RAMP: Real-Time Anomaly Detection in Scientific Workflows

J. Dinal Herath*, Changxin Bai†, Guanhua Yan*, Ping Yang*, and Shiyong Lu†
* State University of New York at Binghamton, Binghamton, NY, USA
† Wayne State University, Detroit, MI, USA

***Abstract*—Research integrity is crucial to ensuring the trustworthiness of scientific discoveries. This work is aimed at detecting misbehaviors targeting scientific workflows, which are computing paradigms widely used to facilitate scientific collaborations across multiple geographically distributed research sites. We develop a new system called *RAMP* (Real-Time Aggregated Matrix Profile) for real-time anomaly detection in scientific workflow systems. RAMP builds upon an existing time series data analysis technique called Matrix Profile to detect anomalous distances among subsequences of event streams collected from scientific workflows in an online manner. Using an adaptive uncertainty function, the anomaly detection model is dynamically adjusted to prevent high false alarm rates. RAMP can incorporate user feedback on reported anomalies and modify model parameters to improve anomaly detection accuracy. Our experimental results from applying RAMP to the logs generated by DATAVIEW, a scientific workflow platform, show that RAMP is able to identify a varied range of anomalies with high accuracy for both interleaved and non-interleaved workflow executions in real time.**

I. Introduction

Research integrity is crucial to ensuring the trustworthiness of scientific discoveries. Scientific misconducts not only cause reputation damages to researchers and institutions involved in the scientific community, but also can have severe real-life implications if dubious research results are transitioned into practical use, such as medicine production and dietary guidelines. Although there have been various regulations to prevent or deter misconducts in scientific research, such misbehaviors are still not uncommon, according to a report in 2009 revealing that 2% of scientists surveyed had falsified, fabricated, or modified their research data [12].

It is thus important to enhance existing cyberinfrastructures used for scientific research activities with capabilities to detect misbehaviors at their early stages before they contaminate the eventual scientific discovery results. This work aims at detecting misbehaviors targeting scientific workflows, which are computing paradigms widely used to facilitate scientific collaborations across multiple geographically distributed research sites. Popular scientific workflows include Montage used by astronomists for image mosaics of the sky [9], CyberShake for generating seismic hazard maps [17], and the myExpriment social network site for bioinformatics researchers [14].

This work aims to develop new techniques that can detect anomalies in scientific workflows in *real time*. The term "real time" is similar to that in [3], [7], which means that the anomaly detection model must observe a data record in a sequential manner and any processing, learning, or anomaly identification must be done before the arrival of the next data record. Real-time anomaly detection has the advantage of catching perpetrators' misbehaviors at their early stages so that the altered or falsified data or code can be prevented from propagating into downstream scientific processes. Although practically appealing, real-time anomaly detection requires us to tackle the following technical challenges. Firstly, the anomaly detection algorithm must be efficiently implemented in order to keep up with the velocity of the event streams observable in scientific workflows. Secondly, the anomaly detection algorithm must be adaptive to situations where there exists only limited supervised information initially. Lastly, the anomaly detection algorithm should be able to adjust its parameters dynamically based on human users' feedback on its reported anomalies.

Against this backdrop, we develop a new system called *RAMP* (*Real-Time Aggregated Matrix Profile*) for real-time anomaly detection in scientific workflow systems. RAMP builds upon an existing time series data analysis technique called Matrix Profile to detect anomalous distances among subsequences of event streams collected from scientific workflows in an online manner. Using an adaptive uncertainty function, the anomaly detection model is dynamically adjusted to avoid repetitive alarms. RAMP can also incorporate user feedback on reported anomalies and retrain model parameters to improve anomaly detection accuracy. We have implemented RAMP to parse logs generated by DATAVIEW, a scientific workflow platform built upon Amazon EC2 [16], and detect anomalous activities in real time.

In a nutshell, our main contributions are as follows:

- We tailor the vanilla Matrix Profile method to real-time anomaly detection with two main modifications: using relative distances among subsequences to avoid inherent biases in Euclidean distance computation and constraining calculation of distance profiles with a small-sized training base to facilitate online model training.
- We introduce a new adaptive training mechanism to reduce false alarm rates commonly plaguing anomaly detection systems in practice. Our method uses a novel uncertainty function that models evolving beliefs in flagged anomalies over time to adjust model parameters. This technique prevents RAMP from triggering repetitive alarms due to the same type of anomalies.
- We empower RAMP with an optional human-in-the-loop training scheme. To improve anomaly detection accuracy, our method carefully modifies RAMP's model parameters based on human users' feedback.
- We compare the performance of RAMP against those of

978-1-7281-0858-2/19/$31.00 © 2019 IEEE

two state-of-the-art real-time anomaly detection models and our results show that RAMP has superior performances in various anomaly situations while achieving real-time responsiveness.

The rest of the paper is organized as follows. Section II provides the background information. Section III discusses the threat model and the types of anomalies RAMP detects. Section IV and Section V present an overview of RAMP and its algorithm details, respectively. Experimental results are given in Section VI. We discuss related work in Section VII and draw concluding remarks in Section VIII.

II. Background

This section provides an overview of scientific workflows and the Matrix Profile, the base model used to build RAMP.

A. Scientific Workflows and DATAVIEW

Scientific workflow is a cyberinfrastructure paradigm for automating and accelerating data processing and sharing in the scientific community. Figure 1 shows a diagnosis recommendation workflow [4] consisting of five workflow tasks $T_1 - T_5$, each of which represents a computational or analytical step.

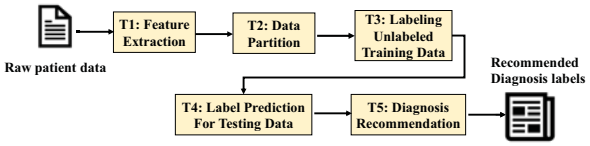


Fig. 1: A Diagnosis Recommendation Workflow

DATAVIEW is a scientific workflow management system running on Amazon EC2 [16]. Its logs record the real-time status (events) of scientific workflows executed on EC2, including the task execution status (e.g., task-start and task-completion), the communication between the local machine and the EC2 VMs (e.g., task-send and task-receive), and the machine provisioning status (e.g., machine-idle and machine-ready). Each log entry is associated with a timestamp, which specifies the date and time of an event. The log also contains the IP address of VMs on which workflow tasks are executed.

B. Matrix Profile

Matrix Profile [27] is a machine learning model that enables the identification of similar (called *motifs*) and dissimilar (called *discords*) patterns in a given time series. When new data is appended to the original time series input, Matrix Profile does not need to re-compute motifs and discords from the beginning using all the data points; instead it is able to change the existing results with little computational overhead. This incrementally updatable nature enables Matrix Profile to identify previously un-identified motifs that appear due to a new data stream with relative ease. Matrix Profile's incremental updatability, fast execution, and low need for parametric tuning (*i.e., only one parameter to tune*) ([5], [26], [27], [29]) makes it an attractive model for real time machine learning applications.

Define a *univariate time series T* which is a sequence of real numbers $T = t_1, t_2, \ldots, t_n$. The Matrix Profile gives insight about the global similarity or dissimilarity in a time series, but this is computed with respect to local *subsequences*. A *subsequence* $T_{i,m}$ of T is a continuous subset of the values from T with a given length m starting at position i (i.e. $T_{i,m} = t_i, t_{i+1}, \ldots, t_{i+m-1}$ where $1 \leq i \leq n - m + 1$). For any given subsequence in a time series, it is possible to compute the Euclidean distance from itself to all other subsequences. An ordered vector of the Euclidean distances between a given subsequence $T_{i,m}$ and an set of all subsequences $[T_{1,m}, T_{2,m}, ..., T_{n-m+1,m}]$ is called a *distance profile* D. By extension, a multivariate time series $\mathbf{T}$ of d dimensions is a set of co-evolving univariate time series where $\mathbf{T} = [T^{(1)}, T^{(2)}, ..., T^{(d)}]$ and a multivariate subsequence is given by $\mathbf{T_{i,m}} = [T^{(1)}_{i,m}, T^{(2)}_{i,m}, ..., T^{(d)}_{i,m}]$.

Let T be a complete univariate time series and m be the length of the subsequence. Matrix Profile works as follows. First, it computes the distance profile D_i for every subsequence $T_{i,m}$. The Matrix profile value at step i is obtained as the minimum recorded value in D_i, excluding the Euclidean distance from $T_{i,m}$ to itself which is trivially 0. Repeating this process for the complete time series results in an ordered vector of minimum Euclidean distance values corresponding to each subsequence, which is called a *Matrix Profile*. A small value in the Matrix Profile indicates that the subsequence pattern is observed elsewhere in the time series (i.e. a *motiff*). An abnormally large Matrix Profile value indicates that the corresponding subsequence is not observed elsewhere in the time series and hence may be a *discord*.

III. Threat Model and Attacks

We assume that workflow logs are protected and are not tampered. We also assume that an attacker may exploit the vulnerabilities in the workflow to modify the workflow tasks and/or the communication between two tasks to alter scientific results. In addition, an attacker can attack workflow systems using Denial Of Service (DOS) attacks. We consider two types of anomaly situations: **(1) Level-1 anomalies**: anomalies resulting from direct attacks or malfunctions of DATAVIEW; **(2) Level-2 anomalies**: attacks that attempt to hide the true intent of an attack and confound the anomaly detection model.

A. Level-1 Anomalies

L1A1-Unexpected scheduler change: DATAVIEW provides several task scheduling options, which deploy the workflow tasks on different numbers of VMs. We consider the situation where tasks are scheduled and executed on the VMs in a pattern that is not predefined. This can be caused by an attacker who intends to increase the load on DATAVIEW or as a part of an adversarial attack (L2A1) described below.

L1A2-DOS attack: In DATAVIEW, after EC2 VMs are provisioned, there are continuous communications between the local machine used by a user and the VMs on which the workflow tasks are executing, including task specifications, task execution status, and VM status. We consider the DOS attacks or spikes in network traffic which may slow down the execution of scientific workflows.

L1A3-Task manipulation: This attack is performed by malicious users who have access to the source code of the workflow tasks. The malicious users modify the task code or inject pre-computed values into the task execution to manipulate the

workflow result. RAMP detects L1A3 attacks that result in an increase/decrease on the task execution time.

L1A4-Workflow structure manipulation: When the task code is unavailable, a malicious user can inject a task or change the workflow structure to manipulate the workflow result.

B. Level-2 Anomalies

L2A1-Workflow structure manipulation with Scheduler change: The attacker performs a scheduler change attack (L1A1) and a workflow structural change attack (L1A4) simultaneously. In this attack, the attacker aims to use the VM provisioning change to mask the code change in the task, which modifies the final results of the scientific workflow.

L2A2-Task manipulation with DOS attack: This combined attack performs a DOS attack (L1A2) and a task manipulation attack (L1A3) simultaneously. Both attacks affect the execution time of workflows. A combined attack of this nature aims to mask task manipulation occurring during a DOS attack.

IV. Overview of RAMP

This section provides an overview of RAMP, which detects anomalies in scientific workflows based on DATAVIEW logs.

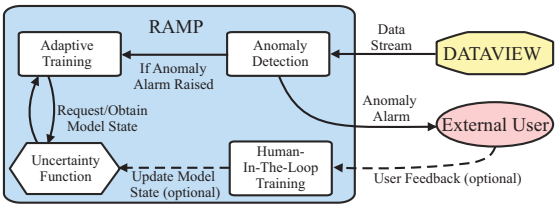


Fig. 2: The Architecture of RAMP

log_1	2019/06/13-12:03:01.340 machine-ready 204.236.200.9
log_2	2019/06/13-12:03:01.763 machine-ready 54.196.14.157
log_3	2019/06/13-12:03:02.184 machine-ready 54.147.255.97
log_4	2019/06/13-12:03:02.601 machine-ready 34.204.71.118

TABLE I: Log entries for VM Provisioning in DATAVIEW

Figure 2 gives the architecture of RAMP. The DATAVIEW log entries are parsed into a time series with three dimensions (features). The first is the time difference between two successive log entries computed in millisecond precision. The second reflects the change in process status between two consecutive logs. In Table I, the term *machine-ready* reflects the machine provisioning occurring at each log step. Therefore, the time series would contain [machine-ready→machine-ready] occurring for three times. The final dimension reflects the change in IP address between two successive log entries as [[204.236.200.9 $\rightarrow$ 54.196.14.157] , [54.196.14.157 $\rightarrow$ 54.147.255.97] , [54.147.255.97 $\rightarrow$ 34.204.71.118]]. Once the log entries are parsed, subsequences are formed and given to RAMP. With a subsequence length of 3, the first subsequence is $[log_1, log_2, log_3]$ and the second one $[log_2, log_3, log_4]$.

RAMP has three main components: an *Anomaly Detection* module, an *Adaptive Training* module, and an optional *Human-in-the-Loop training* module. At each time step, the *Anomaly Detection* module takes as input a subsequence of data stream and computes a weighted aggregated anomaly score, which signifies the possibility that the input subsequence is an anomaly or not. The anomaly score aggregates the prediction results of individual Matrix Profile models operated on different dimensions of the input subsequence. As it is unclear what types of anomalies could happen in the future, RAMP uses a semi-supervised model where RAMP learns the correct behaviour of workflow execution in the first few workflow runs and then identifies instances that heavily deviate from it as potential anomalies. This approach is also used in other real-time machine learning models (e.g., anomaly detection models in the Numenta Anomaly Benchmark [1]), where an initial grace period is needed to learn the correct behaviour of real time data. This grace period, however, may not be sufficient for models to continuously update its internal state and detect anomalies with high accuracy.

The *Adaptive Training* module is designed to facilitate fast state convergence in real time. The *Adaptive Training* module is invoked to update model weights whenever the anomaly detection module flags an anomaly. These weights are updated according to the anomaly score reported by the *Anomaly Detection* module and an *uncertainty function*, which probabilistically captures the model state. The *uncertainty function* assumes that anomalies are less likely to occur in the first few executions of a scientific workflow and the likelihood becomes higher with more runs. The intuition behind this assumption is that an attacker has little information to attack a workflow in its first few execution runs but the situation changes once he gains more knowledge about the workflow. The *Adaptive Training* module ensures that RAMP will continuously learn the temporal behaviour of a time series and subsequently improves the overall performance.

Finally, RAMP uses an optional *human-in-the-loop training* module to improve anomaly detection accuracy based on human feedback. Specifically, given the true positives identified by human users, this module revises the corresponding model weights so that similar anomalies encountered in the future are more likely to be caught by RAMP.

V. Algorithm Details

This section presents the details of RAMP. We assume that, in every M user defined time steps, RAMP invokes the human-in-the-loop training to process human feedback.

A. Anomaly Detection Module

RAMP uses a modified version of Matrix Profile for anomaly detection. Our description here uses the same notations as those in Section II-B. We first explain our key modifications to Matrix Profile and then present the details of our anomaly detection algorithm.

1) Modifications to Matrix Profile: One modification we made to the Matrix Profile is limiting the number of subsequences compared. For a given subsequence, Matrix Profile computes the Euclidean distance with respect to all other subsequences and identifies the minimum distance. Therefore, a repeated anomaly instance would cause false negatives due to the previous anomaly instance being part of the all subsequence set. RAMP, in contrast, uses a semi-supervised model where only the first $M-m+1$ subsequences (i.e., subsequences up to the M-th item in the time series) are considered. We define the *training base* as $\mathbf{T}^{*} = [\mathbf{T_{1,m}}, \mathbf{T_{2,m}}, ..., \mathbf{T_{M\text{-}m+1,m}}]$

with d dimensions (i.e., $\mathbf{T_{1,m}} = [T_{1,m}^{(1)}, T_{1,m}^{(2)}, ..., T_{1,m}^{(d)}]$). We illustrate in Figure 3 the demarcation of subsequences and the initial training base for a univariate time series ($d = 1$) where $m = 3$ and $M = 50$.

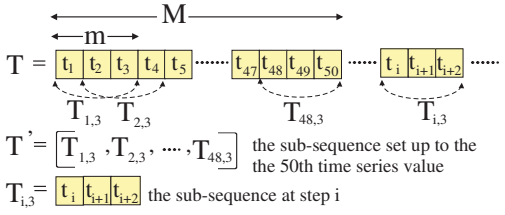


Fig. 3: Demarcation of subsequences

Our second modification is that, instead of computing the absolute Euclidean distance, we compute relative distances among subsequences. The purpose behind this modification is to overcome the inherent bias of Euclidean distance towards numerically larger data points. Given two univariate subsequences $T_{1,m} = [t_1, t_2, ..., t_m]$ (an input subsequence) and $T'_{1,m} = [t'_1, t'_2, ..., t'_m]$ (a subsequence used for comparison), the relative distance is computed by $\frac{\sum_{l=1}^{m} |t_l - t'_l|}{\sum_{l=1}^{m} |t'_l|}$.

Procedure `AnomalyDetection`

1 $\beta_i = 0$, $C_i = \emptyset$, $key = 0$, $D_{min} = \emptyset$
2 **for** $j = 1$ *to* d **do**
3 $min\_rd = +\infty$, $min\_k = -1$
4 **for** $k = 1$ *to* $M - m + 1$ **do**
5 $relativeDistance = \frac{\sum_{l=1}^{m} |T_{k,m}^{(j)}[l] - T_{i,m}^{(j)}[l]|}{\sum_{l=1}^{m} |T_{k,m}^{(j)}[l]|}$
6 **if** $relativeDistance > min\_rd$ **then**
7 $min\_rd = relativeDistance$
8 $min\_k = k$
end
9 $R[j, i\%M + 1] = min\_k$
10 $key \mathrel{+}= (M - m + 1)^{j-1} \cdot (min\_k - 1)$
11 $\beta_i \mathrel{+}= min\_rd$
12 $D_{min}[j] = min\_rd$
end
13 **for** $j = 1$ *to* d **do**
14 $C_i[j] = D_{min}[j]/\beta_i$
end
15 **if** *key exists in* W **then**
16 $\beta_i = W[key] \times \beta_i$
17 $H[i\%M + 1] = \beta_i/\theta$
18 **if** $(\beta_i > \theta)$ $AnomalyDetected = True$
19 **else** $AnomalyDetected = False$
20 **return** $[AnomalyDetected, C_i, R, H]$

Algorithm 1: *Anomaly Detection* procedure

2) Algorithm description: Our anomaly detection algorithm is given in Algorithm 1. For every time step i, the *Anomaly-Detection* procedure computes an aggregated anomaly score, β_i, which indicates the likelihood of an anomaly. In order to compute β_i, for each dimension of the input subsequence $\mathbf{T_{i,m}}$, we first calculate the minimum relative distance between $\mathbf{T_{i,m}}$ and any subsequence in the training base $\mathbf{T}$'. The aggregated anomaly score β_i is updated to be the sum of the minimum relative distances over all d dimensions (Line 2-12). The individual contribution of each dimension towards the aggregated anomaly score, which is calculated as the ratio of the minimum relative distance between $\mathbf{T_{i,m}}$ and any subsequence in the training base $\mathbf{T}$' to β_i, is saved into a contribution list $C_i = [C_i^{(1)}, C_i^{(2)}, ..., C_i^{(d)}]$ (Line 13-14). For example, $C_i^{(1)} = 1$ indicates that the anomaly score is decided only by the first dimension in the time series.

It is noted that the aggregated anomaly score β_i is affected by a *unique* combination of d subsequences in the training base $\mathbf{T}$', each having a minimum relative distance from the current input subsequence $\mathbf{T_{i,m}}$ at one of the d dimensions. As there are $M - m + 1$ subsequences in the training base, there are $(M - m + 1)^d$ possible combinations over d dimensions. Our algorithm keeps a weight for each of these combinations, reflecting the model's confidence level in the aggregated anomaly score if it is derived from this combination. In Algorithm 1, the index of the unique combination responsible for the computation of β_i is stored in variable key (Line 10). The eventual anomaly score β_i is derived by multiplying it by the weight indexed by key (Line 15-16). It is easy to see a challenge due to the curse of dimentionality: if d is large, it is expensive to store all $(M - m + 1)^d$ possible weights. To circumvent this problem, we use a hash table W to store only updated weights, while assuming that those weights not in W take a default value of 1. Our experimental results in Section VI-C show that only a small fraction of possible weights are updated by RAMP in practice.

Recall that the human-in-the-loop training module is called by RAMP every M time steps to process any feedback from users. As users may overrule the anomaly detection results by RAMP in the past M time steps, RAMP should remember the indices of weights used to compute the anomaly scores during this period. To this end, we use a matrix R of dimensions $d \times M$ to record the indices of weights that have been updated in the past M time steps. Entry $R[j, k]$ stores the index of the subsequence within the training base $\mathbf{T}$' that has the minimum relative distance for dimension $j \in [1, d]$ at the k-th time step among the past M ones (i.e., $k \in [1, M]$). This is done by Line 9 in Algorithm 1. Additionally, array H of length M stores the ratios of the anomaly scores to the user-defined threshold (i.e., $\frac{\beta_i}{\theta}$) for the past M time steps (Line 17). The aggregated anomaly score β_i is compared against a user-defined threshold θ (Line 18): if $\beta_i > \theta$, an anomaly is flagged, which triggers the execution of the adaptive learning module; otherwise, the anomaly detection module terminates.

B. Adaptive Training

The original Matrix Profile model does not adjust its model parameters based on the anomalies detected, which can cause repetitive false alarms for a long period of time. To reduce false positive rate, our adaptive training module, which is called when an alarm is raised by the anomaly detection module, adjusts not only the model parameters affected at the current time step but also those that may be affected in the near future.

Algorithm 2 gives the *Adaptive Training* procedure. Recall that there are $(M - m + 1)^d$ possible weights affecting the anomaly scores. To achieve real time training, we use a training heuristic, which is experimentally validated in Section VI-D, to select only $(2m + 1)$ weight values for updating irrespective of the dimensionality d of the input. The middle $m + 1$ weight values, which correspond to $k = m + 1$ in

Procedure `AdaptiveTraining`

1 $keys = zeros[1, 2m+1]$
2 **for** $k = 1$ *to* $2m+1$ **do**
3 **for** $j = 1$ *to* d **do**
4 $keys[k]$ += $(M-m+1)^{j-1} \cdot (R[j, i\%M+1] - m + k - 1)$
  end
5 **if** $keys[k]$ *does not exist in* W **then**
6 $W[keys[k]] = 1$
7 **if** $k == m+1$ **then**
8 $\beta_i^{unweighted} = \frac{H[i\%M+1] \cdot \theta}{W[keys[k]]}$
9 $W[keys[k]] = \frac{W[keys[k]]}{2H[i\%M+1]} \cdot \alpha[k] \cdot (1 - p_i)$
10 $H[i\%M+1] = \frac{W[keys[k]] \cdot \beta_i^{unweighted}}{\theta}$
  else
11 $W[keys[k]] = \frac{W[keys[k]]}{2H[i\%M+1]} \cdot \alpha[k] \cdot (1 - p_i)$
  end
 end
12 **return** $[W, H]$

Algorithm 2: *Adaptive Training* procedure

Algorithm 2, are called the *anomaly-inducing weights*, as they are the same ones used to update the anomaly score by the anomaly detection module (Line 16 in Algorithm 1). The keys indexing the $(2m+1)$ weights in hash table W to be updated are calculated in Lines 3-4.

These selected weights are modified in Lines $7-11$ according to the following equation:

$$W[key] = \frac{W[key]}{2H[i\%M+1]} \cdot \alpha[k] \cdot (1 - p_i), \tag{1}$$

where $W[key]$ is the weight to be updated, $H[i\%M+1]$ is the ratio of the anomaly score computed ($\beta_{i\%M+1}$) to threshold θ, $\alpha[k]$ is a training bias value that captures temporal correlation for the k-th chosen weight, and p_i reflects the uncertainty level at time step i. Due to weight updating, $H[i\%M+1]$, which stores $\frac{\beta_i}{\theta}$, should also be updated. This is done by Line 10 in Algorithm 2 where the unweighted anomaly score $\beta_i^{unweighted}$ (i.e., the unweighted value before Line 16 of Algorithm 1) is first derived and then used to update $H[i\%M+1]$ along with the new weight.

The rationale behind Eq. (1) is that future anomaly score calculations using the same weight indexed by key should result in smaller values for β_i to avoid repetitive alarms. As the adaptive training module is called when an anomaly is detected (i.e., $\beta_i > \theta$), $H[i\%M+1]$, which is calculated as β_i/θ should always be greater than 1. The first portion of the formula $\frac{W[key]}{2H[i\%M+1]}$ aims to reduce the weight value so that a future use of the same weight on the same unweighted anomaly score will result in exactly half of the threshold θ. This is because, assuming that the unweighted anomaly score is β_0 and the old weight is w_0, we have: $H[i\%M=1] = w_0\beta_0/\theta$; as $\frac{W[key]}{2H[i\%M+1]} = \theta/(2\beta_0)$, using it on the same unweighted anomaly score β_0 leads to a weighted anomaly score of $\theta/2$.

Uncertainty function p_i**:** Each weight in Eq. (1) is multiplied by a factor of $1 - p_i$, where $p_i \in [0, 1]$ is an *uncertainty function* capturing the model state:

$$p_i = \begin{cases} 1 - \exp((K_i)^b - (i)^b), & \text{if } i > M \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Here, i is the current time step, K_i is a state variable updated upon user feedback (see Section V-C), and $b \in [0, 1]$ is a user-defined bias parameter, which defines the rate at which p_i converges to 1. When p_i is close to 0, RAMP believes that the reported anomalies are false positives with a high certainty. When p_i is close to 1, RAMP believes that the reported anomalies are likely to be true positives. At the beginning of RAMP, K_i is initialized to M whereas b is given by the user. The greater the b, the faster the convergence. RAMP uses the first M input values to build $\mathbf{T}$'. Therefore, p_i is 0 when $i \leq M$ and starts increasing when $i = M$. Without any user feedback p_i will gradually increase to 1 with no sudden drops. This reflects our assumption that anomalies are more likely to occur with longer running time. In Section V-C, we illustrate how the uncertainty function changes with the user feedback.

Training bias α**:** As the training procedure updates $2m+1$ temporally correlated weights, each one is multiplied by its respective training bias. The training bias α is a vector of size $2m+1$ that captures the temporal correlation among $2m+1$ weights. We assume that the temporal correlation varies according to a normalized Gaussian distribution $N(0, m)$ (i.e., the largest value in distribution at mean 0 is 1) around an identified false positive weight. Note that the sampling process is carried out once for an entire execution of RAMP since $2m+1$ weights are updated at each training cycle. Additionally α is constructed by sampling values from the distribution starting from $-m$ to $+m$ with unit step size such that $\alpha[1]$ and $\alpha[2m+1]$ will have the smallest values and $\alpha[m+1]$ will have the highest value of 1.

C. Human-in-The-Loop Training

Algorithm 3 describes our *Human-in-the-loop training* module. RAMP checks whether there is human feedback every M time steps. The human feedback includes the lists of time step indices with false positives U_{FP} and true positives U_{TP} among the previous M time steps.

Procedure `HumanInTheLoopTraining`

1 $K_i = K_{i-M} + [i - K_{i-M}] \frac{|U_{FP}|}{M}$
2 **for** $i_{TP} \in U_{i,TP}$ **do**
3 $key = 0$
  for $j = 1$ *to* d **do**
4 key += $(M-m+1)^{j-1} \cdot (R[j, i_{TP}\%M+1] - 1)$
  end
5 $W[key]$ = $\frac{2W[key]}{H[i_{TP}\%M+1]}$
 end
6 **return** $[K_i, W]$

Algorithm 3: *Human-In-The-Loop Training* procedure

Recall that in Eq. (2), state variable K_i is used to calculate the uncertainty function p_i. The human-in-the-loop training procedure first updates K_i according to the formula $K_i = K_{i-M} + [i - K_{i-M}] \frac{|U_{FP}|}{M}$ (Line 1). The intuition here is that, if a large fraction of the previous M time steps has raised false alarms, we decrease the confidence in the model prediction results by forcing a sudden drop in uncertainty function p_i.

In addition to updating the state variable K_i, the human-in-the-loop training module also updates the weight values that may have been erroneously trained in the past. As RAMP

reduces the value of each anomaly-inducing weight to avoid high false alarm rates without human feedback (Line 9 in Algorithm 2), RAMP rectifies these weights if their corresponding anomalies are verified to be true positives by human users. Given the user-provided set $U_{i,TP}$, which includes the time step indices of true positives, the keys indexing their corresponding anomaly-inducing weights in hash table W are calculated in Lines 3-4; these weights are then updated in a similar manner as in Eq. (1), except that p_i = 0 because we assume the user feedback to be the ground truth (Line 5). Moreover, for an anomaly-inducing weight, its corresponding training bias α is always 1. Thus, the effect of weight updating is that a future use of the new weight on the same unweighted anomaly score as in time step i_{TP} should result in a weighted anomaly score of exactly 2θ.

VI. Experimental Results

RAMP was implemented using Python. We have measured the performance of RAMP using worklow logs collected by the DATAVIEW system running on Amazon EC2 VMs.

In our experiments, we consider the interleaved execution of three workflows – Ligo [6], Wordcount [8] and Diagnosis Recommendation [4]. At any given point in time, only one of the three workflows runs on EC2 VMs. All workflow executions are recorded on a single log instance and no indication is given to anomaly detection models with respect to which log entry corresponds to which workflow. We conducted experiments to test the robustness of RAMP and its ability to handle noisy data. For each anomaly type, we executed the workflows for six hours and collected the logs containing about 5000 data points in the time series. The parameters used for performance comparison are $m = 10, M = 200, b = 0.8$ for Level-1 anomalies, and are $m = 5, M = 200, b = 0.8$ for Level-2 adversarial attacks.

A. Performance Comparison: Anomaly Detection

This section presents the performance results of RAMP on detecting Level-1 and Level-2 attacks. We compare three RAMP versions based on the extent of user feedback given, namely *RAMP* (RAMP with both adaptive and human-in-the-loop training), *RAMP-no-feedback* (RAMP with only adaptive training), and *RAMP-oracle* (RAMP with feedback from an all-knowing oracle). *RAMP-no-feedback* is unable to rectify any erroneous training or update its internal state due to the lack of user feedback. *RAMP-oracle* updates the weights for both true positives and false negatives identified by the oracle as done in Lines 3-5 of Algorithm 3. In addition, RAMP is compared with two other real-time machine learning models – Hierarchical Temporal Memory (HTM) [3] and KNN-CAD [7] – which are available in the open source Numenta Anomaly Benchmark (NAB) [1]. Among all the machine learning models in NAB, these two were shown to have good performances in detecting anomalies in scientific workflows [20]. As the anomaly detection models in NAB are designed for univariate time series, we consider one dimensional data input where the dimension most affected by a given anomaly type is fed into all the models for Level-1 anomaly situations. As Level-2 adversarial anomalies affect multiple dimensions simultaneously, HTM and KNN-CAD models were not used in performance evaluation in these instances.

1) Level-1 Anomalies: Figures 4(a)–4(d) give the Receiver Operator Characteristics (ROC) for Level-1 anomalies for the interleaved execution of three workflows. The figures show that all the RAMP versions have significantly lower false positives and higher true positives than HTM and KNN-CAD. On average considering the Area Under Curve (AUC) results in Table II for anomalies L1A1-L1A4, RAMP shows an 46.62% and 84.14% increase in AUC compared to HTM and KNN-CAD, respectively. We attribute this increase to the adaptive training module, which adjusts model weights to avoid reporting repetitive anomalies.

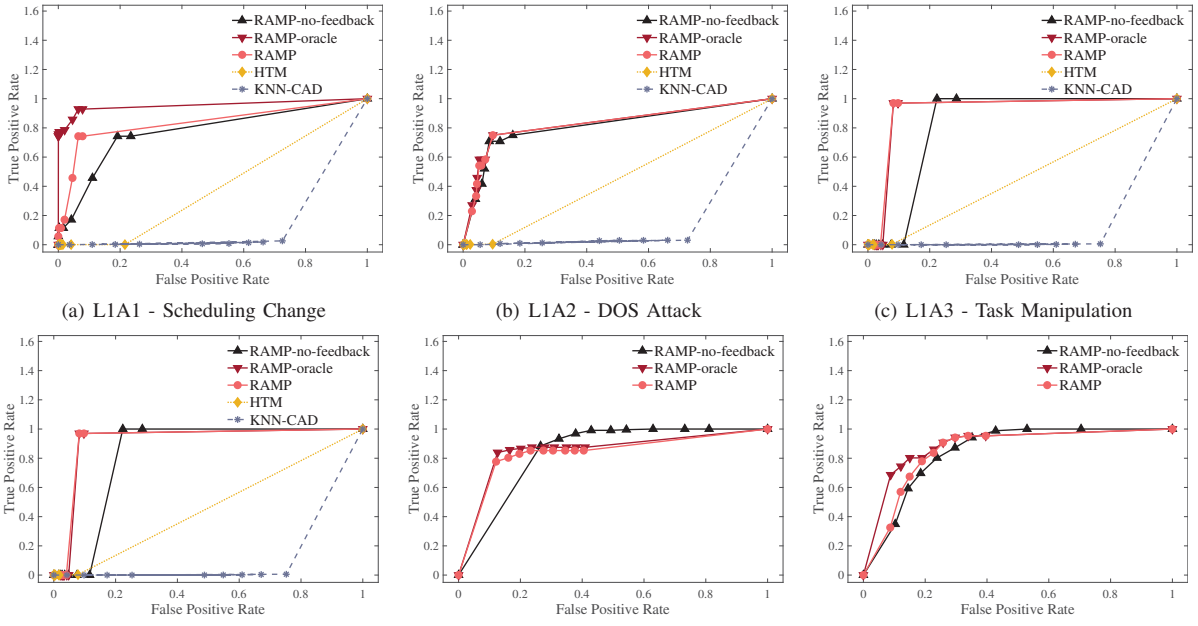
Figure 4(a) shows that, *RAMP-oracle* has higher true positive rates than the other models on L1A1. In addition, *RAMP-no-feedback* has slightly higher false positive rates than the others. We note similar results with respect to other anomaly situations (Figures 4(b)-4(d)). However, when comparing the AUC results of three RAMP versions (Table II), the performance of RAMP is close to that of *RAMP-oracle* and *RAMP-no-feedback* by a difference of $\pm 5\%$.

We note similar results in single workflow executions, where all RAMP versions greatly outperforms HTM and KNN-CAD. Due to lack of space, we present only the AUC results for single workflow execution in Table II. As single workflow execution does not contain noise, RAMP has increased performance in all anomaly situations (Average Level-1 RAMP AUC increases from 0.8883 to 0.9959). We observe that the noise caused by interleaved workflow executions has a higher impact on inter-workflow anomalies L1A1 and L1A2 than intra-workflow anomalies L1A3 and L1A4. Inter-workflow anomalies occurring to a given workflow instance can interfere with other interleaving workflows. RAMP has slightly lower AUC for inter-workflow anomalies, as shown in Figures 4(a) and 4(b), where the maximum true positive rate for RAMP is around 0.8. In contrast, the performance with intra-workflow anomalies is not affected by interleaving workflows because they are local to a single workflow instance.

2) Level-2 Adversarial Anomalies: Figures 4(e) and 4(f) give the performance results of all RAMP versions for Level-2 anomalies. As expected, even though their AUCs decrease slightly, all three RAMP models perform well in adversarial situations (Table II). *RAMP-no-feedback* has its false positive rates increased more significantly than the other two RAMP variants along with a slight increase in its true positive rates. We further note that the performance of RAMP is close to that of *RAMP-oracle*, suggesting that RAMP performs robustly even in adversarial situations.

B. Execution Performance: RAMP Response Time

Figure 5 shows the response time of RAMP in Level-1 anomaly scenarios for the first $3,000$ seconds. The x-axis shows the relative time at which DATAVIEW records a log entry; the average rate is 3.4368 seconds per log entry. The y-axis shows the sum of the RAMP response time and the relative logging time by DataView. The figure shows that for all anomaly scenarios the time closely follows the $y = x$ reference line. More specifically, the execution time for RAMP inclusive of adaptive training is approximately 0.0118 seconds



(a) L1A1 - Scheduling Change (b) L1A2 - DOS Attack (c) L1A3 - Task Manipulation

(d) L1A4 - Workflow Manipulation (e) L2A1 - Scheduler Change (f) L2A2 - Task Manipulation with DOS Attack

Fig. 4: Receiver Operator Characteristics (ROC) for Interleaved Scientific Workflow Anomalies. Each data point is generated between Threshold (θ) values from $0-1$ with a step size of 0.1 where the highest θ of 1 is at the furthest left.

	Anomaly Type	Anomaly Detection Models				
		RAMP-no-feedback	RAMP-oracle	RAMP	HTM	KNN-CAD
Int.	L1A1 - Scheduler Change	0.7745	0.9538	0.8352	0.4927	0.1498
	L1A2 - DOS Attack	0.8164	0.8315	0.8296	0.4540	0.1516
	L1A3 - Task Manipulation	0.8303	0.9200	0.9236	0.4623	0.1265
	L1A4 - Workflow Manipulation	0.9824	0.9784	0.9645	0.4798	0.1348
	L2A1 - Adversarial Attack	0.8417	0.8528	0.8360	-	-
	L2A2 - Adversarial Attack	0.8474	0.8474	0.8538	-	-
Non-Int.	L1A1 - Scheduler Change	0.8708	0.9172	0.8625	0.4706	0.3492
	L1A2 - DOS attack	0.9620	0.9653	0.9653	0.2160	0.3414
	L1A3 - Task Manipulation	0.9968	0.9968	0.9968	0.3367	0.1695
	L1A4 - Workflow Manipulation	0.9963	0.9989	0.9989	0.4440	0.1674

Table II: Area Under the Curve (AUC) results for Reciever Operator Charachteristics (ROC) for Interleaved (Int.) and Non-Interleaved (Non-Int.) workflows.

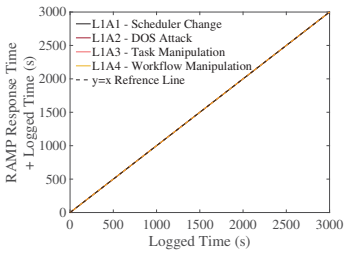


Fig. 5: RAMP Response Time

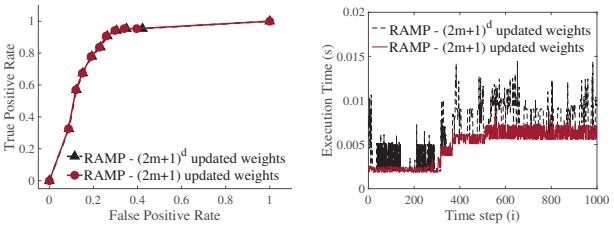
on average, which is about 0.0034 times the data stream speed. The results demonstrate the real-time effectiveness of RAMP, where the time taken to detect an anomaly is negligible.

C. Space Complexity: Sparsity of Modified Weights

As mentioned in Section V-A, the total number of possible weights is $(M-m+1)^d$. To reduce the space complexity, we store only weights modified by RAMP in a hash table. We obtained the weights modified by RAMP for each workflow in the interleaved scenario for L2A1 where $M=200, m=5$ and $d=3$. Our experimental results show that the total number of weights is 7529536 for each workflow, but only 131, 496, and 250 weights are updated for Diagnosis Recommendation, Ligo, and Wordcount workflow, respectively.

D. Time Complexity: Adaptive Training Heuristics

As described in Section V-B, our optimized adaptive training module updates only $2m+1$ weight values at each step instead of $(2m+1)^d$ weights in complete training. Figures 6(a) and 6(b) show the ROC and the execution time for the first 1000 input subsequences received for the complete training and the optimized training, respectively. Figure 6(a) shows that the two training schemes have similar numbers of true



(a) ROC (b) Execution time

Fig. 6: Performance of our optimized adaptive training.

and false positives, and Figure 6(b) shows that our optimized training method is about 2.5 times faster than the complete training one. Note that there is a sudden increase in execution time after about 300–400 time steps for both training methods due to the interleaved execution of workflows.

VII. Related Work

Real-time anomaly detection: Hierarchical Temporal Memory (HTM) is a model designed to replicate the neocortex of mammals and how the neurons learn and predict [3]. While HTM has been used in many real-time anomaly detection applications (e.g., [1], [20]), it interprets only univari-

ate time series data. Bayesian Online Checkpoint Detection (BOCD) [2] is capable of identifying anomalies in streaming data. However, it assumes that the underlying distribution of data is known. KNN-CAD [7] is a model that uses the density and the distance based nearest neighbour algorithm for anomaly detection. Relative Entropy [25] uses Turkey and relative entropy statistics for anomaly detection. EXPoSE [23] is an anomaly detection model that handles multidimenional data. However, as shown in [1], EXPoSE performs poorly when datasets are of moderate sizes and are univariate.

Anomaly detection in distributed systems: Anomaly detection models have been extensively used in the domain of distributed systems, ranging from preventing DOS attacks [18] to detecting anomalous usage in VMs in the cloud [10]. Various Machine Learning models have been proposed in this application domain, including the non-parametric clustering model [28], the Support vector Machine model [19], and the use of Baysian Classifiers and tree augmented networks to identify anomalies in distributed systems [24]. The work in [11] uses deep learning to detect anomalies in system logs. However, none of the above models are real-time models.

Anomaly detection in scientific workflows: Samark *et al.* [22] introduced a Naive Bayes model to predict the likelihood of a job success or failure in scientific workflows. The work in [21] proposes an unsupervised model based on K-means clustering that detects hard and soft anomalies in an online manner. The work in [15] is aimed at identifying time periods where majority of anomalies occur. Gaikwad *et al.* [13] proposed a framework to detect performance anomaly w.r.t. the execution time in scientific workflows. Rodriguez *et al.* [20] conducted a similar study with respect to performance anomalies in scientific workflows. However, none of the above works have considered anomalies resulting from malicious attacks or identified anomalies in multivariate time series.

VIII. Conclusions

This work is aimed at enhancing existing scientific workflow platforms with misbehavior detection capabilities. We develop RAMP, a novel real time anomaly detection model that takes multivariate streaming data input from DATAVIEW logs and produces anomaly alarms in real time. Our experimental results show that RAMP has superior performance while achieving real-time responsiveness in a variety of anomaly situations.

Acknowledgement: This work is supported in part by the National Science Foundation under grant OAC-1738929.

References

[1] Numenta, https://github.com/numenta/nab, 2019.

[2] R. P. Adams and D. J. MacKay. Bayesian online changepoint detection. *arXiv:0710.3742*, 2007.

[3] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134–147, 2017.

[4] I. Ahmed, S. Lu, C. Bai, and F. A. Bhuyan. Diagnosis recommendation using machine learning scientific workflows. In *IEEE Congress on Big Data*, pages 82–90, 2018.

[5] S. D. Anton, L. Ahrens, D. Fraunholz, and H. D. Schotten. Time is of the essence: Machine learning-based intrusion detection in industrial time series data. In *ICDM Workshops*, pages 1–6, 2018.

[6] D. A. Brown, P. R. Brady, A. Dietz, J. Cao, B. Johnson, and J. McNabb. A case study on the use of workflow technologies for scientific analysis: Gravitational wave data analysis. In *Workflows for e-Science*. 2007.

[7] E. Burnaev and V. Ishimtsev. Conformalized density-and distance-based anomaly detection in time-series data. *arXiv:1608.04585*, 2016.

[8] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[9] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good. The cost of doing science on the cloud: the montage example. In *ACM/IEEE conference on Supercomputing*, page 50, 2008.

[10] F. Doelitzscher, M. Knahl, C. Reich, and N. Clarke. Anomaly detection in iaas clouds. In *International Conference on Cloud Computing Technology and Science*, volume 1, pages 387–394, 2013.

[11] M. Du, F. Li, G. Zheng, and V. Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of ACM Conference on Computer and Communications Security*, 2017.

[12] D. Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one*, 4(5), 2009.

[13] P. Gaikwad, A. Mandal, P. Ruth, G. Juve, D. Król, and E. Deelman. Anomaly detection for scientific workflow applications on networked clouds. In *International Conference on High Performance Computing & Simulation*, pages 645–652, 2016.

[14] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, et al. myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(suppl 2):677–682, 2010.

[15] D. Gunter, E. Deelman, T. Samak, C. H. Brooks, M. Goode, G. Juve, G. Mehta, P. Moraes, F. Silva, M. Swany, et al. Online workflow management and performance analysis with stampede. In *International Conference on Network and Service Management*, pages 1–10, 2011.

[16] A. Kashlev and S. Lu. A system architecture for running big data workflows in the cloud. In *2014 IEEE International Conference on Services Computing*, pages 51–58. IEEE, 2014.

[17] M. Malawski, G. Juve, E. Deelman, and J. Nabrzyski. Cost-and deadline-constrained provisioning for scientific workflow ensembles in iaas clouds in: Proceedings of the international conference on high performance computing, networking, storage and analysis, 22, 2012.

[18] A. Navaz, V. Sangeetha, and C. Prabhadevi. Entropy based anomaly detection system to prevent ddos attacks in cloud. *arXiv:1308.6745*, 2013.

[19] H. S. Pannu, J. Liu, and S. Fu. A self-evolving anomaly detection framework for developing highly dependable utility clouds. In *IEEE GLOBECOM*, pages 1605–1610, 2012.

[20] M. A. Rodriguez, R. Kotagiri, and R. Buyya. Detecting performance anomalies in scientific workflows using hierarchical temporal memory. *Future Generation Computer Systems*, 88:624–635, 2018.

[21] T. Samak, D. Gunter, M. Goode, E. Deelman, G. Juve, G. Mehta, F. Silva, and K. Vahi. Online fault and anomaly detection for large-scale scientific workflows. In *International Conference on High Performance Computing and Communications*, pages 373–381, 2011.

[22] T. Samak, D. Gunter, M. Goode, E. Deelman, G. Juve, F. Silva, and K. Vahi. Failure analysis of distributed scientific workflows executing in the cloud. In *8th international conference on network and service management*, pages 46–54, 2012.

[23] M. Schneider, W. Ertel, and F. Ramos. Expected similarity estimation for large-scale batch and streaming anomaly detection. *Machine Learning*, 105(3):305–333, 2016.

[24] Y. Tan et al. Online performance anomaly prediction and prevention for complex distributed systems. 2012.

[25] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, and K. Schwan. Statistical techniques for online anomaly detection in data centers. In *Integrated Network Management*, pages 385–392, 2011.

[26] C.-C. M. Yeh, H. Van Herle, and E. Keogh. Matrix profile iii: the matrix profile allows visualization of salient subsequences in massive time series. In *ICDM*, pages 579–588, 2016.

[27] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *ICDM*, pages 1317–1322, 2016.

[28] L. Yu and Z. Lan. A scalable, non-parametric anomaly detection framework for hadoop. In *ACM Cloud and Autonomic Computing Conference*, page 22, 2013.

[29] Y. Zhu, Z. Zimmerman, N. S. Senobari, C.-C. M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh. Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins. In *ICDM*, pages 739–748, 2016.