,\min}, \infty)$. Under this setting, $\boldsymbol{\theta}_0^{\ell} = 2^{(\ell)}\lambda_0$, $\boldsymbol{\theta}_1^{(\ell)} = (2^\ell - 1)\lambda_0 + \lambda_1$. The minimum rate that is considered anomalous in level ℓ is $\lambda_{1,\min}^{(\ell)} = \frac{\boldsymbol{\theta}_0^{(\ell)}+\boldsymbol{\theta}_1^{(\ell)}}{2}$ such that the anomaly parameter set is $\boldsymbol{\Theta}_1^{(\ell)} = [\lambda_{1,\min}^{(\ell)}, \infty)$. This scenario models the detection of heavy hitters among Poisson flows where the measurements are exponentially distributed packet inter-arrival times. CBRW uses the mean threshold η_ℓ, such that the generalized likelihood ratio is one at η_ℓ and exact bounds for the mean of exponentially distributed random variables with rate $\lambda_\ell = \frac{1}{\eta_\ell}$.

Fig. 5 depicts the risk $\mathcal{R}(\pi)$ as in (8) versus the number of processes M. We can clearly observe that HDS outperforms CBRW and DS for most values, and it is within a minor gap of that of IRW. While for $M \geq 16$, HDS only slightly outperforms CBRW, it notably outperforms DS. However, it is noted that CBRW uses sequential internal tests, which should be more efficient than the fixed size internal tests of HDS. For this reason, in this scenario we also compare an alternative internal test for HDS. The results of this study, depicted in Fig. 6, show that switching to the sequential GLLR statistic for the leaf test instead of the ALLR statistic yields a performance gain for all M. An even greater jump in performance is achieved by using an active test for the internal nodes. The details of the active test are given in Appendix A.

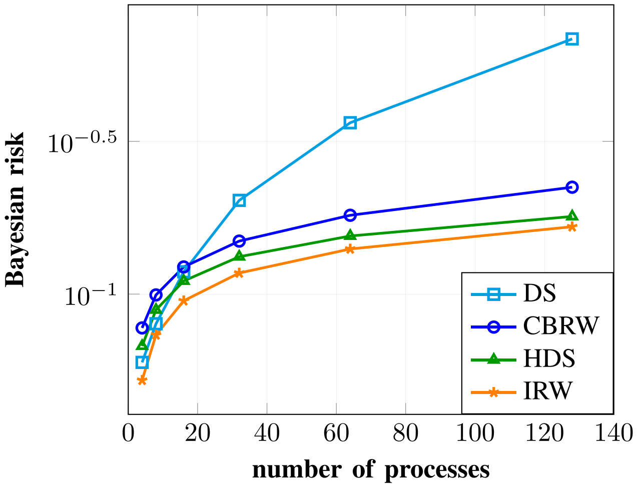


Fig. 5. Risk vs. number of processes, scenario 1.

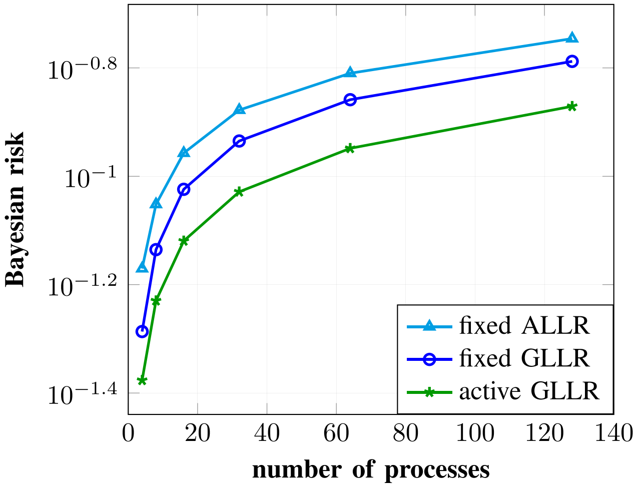


Fig. 6. HDS with different internal tests (fixed sample size vs. active) and leaf test statistics (ALLR vs. GLLR), scenario 1. The active test uses a confidence level of $p = \frac{1}{2} + 10^{-16}$.

Scenario 2: Bernoulli Interference

Next, we simulate our decision making algorithm when considering a set of Poisson point processes with rate $\lambda_0 = 0.1$. Here, the measurements of the nodes that contain the anomaly are corrupted by Bernoulli interference; i.e.,

$$y(i) \sim \text{Exp}(2^\ell \lambda_0) + z \cdot [-6 + (a+6) \cdot \text{Bernoulli}(0.5)]. \quad (27)$$

In (27), $z \in \{0, 1\}$ indicates whether the node is anomalous, and a is unknown. The node parameter $\boldsymbol{\theta}$ is given by the pair (z, a), where $\boldsymbol{\theta}_0^{(\ell)} = (0, 0)$, $\boldsymbol{\theta}_1^{(\ell)} = (1, 10)$, and $\boldsymbol{\Theta}_1^{(\ell)} = \{1\} \times \{1, 5, 10\}$ for all levels $0 \leq \ell \leq \log_2 M$. CBRW uses $\eta_\ell = 1$ and sub-Gaussian bounds with $\xi = 0.05$.

In this case the mean values of the benign and abnormal distribution are close to each other, and the anomalous process is reflected by higher moments of the distributions. The results for this setting, depicted in Fig. 7, show that while CBRW achieves poor performance, HDS detects the anomaly efficiently,

[3] The source code can be found in https://github.com/DrummingBeb/Composite-Anomaly-Detection-via-Hierarchical-Dynamic-Search.

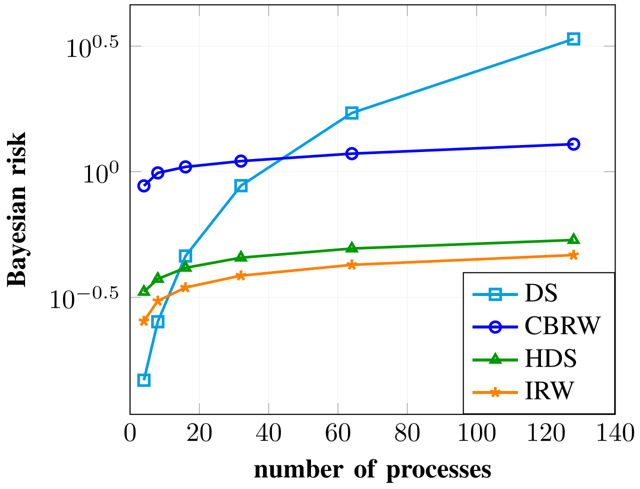


Fig. 7. Risk vs. number of processes, scenario 2.

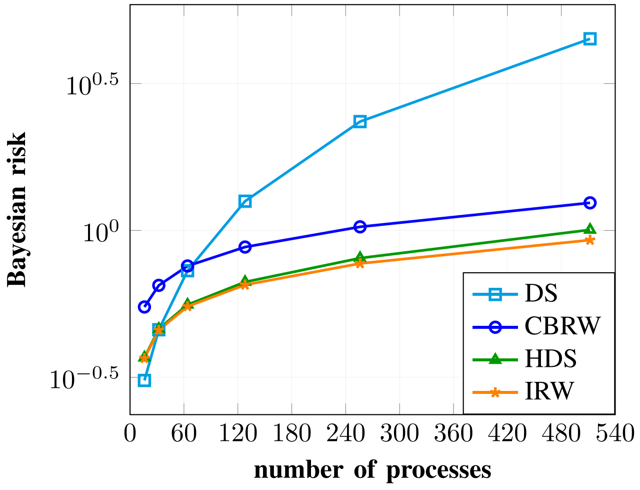


Fig. 8. Bayesian risk vs. number of processes, scenario 3.

resulting in a larger gap between HDS and CBRW than in the first scenario.

Scenario 3: Multiple Anomalies

Here we extend scenario 1 to $K = 5$ anomalies. In this case, $\boldsymbol{\theta}_0^{(\ell)} = 2^\ell \lambda_0, \boldsymbol{\theta}_j^{(\ell)} = (2^\ell - j)\lambda_0 + j \cdot \lambda_1$ for $1 \leq j \leq 5$. Following *AS4*, in our internal tests we consider the parameter set in level ℓ to be $\boldsymbol{\Theta}_1^{(\ell)}$ (i.e., the set that contains 1 anomaly and not j anomalies). HDS and IRW use active internal tests. Additionally HDS uses the GLLR statistic for the leaf tests. Fig. 8 shows a very similar picture as Fig. 5, in which HDS performs close to IRW and better than CBRW and DS. However, the performance HDS surpasses the non-hierarchical DS at $M = 30$ processes as opposed to after already $M = 10$ processes in scenario 1.

Scenario 4: Denial of Service Detection

In this scenario, we detect DoS attacks using the DARPA intrusion detection data set [59]. Every entry in the data set corresponds to a packet arriving at an interface. We only consider the timestamp, packet size and label (either benign or DoS traffic) of each packet. The anomalous process ($K = 1$) corresponds to an interface that receives DoS traffic, so we simulate with permutations the entire data set. The benign processes are simulated by permutations of the packets that are labeled as benign traffic.

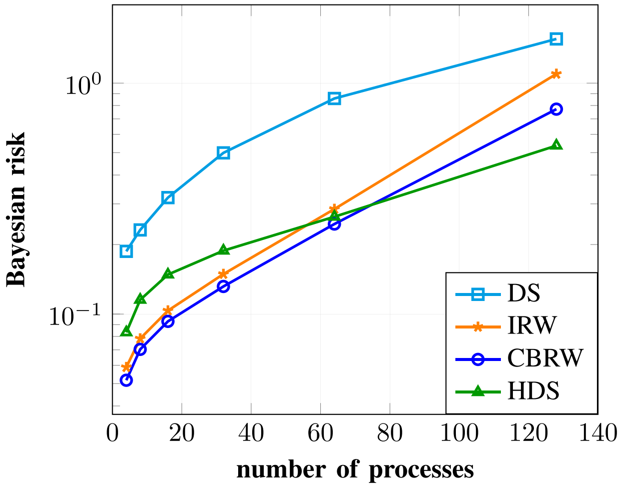


Fig. 9. Bayesian risk vs. number of processes, scenario 4.

We use the model in [41] that considered a sample entropy for packet-size modeling, and demonstrated strong performance in detecting anomalous data on the DARPA data set. Every 100 ms seconds a sample is drawn by calculating the sample entropy of the packet sizes observed in the probed node during the current 100 ms interval. Sampling from an internal node is naturally done by aggregating the packets of the processes within the node. The sample entropy is modeled with a Gaussian distribution that is parametrized by its mean and standard deviation. Using 1000 permutations of the training split (50% of the data), the distribution of the sample entropy is estimated for benign and anomalous nodes at all levels i.e. $\boldsymbol{\theta}_0^{(\ell)} = \left(\mu_0^{(\ell)}, \sigma_0^{(\ell)}\right)$ and $\boldsymbol{\theta}_1^{(\ell)} = \left(\mu_1^{(\ell)}, \sigma_1^{(\ell)}\right)$ are estimated respectively for $\ell = 0, \ldots, \log_2 M - 1$. The anomalous sample entropy is expected to have a smaller mean and variance i.e. $\mu_1^{(\ell)} < \mu_0^{(\ell)}$ and $\sigma_1^{(\ell)} < \sigma_0^{(\ell)}$. For DS and HDS, the anomaly parameter sets are $\boldsymbol{\Theta}_1^{(\ell)} = \left(-\infty, \mu_{\frac{1}{2}}^{(\ell)}\right] \times \left(0, \sigma_{\frac{1}{2}}^{(\ell)}\right]$ where $\mu_{\frac{1}{2}}^{(\ell)} = \frac{\mu_0^{(\ell)} + \mu_1^{(\ell)}}{2}$ and $\sigma_{\frac{1}{2}}^{(\ell)} = \frac{\sigma_0^{(\ell)} + \sigma_1^{(\ell)}}{2}$. HDS and IRW use active internal tests, and HDS uses the sequential GLLR for the leaf tests. CBRW uses thresholds $\eta_\ell = \frac{\mu_0^{(\ell)} + \mu_{\frac{1}{2}}^{(\ell)}}{2}$ and exact confidence intervals for the mean of normally distributed random variables with standard deviation $\sigma^{(\ell)} = \frac{\sigma_0^{(\ell)} + \sigma_{\frac{1}{2}}^{(\ell)}}{2}$. Due to instability of DS, we discarded runs with more than 1000 samples. Therefore, the evaluation of DS is very generous.

Fig. 9 shows the risk as a function of the number processes. Interestingly, HDS scales better with the size of the search space when compared to the other hierarchical algorithms, namely IRW and CBRW. We attribute this to the fact that the estimates can be inaccurate at high levels despite using a large training split and many permutations. IRW loses performance because it relies on the point estimate $\boldsymbol{\theta}_1^{(\ell)}$ while the composite anomaly model of HDS is more robust.

V. Conclusion

In this work we developed a sequential search strategy for the composite hierarchical anomaly detection problem dubbed HDS. HDS uses two variations of the GLLR statistic to ensure a biased random walk for a quick and accurate detection of the anomaly process. HDS is shown to be order optimal with respect to the size of the search space and asymptotically optimal with respect to the detection accuracy. The addition of the hierarchical search significantly improves the performance over linear search methods in the common case of a large number of processes and heavy hitting anomalies. We empirically show that the performance can be further improved by using different statistics and local tests, and that for real-world data the composite anomaly model of HDS is more robust to inaccurate estimates from training than existing algorithms that assume a known anomalous distribution model.

Appendix A
Active Internal Test

Instead of the fixed size internal test described in Section III-A, we can use an active internal test:

Let $S_L(t)$ and $S_R(t)$ be the GLLR of the left and right children respectively at time t and initialize them with zero at $t=0$. As in the IRW active test [11], we define the thresholds

$$v_0 \triangleq -\log\frac{2p}{1-p}, \quad v_1 \triangleq \log\frac{2p}{1-p} \tag{28}$$

where $p > \frac{1}{2}$ is the confidence level. Let child

$$x(t-1) = \arg\max_{i\in\{\mathrm{L,R}\}} S_i(t-1) \tag{29}$$

be the child with the higher GLLR at time $t-1$. Then, in every step t, we draw a sample from child $x(t-1)$ and update $S_{x(t)}(t)$. The other child $\tilde{x}(t) \neq x(t)$ keeps the previous GLLR i.e., $S_{\tilde{x}(t)}(t) = S_{\tilde{x}(t)}(t-1)$. The test terminates at the random time

$$k = \inf\left\{t\in\mathbb{N} \,|\, S_{x(t)}(t) \le v_0 \text{ or } S_{x(t)}(t) \ge v_1\right\}. \tag{30}$$

If $S_{x(k)}(k) \ge v_1$, the random walk zooms into child $x(k)$ and if $S_{x(k)}(k) \le v_0$, the random walk zooms out to the parent.

We observe a significant gain in empirical performance when compared to the fixed sample internal test (Fig. 6).

Appendix B
Proof of Theorem 1

To find an upper bound on the Bayesian risk of HDS, we analyze the case where it is *implemented indefinitely*, meaning that HDS probes the processes indefinitely according to its selection rule, while the stopping rule is disregarded. We divide the trajectory of indefinite HDS into discrete steps at times $t\in\mathbb{N}$. A step is not necessarily associated with every sample as will become clear later. Let τ_∞ mark the first time that indefinite HDS performs a leaf test on the true anomaly and $\tilde{S}_{\mathrm{ALLR}}$ rises above the threshold. It is easy to see that regular HDS terminates no later than τ_∞. We divide the initial trajectory $t = 1, 2, \ldots, \tau_\infty$ of the indefinite random walk into two stages:

- In the *search stage* the random walk explores the high level nodes and eventually concentrates at the true anomaly. This stage ends at time τ_s which is the last time a leaf test is started on the true anomaly before τ_∞.
- The second stage is the *target test* which ends with the declaration of the target. The duration of this stage is τ_0.

Step 1: Bound the sample complexity of the search stage:

We partition the tree $\mathcal{T}$ into a sequence of sub-trees $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_{\log_2 M}$ (Fig. 2) and define the last passage time τ_ℓ as described in Section III-B. Let $G(t)$ indicate the sub-tree of the node tested at time t. The last passage time to $\mathcal{T}_{\log_2 M}$ is

$$\tau_{\log_2 M} = \sup\left\{t\in\mathbb{N} : G(t) = \mathcal{T}_{\log_2 M}\right\} \tag{31}$$

For the smaller sub-trees $\mathcal{T}_1, \ldots, \mathcal{T}_{\log_2 M-1}$ the last passage times are defined recursively such that

$$\tau_i = \sup\{t\in\mathbb{N} : G(t) = \mathcal{T}_i\} - \tau_{i+1}. \tag{32}$$

Notice, that the search time is bounded by

$$\tau_s = \sup_{1\le\ell\le\log_2 M} \tau_\ell \le \sum_{\ell=1}^{\log_2 M} \tau_\ell. \tag{33}$$

Next, we bound the expected last passage times $\mathbb{E}[\tau_\ell]$ for $1 \le \ell \le \log_2 M$. Towards this end, we define a distance D_t from the state of the indefinite random walk at time t to the anomalous leaf. When an internal node is probed, D_t is equal to the discrete distance to the anomaly on the tree. Since the walk starts at the root, we have $D_0 = \log_2 M$. when testing a benign leaf, D_t is equal to the sum of the discrete distance on the tree and the accumulated $\tilde{S}_{\mathrm{ALLR}}$ of the current leaf test. When the true anomaly is probed, the distance is negative i.e. $D_t = -\tilde{S}_{\mathrm{ALLR}}$. Let the step W_t be the random change in the distance at time t such that $D_{t+1} = D_t + W_t$. Internal tests comprise only a single step either towards or away from the anomaly, i.e., $W_t \in \{-1, 1\}$. Because the sample sizes K_ℓ of the internal tests are constructed such that $\mathcal{P}(W_t = 1) < \frac{1}{2}$, we have

$$\mathbb{E}[W_t] = 2\mathcal{P}[W_t = 1] - 1 < 0. \tag{34}$$

We now show that if the sets of anomalous parameters $\boldsymbol{\Theta}_1^{(\ell)}$ are finite, there exists a bounded number of samples K_ℓ such that (34) holds for the internal test at all levels. We identify the two events

$$\mathrm{E}_0 = \text{ the tested node does not contain the anomaly} \tag{35}$$

$$\mathrm{E}_1 = \text{ the tested node contains the anomaly.} \tag{36}$$

The probability of making a step in the wrong direction with an internal test is upper bounded by

$$\mathcal{P}[W_t = 1] \le \max\left(\mathcal{P}[W_t = 1 \,|\, \mathrm{E}_0], \mathcal{P}[W_t = 1 \,|\, \mathrm{E}_1]\right). \tag{37}$$

We first bound the first term in the maximization of (37). Let $\mathcal{P}_{\boldsymbol{\theta}_i}$ be the probability measure when the true state of nature is $\boldsymbol{\theta}_i, i = 0, 1$, and let $\mathbb{E}_{\boldsymbol{\theta}_i}$ be the operator of expectation with respect to the measure $\mathcal{P}_{\boldsymbol{\theta}_i}$. Let $S_{\boldsymbol{\theta}_0}$ and $S_{\boldsymbol{\theta}_1}$ be the random GLLRs based on K samples from a benign node and an anomalous node respectively, where we omit the level ℓ for readability. Then, under $\mathbb{E}_{\boldsymbol{\theta}_0}$ an error implies that at least one of the GLLRs is strictly positive. By applying the union bound we get

$$\mathcal{P}[W_t = 1 \,|\, \mathrm{E}_0] \le 2\mathcal{P}[S_{\boldsymbol{\theta}_0} > 0]. \tag{38}$$

Let $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \prod_{i=1}^{K} f(y(i)\,|\,\boldsymbol{\theta})$ be the maximum likelihood estimate (MLE) in the set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_0\} \cup \boldsymbol{\Theta}_1$. The event that $S_{\boldsymbol{\theta}_0}$ is strictly positive implies that $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$ via the definition of the MLE. Therefore, we find that

$$\mathcal{P}[S_{\boldsymbol{\theta}_0} > 0] = \sum_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_1} \mathcal{P}_{\boldsymbol{\theta}_0}\left[\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_1\right]. \tag{39}$$

Applying the definition of the MLE, the Chernoff bound and the independent and identically distributed (i.i.d.) property yields

$$\begin{aligned}\mathcal{P}_{\boldsymbol{\theta}_0}\left[\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_1\right] &\le \mathcal{P}_{\boldsymbol{\theta}_0}\left[\sum_{i=1}^{K} \log\frac{f(y(i)\,|\,\boldsymbol{\theta}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)} \ge 0\right]\\ &\le \left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\exp\left(-s\log\frac{f(y(i)\,|\,\boldsymbol{\theta}_0)}{f(y(i)\,|\,\boldsymbol{\theta}_1)}\right)\right]\right)^{K}\end{aligned} \tag{40}$$

for all $s \ge 0$. Notice, that the derivative of the expectation on the RHS of (40) with respect to s, $-\mathcal{D}(\boldsymbol{\theta}_0||\boldsymbol{\theta}_1) \le -\Delta < 0$, is strictly negative for all $\boldsymbol{\theta}_1$ due to the assumption in (3). Thus, for all $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$ there exists a $s > 0$ such that the RHS of (40) decays exponentially meaning that there exist a bounded $C > 0$ and a $\gamma > 0$ such that

$$\mathcal{P}_{\boldsymbol{\theta}_0}\left[\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_1\right] \le Ce^{-\gamma K}. \tag{41}$$

Combining (38), (39) and (41), we find that $\mathcal{P}[W_t = 1|\mathrm{E}_0]$ decays exponentially with the number of samples K.

Next, we show that $\mathcal{P}[W_t = 1|\mathrm{E}_1]$ also decays exponentially. Under E_1, the event that the GLLR of the anomalous child is strictly positive and the GLLR of the benign child is negative implies, that we move towards the anomaly, resulting in

$$\begin{aligned}&\mathcal{P}[W_t = 1\,|\,\mathrm{E}_1] = 1 - \mathcal{P}[W_t = -1\,|\,\mathrm{E}_1]\\ &\quad\le 1-\mathcal{P}[S_{\boldsymbol{\theta}_1} > 0]\cdot\mathcal{P}[S_{\boldsymbol{\theta}_0} \le 0] \le \mathcal{P}[S_{\boldsymbol{\theta}_1} \le 0] + \mathcal{P}[S_{\boldsymbol{\theta}_0} > 0].\end{aligned}$$

We already showed that $\mathcal{P}[S_{\boldsymbol{\theta}_0} > 0]$ decays exponentially with K, it remains to show the same for $\mathcal{P}[S_{\boldsymbol{\theta}_1} \le 0]$. Using the definition of the MLE, the Chernoff bound and the i.i.d. property find

$$\begin{aligned}\mathcal{P}[S_{\boldsymbol{\theta}_1} \le 0] &\le \mathcal{P}_{\boldsymbol{\theta}_1}\left[\sum_{i=1}^{K} \log\frac{f(y(i)\,|\,\hat{\boldsymbol{\theta}}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)} \le 0\right]\\ &\le \mathcal{P}_{\boldsymbol{\theta}_1}\left[\sum_{i=1}^{K} \log\frac{f(y(i)\,|\,\boldsymbol{\theta}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)} \le 0\right]\\ &\le \left(\mathbb{E}_{\boldsymbol{\theta}_1}\left[\exp\left(-s\log\frac{f(y(i)\,|\,\boldsymbol{\theta}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)}\right)\right]\right)^{K}.\end{aligned} \tag{42}$$

for all $s \ge 0$. Once again, the derivative of the expectation on the RHS of (42) with respect to s, $-\mathcal{D}(\boldsymbol{\theta}_1||\boldsymbol{\theta}_0) \le -\Delta < 0$, is strictly negative for all $\boldsymbol{\theta}_1$ due to the assumption in (3). It follows that $\mathcal{P}[W_t = 1\,|\,\mathrm{E}_1]$ decays exponentially with the number of samples K. Thus, there exists a bounded K such that (34) holds.

On leaf nodes, every single sample of the sequential test comprises a step. A step is therefore the change in $\tilde{S}_{\mathrm{ALLR}}$. Using the assumption in (3) and the independence of $\hat{\boldsymbol{\theta}}_1^{(0)}(i-1)$ and $y(i)$ we find that for *benign* leafs

$$\mathbb{E}[W_t] = \mathbb{E}_{\boldsymbol{\theta}_0^{(0)}}\left[\log\frac{f_0\left(y(t)\,|\hat{\boldsymbol{\theta}}_1^{(0)}(t-1)\right)}{f_0\left(y(t)\,|\,\boldsymbol{\theta}_0^{(0)}\right)}\right] \le -\Delta < 0. \tag{43}$$

Similarly, we want to show that for the *anomalous* leaf that

$$\mathbb{E}[W_t] = \mathbb{E}_{\boldsymbol{\theta}_1^{(0)}}\left[-\log\frac{f_0\left(y(t)\,|\hat{\boldsymbol{\theta}}_1^{(0)}(t-1)\right)}{f_0\left(y(t)\,|\,\boldsymbol{\theta}_0^{(0)}\right)}\right] < 0. \tag{44}$$

Denoting $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_1^{(0)}(t-1)$, we split the term and use the law of total expectation to find that

$$\begin{aligned}\mathbb{E}[W_t] &= \mathbb{E}_{\boldsymbol{\theta}_1^{(0)}}\left[-\log\frac{f_0\left(y(t)\,|\hat{\boldsymbol{\theta}}\right)}{f_0\left(y(t)\,|\,\boldsymbol{\theta}_0^{(0)}\right)} + \underbrace{\log\frac{f_0\left(y(t)\,|\,\boldsymbol{\theta}_1^{(0)}\right)}{f_0\left(y(t)\,|\,\boldsymbol{\theta}_1^{(0)}\right)}}_{=0}\right]\\ &= -\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}\,||\,\boldsymbol{\theta}_0^{(0)}\right) + \mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}\left[\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_1^{(0)}\right]\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}\,||\hat{\boldsymbol{\theta}}\right)\end{aligned} \tag{45}$$

where we used the fact that $\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}\,||\,\boldsymbol{\theta}_1^{(0)}\right) = 0$. For (44) to hold, it remains to be shown that

$$\mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}\left[\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_1^{(0)}\right] < \inf_{\hat{\boldsymbol{\theta}}\in\boldsymbol{\Theta}_1^0} \frac{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}\,||\,\boldsymbol{\theta}_0^{(0)}\right)}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}\,||\hat{\boldsymbol{\theta}}\right)} \triangleq \lambda_{\boldsymbol{\theta}_1^{(0)}}. \tag{46}$$

Notice, that the $\lambda_{\boldsymbol{\theta}_1^{(0)}}$ are strictly positive due to the assumption in (3) and assuming that $\sup_{\boldsymbol{\theta}_1^{(0)},\hat{\boldsymbol{\theta}}\in\boldsymbol{\Theta}_1^0} \mathcal{D}_0(\boldsymbol{\theta}_1^{(0)}\,||\hat{\boldsymbol{\theta}}) < \infty$. For this purpose, we first introduce the following Lemma:

Lemma 1: Let $\boldsymbol{\Theta}_1^{(0)}$ be finite, i.e., $R = |\boldsymbol{\Theta}_1^{(0)}| < \infty$ and let $\hat{\boldsymbol{\theta}}_1^{(0)}(n)$ be the ML estimate of $\boldsymbol{\theta}_1^{(0)}$ using n samples. Let τ_{ML} be the smallest integer such that $\hat{\boldsymbol{\theta}}_1^{(0)}(n) = \boldsymbol{\theta}_1^{(0)}$ for all $n > \tau_{ML}$. Then, there exist a bounded $C > 0$ and a $\gamma > 0$ independent of M and c such that

$$\mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}[\tau_{ML} > n] \le Ce^{-\gamma n}. \tag{47}$$

Proof: The event $\tau_{ML} > n$ implies that there exists a time $t > n$ such that $\hat{\boldsymbol{\theta}}_1^{(0)}(t) \neq \boldsymbol{\theta}_1^{(0)}$ and therefore we have

$$\mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}[\tau_{ML} > n] \le \sum_{t=n}^{\infty} \mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}\left[\hat{\boldsymbol{\theta}}_1^{(0)}(t) \neq \boldsymbol{\theta}_1^{(0)}\right]. \tag{48}$$

By definition of the maximum likelihood estimate, the event $\hat{\boldsymbol{\theta}}_1^{(0)}(t) \neq \boldsymbol{\theta}_1^{(0)}$ implies $\sum_{i=1}^{t} S_{\tilde{\boldsymbol{\theta}}}(i) \ge 0$ for some $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_1^{(0)}$, where $S_{\tilde{\boldsymbol{\theta}}}(i) = \log\frac{f(y(i)|\tilde{\boldsymbol{\theta}})}{f(y(i)|\boldsymbol{\theta}_1^{(0)})}$. Applying the Chernoff bound and using the i.i.d. property yields

$$\mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}\left[\sum_{i=1}^{t} S_{\tilde{\boldsymbol{\theta}}}(i) \geq 0\right] \leq \left(\mathbb{E}_{\boldsymbol{\theta}_1^{(0)}}\left[e^{sS_{\tilde{\boldsymbol{\theta}}}(i)}\right]\right)^t \tag{49}$$

for all $s \geq 0$. The moment generating function (MGF) $e^{sS_{\tilde{\boldsymbol{\theta}}}(i)}$ is equal to one at $s = 0$. The derivative of the MGF at $s = 0$ is $\mathbb{E}_{\boldsymbol{\theta}_1^{(0)}}[S_{\tilde{\boldsymbol{\theta}}}(i)] = -\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\tilde{\boldsymbol{\theta}}\right) < 0$. Because the derivative is negative and assuming that the distribution of $S_{\tilde{\boldsymbol{\theta}}}(i)$ is light-tailed[4], there exist $s > 0$ and $\gamma > 0$ such that $\mathbb{E}[e^{sS_{\tilde{\boldsymbol{\theta}}}(i)}] = e^{-\gamma} < 1$ and the RHS of (49) decays exponentially with t. Summing over all $\tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_1^{(0)}$, we get $\mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}[\hat{\boldsymbol{\theta}}_1^{(0)}(t) \neq \boldsymbol{\theta}_1^{(0)}] \leq Re^{-\gamma t}$, and thus the RHS of (48) is bounded by $\sum_{t=n}^{\infty} Re^{-\gamma t} = \frac{R}{1-e^{-\gamma}} e^{-\gamma n}$. ■

In light of lemma 1, we propose the following mechanism to ensure that (46) holds: Whenever a leaf test is started, before beginning with the sequential test described in Section III-A, a fixed number $N_{\text{leaf}} \geq 0$ of samples $\{y_i\}_{i=-N_{\text{leaf}}+1}^{0}$ is drawn from the leaf to initialize the estimate $\hat{\boldsymbol{\theta}}_1^{(0)}$, meaning, instead of (16) we write

$$\hat{\boldsymbol{\theta}}_1^{(0)}(i-1) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1^{(0)}}{\arg\max} \prod_{j=-N_{\text{leaf}}+1}^{i-1} f_0(y(j)\,|\,\boldsymbol{\theta}). \tag{50}$$

This has the effect, that at every step of the subsequent sequential test, the estimate $\hat{\boldsymbol{\theta}}_1^{(0)}$ is based on at least N_{leaf} samples. Since $\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_1^{(0)}$ implies that $\tau_{ML} > N_{\text{leaf}}$, we have

$$\mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}\left[\hat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_1^{(0)}\right] \leq \mathcal{P}_{\boldsymbol{\theta}_1^{(0)}}[\tau_{ML} > N_{\text{leaf}}]. \tag{51}$$

Using $\lambda = \inf_{\boldsymbol{\theta}_1^{(0)} \in \boldsymbol{\Theta}_1^{(0)}} \lambda_{\boldsymbol{\theta}_1^{(0)}}$ and lemma 1 we find that (46) is satisfied if $N_{\text{leaf}} > -\frac{\log \frac{\lambda}{C}}{\gamma}$. Notice, that N_{leaf} is chosen independent of the size of search space M and the cost c.

With (34), (43) and (44) we established that HDS has the same drift behavior as IRW. Furthermore, we assume that the distribution of $\log \frac{f_0(y(i)\,|\,\tilde{\boldsymbol{\theta}})}{f_0(y(i)\,|\,\boldsymbol{\theta}_0^{(0)})}$ is light-tailed for all $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}_1^{(0)}$.

Thus, we can apply [11, Lemma 1, 2] and find that the expected last passage times $\mathbb{E}[\tau_i]$ for $1 \leq i \leq \log_2 M$ are bounded by a constant β independent of M and c. Applying (33) yields

$$\mathbb{E}[\tau_s] \leq \beta \log_2 M. \tag{52}$$

Let $K_{\max} = \sup_{0 \leq \ell \leq \log_2 M - 1}\{K_\ell\}$ be the maximum number of samples taken from a child during an internal test. Then every step W_t takes at most $N_{\max} = \max\{2K_{\max}, N_{\text{leaf}} + 1\}$ samples and the complexity of the search stage $\mathcal{Q}_s$ is bounded by

$$\mathcal{Q}_s \leq N_{\max}\mathbb{E}[\tau_s] \leq B \log_2 M \tag{53}$$

where $B = \beta N_{\max}$ is a constant independent of M and c.

Step 2: Bound the sample complexity of the target test:

In the analysis of the target test we associate a time step $n = 1, 2, \ldots, \tau_0$ with every sample. Using lemma 1 and the tail sum for expectation we find

$$\mathbb{E}[\tau_{ML}] = O(1). \tag{54}$$

At all times $n > \tau_{ML}$, we necessarily have $\hat{\boldsymbol{\theta}}_1^{(0)} = \boldsymbol{\theta}_1^{(0)}$. From the definition of $\tilde{S}_{\text{LALLR}}$ in (15) it is easy to see, that after $n = \tau_{ML} + 1$, the leaf test is essentially a sequential likelihood ratio test. The expected time until the threshold $\log \frac{\log_2 M}{c}$ is reached $\tau_f = \tau_0 - \tau_{ML}$ is bounded by

$$\mathbb{E}[\tau_f] \leq \frac{\log \frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\boldsymbol{\theta}_0^{(0)}\right)} + O(1) \tag{55}$$

where we used Wald's equation [38] and Lorden's inequality [58] and assumed that the first two moments of the log-likelihood ratio are finite. Combining (54) and (55) yields the sample complexity of the target test

$$\mathcal{Q}_t = \mathbb{E}[\tau_0] \leq \frac{\log \frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)}||\boldsymbol{\theta}_0^{(0)}\right)} + O(1). \tag{56}$$

Step 3: Bound the error rate:

Notice, that detection errors can only occur in the search stage. The expected number of times a benign leaf is tested $\mathbb{E}[N]$ is bounded by the number of steps in the search stage. Thus, using (52) we get

$$\mathbb{E}[N] \leq \mathbb{E}[\tau_s] \leq \beta \log_2 M. \tag{57}$$

Let $Z(n) = \exp(\tilde{S}_{\text{ALLR}}(n))$ be adaptive likelihood ratio at time n. In the following, we use the properties of the ALLR to bound the false positive rate of the leaf test

$$\alpha = \mathcal{P}_{\boldsymbol{\theta}_0^{(0)}}\left[Z(n) \geq \frac{\log_2 M}{c} \text{ for some } n \geq 1\right]. \tag{58}$$

Note that on benign leafs $Z(n)$ is a non-negative martingale, i.e.,

$$\mathbb{E}_{\boldsymbol{\theta}_0^{(0)}}[Z(n+1)\,|\,\{y(i)\}_{i=1}^{n}] \tag{59}$$

$$= Z(n)\mathbb{E}_{\boldsymbol{\theta}_0^{(0)}}\left[\frac{f\left(y(n+1)\,|\hat{\boldsymbol{\theta}}_1^{(0)}(n)\right)}{f\left(y(n+1)\,|\,\boldsymbol{\theta}_0^{(0)}\right)}\right] = Z(n) \tag{60}$$

where we used the independence of $\hat{\boldsymbol{\theta}}_1^{(0)}(n)$ and $y(n+1)$ in the last step. Using a lemma for nonnegative supermartingales [61] we find

$$\mathcal{P}_{\boldsymbol{\theta}_0^{(0)}}\left[Z(n) \geq \frac{\log_2 M}{c} \text{ for some } n \geq 1\right] \leq \frac{c}{\log_2 M}\mathbb{E}_{\boldsymbol{\theta}_0^{(0)}}[Z(1)].$$

Since $Z(1) = \mathbb{E}_{\boldsymbol{\theta}_0^{(0)}}\left[\frac{f(y(1)\,|\hat{\boldsymbol{\theta}}_1^{(0)}(0))}{f(y(1)\,|\,\boldsymbol{\theta}_0^{(0)})}\right] = 1$, the false positive rate is bounded by

$$\alpha \leq \frac{c}{\log_2 M}. \tag{61}$$

[4] A distribution with density f is light-tailed if $\int_{-\infty}^{\infty} e^{\lambda x} f(x)dx < \infty$ for some $\lambda > 0$ [60].

Finally, combining (57) and (61) yields the bound on the error rate

$$\mathrm{P}_{\mathrm{Err}}(\pi_{\mathrm{HDS}}) \leq \alpha \cdot \mathbb{E}[N] \leq \beta c = O(c) \tag{62}$$

Theorem 1 follows from (53), (56) and (62).

Appendix C
Proof of Theorem 2

To find an upper bound on the Bayesian risk of HDS in the multi-target scenario, we analyze the K random walks separately. This can be done because there is at least one undeclared anomalous leaf in the tree $\mathcal{T}$ during each random walk.

Step 1: Bound the sample complexity of the search stage:

Similar to the proof in Appendix B1, we divide the tree $\mathcal{T}$ as described in Section III-C and Fig. 4. The last passage times are defined recursively by (31)–(32) and the search time is bounded by (33). Let $D_t^{(i)}$ be the distance to the i-th anomalous leaf at time t, where the distance is defined as in Appendix B. Now consider the change in the distance to the *closest* anomaly $W_t = D_{t+1} - D_t$ where $D_t = \min_i D_t^{(i)}$. We want to show that in expectation the minimum distance decreases at all times during the random walk i.e.

$$\mathbb{E}[W_t] < 0. \tag{63}$$

As the leaf test is unaffected by additional anomalies and the currently tested leaf is also the closest, it only remains to show that (63) holds for the internal test. Recall, that the number of samples K_ℓ of an internal test is chosen such that (63) holds. In Appendix B, we have proven that such a K_ℓ exists for the two events E_0 and E_1 defined in (35)–(36). Notice, that under E_0 the closest anomaly lies outside the tested node and the distance to it is in expectation reduced by moving to the parent by following the same argument as for a single anomaly. Now, we recognize the events

$$\mathrm{E}_j = \text{ the tested node contains } j \text{ anomalies} \tag{64}$$

for $j \geq 1$. Notice, that the j anomalies within the node are the closest anomalies and they are equally close. Moving to a child that contains at least one anomaly reduces D_t by 1. We distinguish the two events

$$\mathrm{E}_j^{(1)} = \text{ one of the children contains anomalies} \tag{65}$$

$$\mathrm{E}_j^{(2)} = \text{ both of the children contains anomalies.} \tag{66}$$

Let $S_{\boldsymbol{\theta}_j}$ be the random GLLRs based on K' samples from a node containing j anomalies, where we omit the level ℓ for readability. Then under $\mathrm{E}_j^{(1)}$, the event that the GLLR of the anomalous child is strictly positive and the GLLR of the benign child is negative, implies $W_t = -1$ such that

$$\begin{aligned}\mathcal{P}\left[W_t = 1 \,|\, \mathrm{E}_j^{(1)}\right] &= 1 - \mathcal{P}\left[W_t = -1 \,|\, \mathrm{E}_j^{(1)}\right] \\ &\leq 1 - \mathcal{P}\left[S_{\boldsymbol{\theta}_j} > 0\right] \cdot \mathcal{P}[S_{\boldsymbol{\theta}_0} \leq 0] \leq \mathcal{P}\left[S_{\boldsymbol{\theta}_j} \leq 0\right] + \mathcal{P}[S_{\boldsymbol{\theta}_0} > 0].\end{aligned}$$

We already showed that $\mathcal{P}[S_{\boldsymbol{\theta}_0} > 0]$ and $\mathcal{P}[S_{\boldsymbol{\theta}_1} \leq 0]$ decay exponentially with K' (Appendix B1), it remains to show the same for $\mathcal{P}[S_{\boldsymbol{\theta}_j} \leq 0]$ with $j > 1$. Using the definition of the MLE, the Chernoff bound and the i.i.d. property find

$$\begin{aligned}\mathcal{P}\left[S_{\boldsymbol{\theta}_j} \leq 0\right] &\leq \mathcal{P}_{\boldsymbol{\theta}_j}\left[\sum_{i=1}^{K'} \log \frac{f(y(i)\,|\hat{\boldsymbol{\theta}}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)} \leq 0\right] \\ &\leq \mathcal{P}_{\boldsymbol{\theta}_j}\left[\sum_{i=1}^{K'} \log \frac{f(y(i)\,|\,\boldsymbol{\theta}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)} \leq 0\right] \\ &\leq \left(\mathbb{E}_{\boldsymbol{\theta}_j}\left[\exp\left(-s \log \frac{f(y(i)\,|\,\boldsymbol{\theta}_1)}{f(y(i)\,|\,\boldsymbol{\theta}_0)}\right)\right]\right)^{K'}.\end{aligned} \tag{67}$$

for all $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$ and $s \geq 0$. Due to the assumption in (24), for all $\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j$ there exists a $\boldsymbol{\theta}_1$ such that the derivative of the expectation on the RHS of (67) with respect to s

$$\mathcal{D}_\ell(\boldsymbol{\theta}_j || \boldsymbol{\theta}_1) - \mathcal{D}_\ell(\boldsymbol{\theta}_j || \boldsymbol{\theta}_0) \leq -\Delta < 0. \tag{68}$$

is strictly negative. Therefore $\mathcal{P}[S_{\boldsymbol{\theta}_j} \leq 0]$ and $\mathcal{P}[W_t = 1 \,|\, \mathrm{E}_j^{(1)}]$ decay exponentially with K'.

Next, we consider $\mathrm{E}_j^{(2)}$. Moving away from the closest anomalies implies that the GLLR of both children is negative such that

$$\mathcal{P}\left[W_t = 1 \,|\, \mathrm{E}_j^{(2)}\right] = \mathcal{P}\left[S_{\boldsymbol{\theta}_{j_l}} \leq 0\right] \cdot \mathcal{P}\left[S_{\boldsymbol{\theta}_{j_r}} \leq 0\right]. \tag{69}$$

where $\boldsymbol{\theta}_{j_l}$ and $\boldsymbol{\theta}_{j_r}$ are the parameters of the left and right child containing j_l and j_r anomalies respectively. The factors on the RHS of (69) decay exponentially with K'. It follows that there exists a bounded number of samples K' such that (63) holds.

Following the same arguments as in step 1 of Appendix B, we find that the sample complexity of a single random walk is bounded by (53). Consequently, the complexity of the search stages of the K random walks is bounded by

$$\mathcal{Q}_s \leq KB \log_2 M. \tag{70}$$

Step 2: Bound the sample complexity of the target test:

Since, the leaf target test is unaffected by additional anomalies, its sample complexity is bounded by (56) and summing over the K random walks yields

$$\mathcal{Q}_t \leq K\mathbb{E}[\tau_0] \leq \frac{K \log \frac{\log_2 M}{c}}{\mathcal{D}_0\left(\boldsymbol{\theta}_1^{(0)} || \boldsymbol{\theta}_0^{(0)}\right)} + O(1). \tag{71}$$

Step 3: Bound the error rate:

Applying the reasoning in step 3 of Appendix B we find that the error rate is bounded by (62) and applying the union bound over the K random walks yields

$$\mathrm{P}_{\mathrm{Err}}(\pi_{\mathrm{HDS}}) = K\alpha\mathbb{E}[N] \leq K\beta c = O(c). \tag{72}$$

Theorem 2 follows from (70), (71) and (72).

Acknowledgment

In this journal version we include: (i) A more general model that includes multiple anomalies; (ii) a detailed discussion on the implementation of the algorithm; (iii) a rigorous theoretical analysis of the algorithm with detailed proofs; (iv) more

extensive simulation results including new experiments using synthetic and real data; and (v) a detailed discussion of the results, and comprehensive discussion and comparison with the existing literature.

References

[1] B. Wolff, T. Gafni, G. Revach, N. Shlezinger, and K. Cohen, "Composite anomaly detection via hierarchical dynamic search," in *Proc. IEEE Int. Symp. Inf. Theory*, 2022, pp. 2421–2426.

[2] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[3] J. Zhang and M. Zulkernine, "Anomaly based network intrusion detection with unsupervised outlier detection," in *Proc. IEEE Int. Conf. Commun.*, 2006, vol. 5, pp. 2388–2393.

[4] B. Genge, D. A. Rusu, and P. Haller, "A connection pattern-based approach to detect network traffic anomalies in critical infrastructures," in *Proc. Eur. Workshop Syst. Secur.*, 2014, pp. 1–6.

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

[6] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Gener. Comput. Syst.*, vol. 55, pp. 278–288, 2016.

[7] K. Singh, S. Rajora, D. K. Vishwakarma, G. Tripathi, S. Kumar, and G. S. Walia, "Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets," *Neurocomputing*, vol. 371, pp. 188–198, 2020.

[8] K. Thompson, G. J. Miller, and R. Wilder, "Wide-area internet traffic patterns and characteristics," *IEEE Netw.*, vol. 11, no. 6, pp. 10–23, Nov./Dec. 1997.

[9] S.-E. Chiu, N. Ronquillo, and T. Javidi, "Active learning and CSI acquisition for mmwave initial alignment," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2474–2489, Nov. 2019.

[10] T. Simsek, R. Jain, and P. Varaiya, "Scalar estimation and control with noisy binary observations," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1598–1603, Sep. 2004.

[11] C. Wang, K. Cohen, and Q. Zhao, "Information-directed random walk for rare event detection in hierarchical processes," *IEEE Trans. Inf. Theory*, vol. 67, no. 2, pp. 1099–1116, Feb. 2021.

[12] S. Vakili, Q. Zhao, C. Liu, and C.-N. Chuah, "Hierarchical heavy hitter detection under unknown models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6917–6921.

[13] S. Vakili and Q. Zhao, "A random walk approach to first-order stochastic convex optimization," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 395–399.

[14] T. Gafni, K. Cohen, and Q. Zhao, "Searching for unknown anomalies in hierarchical data streams," *IEEE Signal Process. Lett.*, vol. 28, pp. 1774–1778, 2021.

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

[16] K. P. Tognetti, "An optimal strategy for a whereabouts search," *Operations Res.*, vol. 16, no. 1, pp. 209–211, 1968.

[17] J. B. Kadane, "Optimal whereabouts search," *Operations Res.*, vol. 19, no. 4, pp. 894–904, 1971.

[18] Y. Zhai and Q. Zhao, "Dynamic search under false alarms," in *Proc. IEEE Glob. Conf. Signal Inf. Process.*, 2013, pp. 201–204.

[19] D. A. Castanon, "Optimal search strategies in dynamic hypothesis testing," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, no. 7, pp. 1130–1138, Jul. 1995.

[20] K. S. Zigangirov, "On a problem in optimal scanning," *Theory Probability Appl.*, vol. 11, no. 2, pp. 294–298, 1966.

[21] E. Klimko and J. Yackel, "Optimal search strategies for wiener processes," *Stochastic Processes Their Appl.*, vol. 3, no. 1, pp. 19–33, 1975.

[22] 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.

[23] B. Huang, K. Cohen, and Q. Zhao, "Active anomaly detection in heterogeneous processes," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2284–2301, Apr. 2019.

[24] A. Gurevich, K. Cohen, and Q. Zhao, "Sequential anomaly detection under a nonlinear system cost," *IEEE Trans. Signal Process.*, vol. 67, no. 14, pp. 3689–3703, Jul. 2019.

[25] T. Lambez and K. Cohen, "Anomaly search with multiple plays under delay and switching costs," *IEEE Trans. Signal Process.*, vol. 70, pp. 174–189, 2022.

[26] N. K. Vaidhiyan and R. Sundaresan, "Learning to detect an oddball target," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 831–852, Feb. 2018.

[27] S. Nitinawarat and V. V. Veeravalli, "Universal scheme for optimal search and stop," in *Proc. Inf. Theory Appl. Workshop*, 2015, pp. 322–328.

[28] 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.

[29] A. G. Tartakovsky, "Nearly optimal sequential tests of composite hypotheses revisited," *Proc. Steklov Inst. Math.*, vol. 287, no. 1, pp. 268–288, 2014.

[30] A. G. Tartakovsky, G. Sokolov, and Y. Bar-Shalom, "Nearly optimal adaptive sequential tests for object detection," *IEEE Trans. Signal Process.*, vol. 68, pp. 3371–3384, 2020.

[31] R. Caromi, Y. Xin, and L. Lai, "Fast multiband spectrum scanning for cognitive radio systems," *IEEE Trans. Commun.*, vol. 61, no. 1, pp. 63–75, Jan. 2013.

[32] J. Heydari, A. Tajer, and H. V. Poor, "Quickest linear search over correlated sequences," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5786–5808, Oct. 2016.

[33] J. Geng, W. Xu, and L. Lai, "Quickest sequential multiband spectrum sensing with mixed observations," *IEEE Trans. Signal Process.*, vol. 64, no. 22, pp. 5861–5874, Nov. 2016.

[34] A. Tajer and H. V. Poor, "Quick search for rare events," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4462–4481, Jul. 2013.

[35] A. Tsopelakos, G. Fellouris, and V. V. Veeravalli, "Sequential anomaly detection with observation control," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 2389–2393.

[36] A. Tsopelakos and G. Fellouris, "Sequential anomaly detection under sampling constraints," *IEEE Trans. Inf. Theory*, early access, May 23, 2022, doi: 10.1109/TIT.2022.3177142.

[37] 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.

[38] A. Wald, *Sequential Analysis*. Mineola, NY, USA: Dover Publications, 2004.

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

[40] M. Egan, J.-M. Gorce, and L. Cardoso, "Fast initialization of cognitive radio systems," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.*, 2017, pp. 1–5.

[41] B. Hemo, T. Gafni, K. Cohen, and Q. Zhao, "Searching for anomalies over composite hypotheses," *IEEE Trans. Signal Process.*, vol. 68, pp. 1181–1196, 2020.

[42] V. Raghavan and V. V. Veeravalli, "Quickest change detection of a Markov process across a sensor array," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1961–1981, Apr. 2010.

[43] H. Zhang, O. Hadjiliadis, T. Schafer, and H. V. Poor, "Quickest detection in coupled systems," *SIAM J. Control Optim.*, vol. 52, no. 3, pp. 1567–1596, 2014.

[44] G. Fellouris and V. V. Veeravalli, "Quickest change detection with controlled sensing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2022, pp. 1921–1926.

[45] G. V. Moustakides, "Optimal stopping times for detecting changes in distributions," *Ann. Statist.*, vol. 14, no. 4, pp. 1379–1387, 1986.

[46] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[47] D. D. Sleator and R. E. Tarjan, "Self-adjusting binary search trees," *J. ACM*, vol. 32, no. 3, pp. 652–686, 1985.

[48] G. K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, Mar. 2012.

[49] V. Y. Tan and G. K. Atia, "Strong impossibility results for noisy group testing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 8257–8261.

[50] Y. Kaspi, O. Shayevitz, and T. Javidi, "Searching for multiple targets with measurement dependent noise," in *Proc. IEEE Int. Symp. Inf. Theory*, 2015, pp. 969–973.

[51] J. Scarlett, "Noisy adaptive group testing: Bounds and algorithms," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3646–3661, Jun. 2019.

[52] A. Rajwade, N. Shlezinger, and Y. C. Eldar, "Ai for pooled testing of COVID-19 samples," in *Proc. Artif. Intell. COVID-19*. Cham, Switzerland, 2022, pp. 27–58.

[53] P. I. Frazier, S. G. Henderson, and R. Waeber, "Probabilistic bisection converges almost as quickly as stochastic approximation," *Math. Operations Res.*, vol. 44, no. 2, pp. 651–667, 2019.

[54] R. Waeber, P. I. Frazier, and S. G. Henderson, "Bisection search with noisy responses," *SIAM J. Control Optim.*, vol. 51, no. 3, pp. 2261–2279, 2013.

[55] M. Ben-Or and A. Hassidim, “The Bayesian learner is optimal for noisy binary search (and pretty good for quantum as well),” in *Proc. IEEE Symp. Found. Comput. Sci.*, 2008, pp. 221–230.

[56] H. Robbins and D. Siegmund, “A class of stopping rules for testing parametric hypotheses,” in *Proc. 6th Berkeley Symp. Math. Statist. Probability, Volume 4: Biol. Health*, 1972, pp. 37–41.

[57] H. Robbins and D. Siegmund, “The expected sample size of some tests of power one,” *Ann. Statist.*, vol. 2, no. 3, pp. 415–436, 1974.

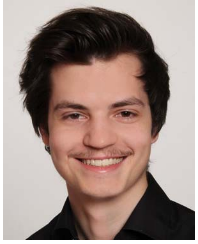
[58] G. Lorden, “On excess over the boundary,” *Ann. Math. Statist.*, vol. 41, no. 2, pp. 520–527, 1970.

[59] “Darpa intrusion detection data sets,” 2000, Accessed: Apr. 18, 2022. [Online]. Available: https://archive.ll.mit.edu/ideval/data/2000data.html

[60] S. Foss, D. Korshunov, and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Distributions* (Springer Series in Operations Research and Financial Engineering Series). New York, NY, USA: Springer, 2011.

[61] H. Robbins and D. Siegmund, “A class of stopping rules for testing parametric hypotheses,” in *Proc. 16th Berkeley Symp. Math. Statist. Probability*, Berkeley, CA, USA, 1972, pp. 37–41.

Tomer Gafni received the B.Sc. and M.Sc. degrees in electrical and computer engineering in 2019 and 2020, respectively, from Ben-Gurion University, Beersheba, Israel, where he is currently working toward the Ph.D. degree with the School of Electrical and Computer Engineering. His main research interests include sequential learning, federated learning, decision theory, and statistical inference and learning, with applications in large-scale systems and wireless networks. He was the recipient of the Kaufmann Award for the Highest Student Achievement in Electrical Engineering, Ben-Gurion university.



Benjamin Wolff received the B.Sc. and M.Sc. degrees in electrical engineering from the Department of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland, in 2020 and 2022, respectively. His main research interests include statistical signal processing, machine learning, and information theory.

Guy Revach received the B.Sc. (*cum laude*), and M.Sc. degrees in electrical and computer engineering from the Andrew and Erna Viterbi Department of Electrical and Computer Engineering, Technion - Israel Institute of Technology, Haifa, Israel, in 2008 and 2017, respectively. He did his master’s thesis under the supervision of Prof. Nahum Shimkin on planning for cooperative multi-agents. Since 2019, he has been working toward the Ph.D. degree in electrical and computer engineering with the Institute for Signal and Information Processing (ISI), ETH Zürich, Zürich, Switzerland, supervised by Prof. Dr. Hans-Andrea Loeliger. He is currently a Researcher with a proven industry track record as an Innovator and System Engineer. His main research interests include intersection of machine learning with statistical signal processing, and more specifically combining sound theoretical principles from classical statistical signal processing with state-of-the-art machine learning algorithms for tracking and detection problems. Before coming to ETH Zürich he was with the Israeli wireless communication industry for more than ten years, first as a real-time Embedded Software Engineer and then a Software Manager. He was the main Innovator behind state-of-the-art, software-defined radio (SDR) for wireless communication, game-changing and groundbreaking in terms of size, weight, and power. As a System Engineer, he defined major aspects of SDR requirements and architecture for hardware, software, network, cyber defense, signal processing, data analysis, and control algorithms.

Nir Shlezinger (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and computer engineering, from Ben-Gurion University, Beersheba, Israel, in 2011, 2013, and 2017, respectively. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Ben-Gurion University. From 2017 to 2019 he was a Postdoctoral Researcher with the Technion, Haifa, Israel, and from 2019 to 2020, he was a Postdoctoral Researcher with the Weizmann Institute of Science, Rehovot, Israel. His research interests include communications, information theory, signal processing, and machine learning. He was the recipient of the FGS Prize for outstanding research achievements from the Weizmann Institute of Science.

Kobi Cohen (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in electrical engineering from Bar-Ilan University, Ramat Gan, Israel, in 2007 and 2013, respectively. He was with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Champaign, IL, USA, from August 2014 to July 2015, and the Department of Electrical and Computer Engineering, University of California, Davis, Davis, CA, USA, November 2012 to July 2014, as a Postdoctoral Research Associate. In October 2015, he joined the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev (BGU), Beer Sheva, Israel, where he is currently an Associate Professor. He is also a Member of the Cyber Security Research Center, and the Data Science Research Center at BGU. His main research interests include statistical inference and learning, signal processing, communication networks, decision theory and stochastic optimization with applications to large-scale systems, cyber systems, wireless and wireline networks. Since 2021, he has been an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Other selected Awards and Honors include highlighting in top 50 popular paper list, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2019, 2020) for paper: Deep multi-user reinforcement learning for distributed dynamic spectrum access, highlighting in popular paper list, IEEE Signal Processing Magazine (2022) for paper: Federated Learning: A signal processing perspective, receiving the Best Paper Award in the International Symposium on Modeling and Optimization in Mobile, Ad hoc and Wireless Networks (WiOpt) 2015, the Feder Family Award (second prize), awarded by the Advanced Communication Center at Tel Aviv University (2011), and President Fellowship (2008-2012) and top Honor List’s prizes (2006, 2010, 2011) from Bar-Ilan University.