Online Privacy-Preserving Data-Driven Network Anomaly Detection

Mehmet Necip Kurt, *Member, IEEE*, Yasin Yılmaz, *Senior Member, IEEE*, Xiaodong Wang, *Fellow, IEEE*, and Pieter J. Mosterman, *Member, IEEE*

***Abstract*— We study online privacy-preserving anomaly detection in a setting in which the data are distributed over a network and locally sensitive to each node, and a probabilistic data model is unknown. We design and analyze a data-driven solution scheme where each node observes a high-dimensional data stream for which it computes a local outlierness score. This score is then perturbed, encrypted, and sent to a network operator. The network operator then decrypts an aggregate statistic over the network and performs online network anomaly detection via the proposed generalized cumulative sum (CUSUM) algorithm. We derive an asymptotic lower bound and an asymptotic approximation for the average false alarm period of the proposed algorithm. Additionally, we derive an asymptotic upper bound and asymptotic approximation for the average detection delay of the proposed algorithm under a certain anomaly. We show the analytical tradeoff between the anomaly detection performance and the differential privacy level, controlled via the local perturbation noise. Experiments illustrate that the proposed algorithm offers a good tradeoff between privacy and quick anomaly detection against the UDP flooding and spam attacks in a real Internet of Things (IoT) network.**

***Index Terms*— Network anomaly detection, online, data-driven, distributed differential privacy, privacy-anomaly detection tradeoff.**

I. Introduction

IN REAL-TIME monitoring of safety-critical systems, anomalies should be quickly identified for a timely response [1], [2]. Moreover, in distributed networks where each node/user/device has privacy-sensitive data, data-driven statistical inference should not violate the confidentiality of data providers [3], [4]. Further, since real-world data often exhibit arbitrary statistical characteristics [5], [6], this study aims to develop an effective online privacy-preserving network anomaly detection scheme that is free of data model assumptions and hence applicable to a variety of real-world networks such as vehicular [7], power (i.e., smart grid) [8], Internet of Things (IoT) [9], and cellular networks [10]. For example, in distribution smart grids, smart meters are subject to false data injection (FDI) attacks [11]. The network operator needs early detection of the FDI attacks for the security and reliability of the network, however, raw smart meter data reveal privacy-sensitive user electricity consumption patterns [8]. This study proposes a generic solution that can be used to timely detect FDI attacks in the smart grid while effectively maintaining the privacy of the smart meter data of users.

A. Background

To process privacy-sensitive data, various techniques have been developed such as homomorphic encryption, secure multi-party computation, federated learning, and differential privacy (DP) [4], [12]. Homomorphic encryption [13] transforms the data such that arithmetic operations performed on the encrypted data correspond to the same arithmetic operations on the raw data. This enables processing encrypted data directly without having access to the raw data. In a distributed setting, this method requires a coordination among multiple parties, particularly a centralized key management authority. Moreover, homomorphic encryption is computationally intensive, which might limit its practical use. Alternatively, secure multi-party computation [14] enables a set of nodes to compute a desired function collaboratively without revealing their own raw data. Although promising, this method also requires a coordination among nodes via peer-to-peer communication, which might be costly in practice, especially over large-scale networks.

Federated learning is an iterative distributed learning procedure where a set of nodes is coordinated by a central node with the goal of learning a common model in a privacy-preserving manner [15]. The central node sends the latest model to the nodes and each node feeds back a model update computed through its own data. The central node then updates the model by fusing the local updates. In [15], a deep neural network model is trained where at each iteration the nodes compute the local gradients and the central node updates the neural network weights via the average of local gradients. In terms of privacy, the main concern is that the local updates may still reveal privacy-sensitive information to the central node [16].

DP [3] is a probabilistic framework based on the notion of indistinguishability. In particular, observing the output of

Manuscript received February 27, 2021; revised November 1, 2021; accepted December 21, 2021. Date of publication January 21, 2022; date of current version February 17, 2022. This work was supported in part by the U.S. National Science Foundation under Grant ECCS-2040500 and Grant ECCS-2040572. *(Corresponding author: Mehmet Necip Kurt.)*

Mehmet Necip Kurt was with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: m.n.kurt@columbia.edu).

Yasin Yılmaz is with the Department of Electrical Engineering, University of South Florida, Tampa, FL 33620 USA (e-mail: yasiny@usf.edu).

Xiaodong Wang is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: wangx@ee.columbia.edu).

Pieter J. Mosterman is with MathWorks, Natick, MA 01760 USA (e-mail: pmosterm@mathworks.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3142302.

Digital Object Identifier 10.1109/JSAC.2022.3142302

a differentially private algorithm, one cannot infer if any specific node/user/device contributed to the data. This ensures roughly the same level of privacy to each data provider. More specifically, change or removal of the data of any single node does not significantly alter the output likelihood. In this framework, privacy is achieved by randomizing the released statistics from a database, where the worst-case privacy risk can be quantified and calibrated with the level of randomization. The randomization can be achieved in many manners such as input perturbation (via additive noise), output perturbation, objective function perturbation, and exponential selection mechanism [4].

B. Related Work

There has been a growing research interest in differentially private machine learning and signal processing [4], [17]–[24]. For a differentially private algorithm, mainly the effect of privacy constraints on the algorithm performance should be analyzed, which allows the computation of achievable performance given the desired privacy level. The DP literature commonly presumes a centralized trusted data collector that has access to all the raw data and releases statistics from this database privately.

Albeit to a lesser extent, distributed DP has been also studied. In [25], distributed implementation of privacy-preserving databases through distributed noise generation has been studied where each node generates a share of overall random noise. This scheme requires cooperation and coordination among nodes. In [12], privacy-preserving aggregation of discrete time-series data is studied and a differentially private stream aggregation algorithm is presented, where a group of nodes periodically send encrypted randomized messages to an untrusted data aggregator. This mechanism achieves aggregator obliviousness, that is, the aggregator is not able to obtain any unintended information about the individual nodes other than the intended aggregate statistic over the network. Similarly, in [26] a private stream aggregation algorithm is presented where, different from [12], neither coordination among nodes nor a centralized key management authority is required to achieve the aggregator obliviousness.

Privacy-preserving change and anomaly detection have also been studied recently. A subset of these studies claim practical privacy benefits without rigorous privacy analysis. Such algorithms can be mainly motivated by the data processing inequality, that is, the processed or transformed form of data carry less information than the raw data. For instance, in [27] a privacy-preserving anomaly detection scheme is presented where the raw sensitive data is firstly transformed for privacy concerns and then an anomaly detection algorithm is employed based on the Gaussian mixture model and the Kalman filter, however, no theoretical privacy analysis is provided.

Federated learning-based anomaly detection algorithms [28]–[35] also provide practical privacy benefits as they only allow the exchange of model updates while keeping the raw data private. However, local model updates are still subject to leakage of privacy-sensitive information [16]. In [28], a federated learning-based algorithm is proposed, which depends on a local gated recurrent unit (GRU) model at every node and a global GRU model at the network center. The network center aggregates the local model weights shared by the nodes and updates the global model for attack detection and classification in IoT networks. In [29], a network anomaly detection algorithm is proposed based on the federated learning and transfer learning mainly against data scarcity, which is motivated by the fact that deep learning models usually require a large dataset for training. In [30], a distributed self-learning system is proposed for the detection of compromised devices (e.g., by the Mirai malware) in an IoT network. The algorithm detects anomalous communication behaviors of IoT devices and uses federated learning to aggregate anomalous behavior profiles privately. In [32], a federated learning-based intrusion detection scheme is proposed based on the convolutional neural network (CNN) and the GRU models along with a secure communication protocol using the homomorphic encryption. The secure communication protocol requires a centralized key management authority and the communication overhead associated with this protocol might be prohibitive for a real-time implementation.

Another subset of studies provide provable privacy guarantees, particularly DP, but with restrictive assumptions such as fully-known data models or bounded log-likelihood ratios. For instance, in [21] a window-based differentially private variant of the cumulative sum (CUSUM) algorithm is presented in a setting where the pre- and post-change data models are known. In this method, the log-likelihood ratio is perturbed to provide a private estimate of the change-point. It is a centralized detection method as the entire raw data is assumed to be available at the decision maker. Similarly, in [22] the differentially private hypothesis testing problem is studied in a model-based centralized setting and a window-based private change detection algorithm is presented.

Finally, in [24] a differentially private distributed intrusion detection scheme is proposed for vehicular networks based on the alternating direction method of multipliers. In this method, vehicles collaborate via vehicle-to-vehicle communications to train a network-wide classifier for attacks or intrusions while protecting the privacy of training data.

C. Contributions

We propose a generic privacy-preserving data-driven online network anomaly detection algorithm. Our main contributions are in analyzing the performance of the proposed algorithm. In particular,

- We derive an asymptotic approximation and an asymptotic lower bound for the average false alarm period (FAP) of the proposed algorithm.
- We derive an asymptotic approximation and an asymptotic upper bound for the average detection delay (ADD) of the proposed algorithm under a certain given anomaly. Additionally, we derive the worst-case asymptotic upper bound on the ADD of the proposed algorithm without needing any knowledge about the anomaly.
- We show the analytical tradeoff between the DP level and the network anomaly detection performance, where the

TABLE I
COMMON SYMBOLS AND PARAMETERS

Symbol	Definition
τ	Change-point
Γ	Stopping time
m_n	Data dimensionality at node n
W_n	Size of the historical dataset at node n
N	Network size
σ	Standard deviation of local Gaussian perturbation noise
θ	$\sigma/\sqrt{N}$
η	Detector sensitivity
ρ	η/θ
γ	Mean decrease in case of an anomaly
h	Test threshold
ϵ, δ	Differential privacy parameters

tradeoff is controlled via the variance of local perturbation noise.

D. Organization

The remainder of the paper is organized as follows. Section II describes the problem. Section III explains our solution approach. Section IV presents analysis of the proposed solution. Section V presents numerical evaluation of the proposed solution over an IoT network. Finally, Section VI concludes the paper. Boldface letters denote vectors and matrices, all vectors are column vectors, and $\cdot^{\mathrm{T}}$ denotes the transpose operator. $\mathbb{P}$ and $\mathbb{E}$ denote the probability and the expectation operators, respectively. Table I summarizes the common symbols and parameters in the paper.

II. PROBLEM DESCRIPTION

Consider a distributed network with N nodes and each node $n \in \{1, 2, \ldots, N\}$ has a high-dimensional observation $\mathbf{x}_{t,n} \in \mathbb{R}^{m_n}$ at time $t \geq 1$, where $m_n \gg 1$ denotes the data dimensionality of node n. At an unknown time τ, called the change-point, an anomaly happens over the network, such as a cyber attack, a cyberphysical attack, or a random fault, and the network deviates from its nominal operation. Anomalies can be attributed to a change in the statistical properties of the data generating process and hence for the network data $\mathbf{x}_t \triangleq [\mathbf{x}_{t,1}^{\mathrm{T}}, \mathbf{x}_{t,2}^{\mathrm{T}}, \ldots, \mathbf{x}_{t,N}^{\mathrm{T}}]^{\mathrm{T}}$, we can write

$$\mathbf{x}_t \sim \begin{cases} f_0^{\mathbf{x}}, & \text{if } t < \tau, \\ f_1^{\mathbf{x}} \neq f_0^{\mathbf{x}}, & \text{if } t \geq \tau, \end{cases}$$

where $f_0^{\mathbf{x}}$ denotes the probability density function (pdf) of $\mathbf{x}_t$ under nominal conditions and $f_1^{\mathbf{x}}$ denotes the pdf of $\mathbf{x}_t$ after the anomaly.

Our goal is to detect network anomalies timely and reliably based on the observed data sequence, corresponding to a sequential change detection problem [1], [36], [37], where at each time t, after new observations, a decision is made: either a change (anomaly) is declared or it is continued to observe more data in the next time interval. The design goal is to detect the changes as quickly as possible while limiting the risk of false alarm.

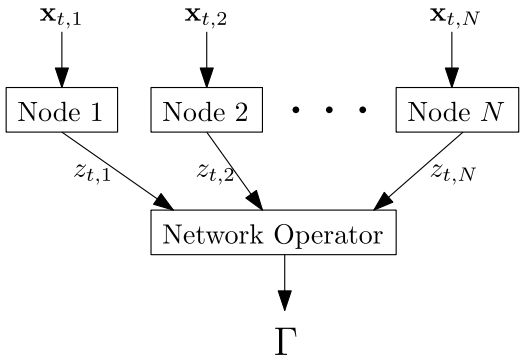


Fig. 1. A graphical description of the problem.

The Lorden's minimax formulation aims to minimize the worst-case ADD subject to a lower bound on the FAP [38]. In particular, letting Γ be the stopping time at which a change is declared and $J(\Gamma)$ be the worst-case ADD, given by

$$J(\Gamma) \triangleq \sup_{\tau} \operatorname*{ess\,sup}_{\mathcal{F}_\tau} \mathbb{E}_\tau\big[(\Gamma - \tau)^+ \,|\mathcal{F}_\tau\,\big],$$

where $(\cdot)^+ = \max\{0, \cdot\}$, $\mathcal{F}_\tau$ denotes the set of observations up to the change-point τ, and $\operatorname{ess\,sup}$ denotes the essential supremum, the minimax problem can be stated as [38]

$$\inf_{\Gamma} \; J(\Gamma) \quad \text{subject to } \mathbb{E}_\infty[\Gamma] \geq \zeta, \tag{1}$$

where $\mathbb{E}_\infty[\Gamma]$ denotes the FAP and ζ denotes the desired lower bound on the FAP.

If the data models $f_0^{\mathbf{x}}$ and $f_1^{\mathbf{x}}$ are known and the network data $\{\mathbf{x}_t\}_t$ are fully available at a decision maker, the CUSUM algorithm is the optimal solution to the minimax problem in Eq. (1) [39]. If the data models are known except for some unknown fixed parameters, the generalized likelihood ratio (GLR) test, making use of the estimates of unknown parameters, has asymptotic optimality properties [1, Sec. 5.3]. However, for high-dimensional real-world data streams, usually the nominal pdf $f_0^{\mathbf{x}}$ is difficult to model or intractable to estimate. For example, in a large-scale real network it is difficult to model complex interactions between nodes. Moreover, the anomalous pdf $f_1^{\mathbf{x}}$ might take arbitrary unknown and time-varying forms depending on the type and cause of anomalies [40], [41]. Hence, in this study, we assume both $f_0^{\mathbf{x}}$ and $f_1^{\mathbf{x}}$ are unknown and we look for a model-free (i.e., data-driven) solution.

In distributed networks with privacy-sensitive data, data-driven change/anomaly detection procedure should not violate the data privacy. Basically, if the local data are sensitive to a node, no other entity should be allowed to access the raw local data. In this case, since the network anomaly detection is critical for the network security and reliability, the network operator may only request from nodes to disclose some minimal information aligned with the anomaly detection task. Considering such a setting, let every node n, based on its local data $\mathbf{x}_{t,n}$ at time t, send a univariate message $z_{t,n}$ to the network operator, as illustrated in Fig. 1. The decision maker

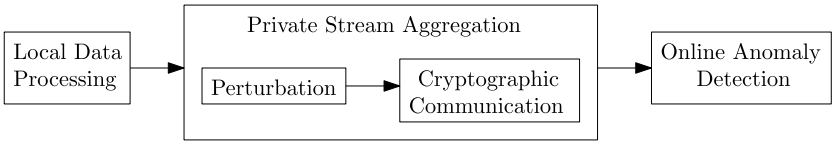


Fig. 2. An overview of the proposed solution approach.

(network operator) then receives $\mathbf{z}_t \triangleq [z_{t,1}, z_{t,2}, \ldots, z_{t,N}]$ from all nodes and decides on the anomaly based on the sequence of $\{\mathbf{z}_t\}_t$. Using this general architecture, we aim to design an effective solution scheme that achieves

- model-free online processing of the local data streams and disclosure of minimal task-oriented information from each node,
- differentially private aggregation of node messages at the network operator, and
- quick and reliable low-complexity network anomaly detection.

III. Solution Approach

Our solution consists of three functional modules: local data processing, private stream aggregation, and online anomaly detection (see Fig. 2). In the first module, sensitive data is analyzed and processed locally at the node it belongs to and some useful information is extracted for the anomaly detection task. In the second module, for DP, instead of releasing the extracted information directly, it is first perturbed via additive noise and then a form of cryptographic communication is utilized between the nodes and the network operator to ensure that the network operator can only decrypt an aggregate statistic over the network but not the individual node information. In the third module, the network operator performs online anomaly detection. The local data processing, perturbation, and encryption are carried out at the nodes while the decryption and online anomaly detection are performed at the network operator. Next, we explain these functional modules in more detail.

A. Local Data Processing

At time t, after acquiring a new sample $\mathbf{x}_{t,n}$, node n computes a local outlierness score $p_{t,n} \in [0,1]$ corresponding to $\mathbf{x}_{t,n}$. Since the data models are unknown, a model-free (i.e., nonparametric) method is used to compute the local outlierness score. In our solution, we describe a generic method, where each node can use a different (e.g., distance-based, neural networks-based, or subspace-based) method under this generic scheme based on the local data characteristics.

We assume each node n stores a local historical dataset $\mathcal{X}_n \triangleq \{\mathbf{x}_{1,n}, \mathbf{x}_{2,n}, \ldots, \mathbf{x}_{W_n,n}\}$ of size W_n consisting of nominal (anomaly-free) samples. The local data processing at node n consists of offline and online phases. In the offline phase, a set of useful univariate summary statistics is extracted from $\mathcal{X}_n$. These statistics are used to represent the nominal behavior of node n. In the online phase, when a new sample $\mathbf{x}_{t,n}$ is acquired, the online summary statistic corresponding to $\mathbf{x}_{t,n}$, denoted with $s_{t,n}$, is computed and the corresponding p-value estimate $p_{t,n}$ is computed as the local outlierness score based on how likely it is to observe $s_{t,n}$ under nominal conditions. As a useful and illustrative example, Algorithm 1 explains a subspace-based local data processing method based on the principal component analysis (PCA). We next briefly explain the principle behind the proposed local data processing method in Algorithm 1.

TABLE II
Space and Time Complexity of the PCA-Based Procedure at Node n

	Space	Time
Offline	$O\left(m_n^2 + W_n\right)$	$O\left(m_n^3 + W_n m_n^2 + W_n \log(W_n)\right)$
Online	$O\left(m_n\right)$	$O\left(m_n^2 + \log(W_n)\right)$

If the observed high-dimensional data stream at node n exhibits a low intrinsic dimensionality, we can write:

$$\mathbf{x}_{t,n} = \mathbf{y}_{t,n} + \mathbf{r}_{t,n},$$

where $\mathbf{y}_{t,n}$ is the representation of $\mathbf{x}_{t,n}$ in a lower-dimensional subspace and $\mathbf{r}_{t,n}$ is the residual term. The PCA is a well-known nonparametric method to learn linear manifolds [42, Sec. 12.1]. If the local nominal data can be well represented in a linear subspace, we can use the PCA to learn a nominal subspace for node n using $\mathcal{X}_n$. Since anomalous data are expected to deviate from the nominal subspace, the magnitude of the residual term, that is, $\|\mathbf{r}_{t,n}\|_2$, is expected to take higher values for anomalous data compared to the nominal data. We can hence use the magnitude of the residual term as a useful summary statistic (i.e., $s_{t,n} = \|\mathbf{r}_{t,n}\|_2$) to make a distinction between anomalous and nominal data. First, in an offline phase, we employ the PCA [42, Sec. 12.1] to determine a representative nominal subspace and compute a set of residual magnitudes. From this set of nominal summary statistics, we then form a nominal empirical distribution function (edf). In the online phase, we estimate the p-value $p_{t,n}$ corresponding to a new sample $\mathbf{x}_{t,n}$ based on the nominal edf of summary statistics, as summarized in Algorithm 1. The space and time complexity of Algorithm 1 is provided in Table II (for both offline and online phases).

The p-value estimate $p_{t,n}$ given in Algorithm 1 almost surely converges to the actual p-value as the sample size grows to infinity, that is, as $W_n \to \infty$. This is because the nominal edf formed by $\{\|\mathbf{r}_{i,n}\|_2 : \mathbf{x}_{i,n} \in \mathcal{X}_n\}$ pointwise almost surely converges to the nominal cumulative distribution function (cdf) as the sample size grows, by the Glivenko-Cantelli theorem [43]. Under nominal conditions (no anomaly), the p-value is uniformly distributed $\mathcal{U}[0,1]$. The p-value estimate $p_{t,n}$ is hence asymptotically (as $W_n \to \infty$) uniform for $t < \tau$. Moreover, the p-value is expected to take smaller

Algorithm 1 PCA-Based Procedure at Node n

Offline Phase

Compute the sample mean $\bar{\mathbf{x}}_n = \frac{1}{W_n}\sum_{\mathbf{x}_{i,n}\in\mathcal{X}_n}\mathbf{x}_{i,n}$.
Compute the sample data covariance matrix $\mathbf{Q}_n = \frac{1}{W_n}\sum_{\mathbf{x}_{i,n}\in\mathcal{X}_n}(\mathbf{x}_{i,n}-\bar{\mathbf{x}}_n)(\mathbf{x}_{i,n}-\bar{\mathbf{x}}_n)^{\mathrm{T}}$.
Compute the eigenvalues $\{\lambda_{i,n} : i = 1, 2, \ldots, m_n\}$ and the eigenvectors $\{\mathbf{v}_{i,n} : i = 1, 2, \ldots, m_n\}$ of $\mathbf{Q}_n$.
Determine the dimensionality of the submanifold, r_n, based on the desired fraction of data variance retained in the linear manifold, written by $\sum_{i=1}^{r_n}\lambda_{i,n}/\sum_{i=1}^{m_n}\lambda_{i,n}$, where $\lambda_{1,n}$, $\lambda_{2,n}$, ..., $\lambda_{r_n,n}$ denote the r_n largest eigenvalues $\mathbf{Q}_n$.
For a chosen dimensionality $r_n < m_n$, form the matrix $\mathbf{V}_n \triangleq [\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_{r_n}]$.
for $i : \mathbf{x}_{i,n} \in \mathcal{X}_n$ **do**
  $\mathbf{r}_{i,n} = (\mathbf{I}_{m_n} - \mathbf{V}_n\mathbf{V}_n^{\mathrm{T}})(\mathbf{x}_{i,n} - \bar{\mathbf{x}}_n)$.
  Compute $\|\mathbf{r}_{i,n}\|_2$.
end for
Sort the nominal summary statistics $\{\|\mathbf{r}_{i,n}\|_2 : \mathbf{x}_{i,n} \in \mathcal{X}_n\}$ in ascending order and store them in the local memory.

Online Phase

Initialization: $t \leftarrow 0$
while $t < \Gamma$ **do**
  $t \leftarrow t + 1$
  Acquire a new sample $\mathbf{x}_{t,n}$.
  $\mathbf{r}_{t,n} = (\mathbf{I}_{m_n} - \mathbf{V}_n\mathbf{V}_n^{\mathrm{T}})(\mathbf{x}_{t,n} - \bar{\mathbf{x}}_n)$ and compute $\|\mathbf{r}_{t,n}\|_2$.
  Compute the local outlierness score: $p_{t,n} = \frac{1}{W_n}\sum_{\mathbf{x}_{i,n}\in\mathcal{X}_n}\mathbb{1}\{\|\mathbf{r}_{i,n}\|_2 > \|\mathbf{r}_{t,n}\|_2\}$.
  Perturbation: $\tilde{p}_{t,n} = p_{t,n} + v_{t,n}$.
  Encryption: $z_{t,n} = \tilde{p}_{t,n} + k_{t,n}$.
  Send $z_{t,n}$ to the network operator and $k_{t,n}$ to the auxiliary node.
end while

(close to zero) values for anomalous samples, and hence we expect smaller $p_{t,n}$ for $t \geq \tau$.

The output of the local data processing module, that is, the p-value estimate, while useful to infer possible anomalies, is not quite as informative to interpret the raw local data. This is because, irrespective of the local data processing method and the observed data, the p-value estimate is always (asymptotically) a uniform random variable under nominal conditions. Hence, for network anomaly detection, on the one hand, releasing this task-oriented statistic from the nodes can be useful to preserve the data privacy. On the other hand, this does not provide any privacy guarantee as releasing such summary statistics is still subject to side information leakages and reconstruction attacks [3]. To further improve the data privacy and to provide provable privacy guarantees, we next present our private stream aggregation module.

B. Private Stream Aggregation

This module ensures that the network operator privately aggregates the local outlierness scores at nodes for a differentially private network anomaly detection.

1) Perturbation: Each node n perturbs its local outlierness score $p_{t,n}$ via additive white Gaussian noise (AWGN), i.i.d. over time and space (i.e., between the nodes). Denoting the perturbed statistic by $\tilde{p}_{t,n}$ yields

$$\tilde{p}_{t,n} = p_{t,n} + v_{t,n},$$

where $v_{t,n} \sim \mathcal{N}(0, \sigma^2)$ is zero-mean AWGN with variance σ^2. Higher σ^2 increases uncertainty of the released statistic. Note that each node employs the output perturbation for DP as the output of the local data processing is $p_{t,n}$ at each node n.

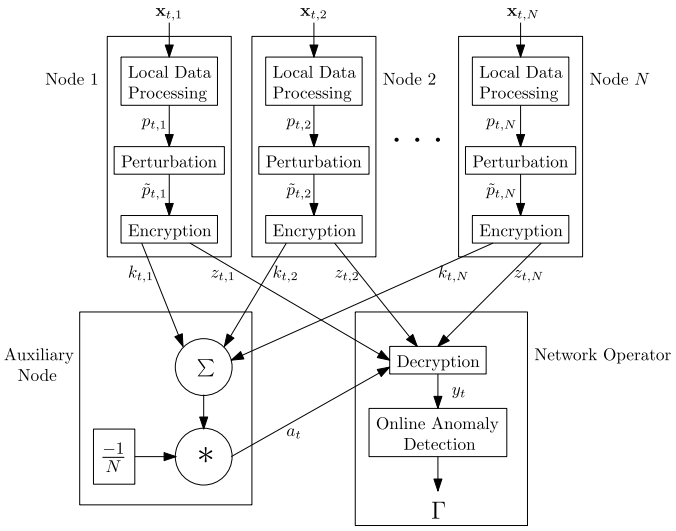


Fig. 3. A graphical illustration of the proposed solution scheme.

2) Cryptographic Communication: The purpose of cryptographic communication between nodes and the network operator is that the network operator can only decrypt the noisy mean of the local outlierness scores, given by

$$y_t \triangleq \frac{1}{N}\sum_{n=1}^{N}\tilde{p}_{t,n}. \tag{2}$$

To achieve this without coordination between nodes in both static and dynamic networks,[1] we make use of an auxiliary node [26], as illustrated in Fig. 3. In this mechanism, each node n generates a private key $k_{t,n} > 0$ at each time t and obtains the encrypted message to be sent to the network operator as follows:

$$z_{t,n} = \tilde{p}_{t,n} + k_{t,n}.$$

The key generation process is specific to each node, kept secret, and independent from the other nodes. Ideally, it should be difficult to track the generated key pattern and hence to infer $\tilde{p}_{t,n}$ from $z_{t,n}$. Moreover, the private keys should not carry any information relevant to the local data. The generated keys are also sent to an auxiliary node that computes the negative average of the received keys from the nodes:

$$a_t \triangleq -\frac{1}{N}\sum_{n=1}^{N}k_{t,n},$$

and sends the result to the network operator. The auxiliary node does not share any other information with the network

[1]In dynamic networks, new nodes can join and leave in time, which is especially relevant to the case of node failures and to networks dynamic in nature such as mobile and vehicular networks.

Algorithm 2 Procedure at the Auxiliary Node

1: Initialization: $t \leftarrow 0$
2: **while** $t < \Gamma$ **do**
3: $t \leftarrow t + 1$
4: Receive $\{k_{t,n}, n = 1, 2, \ldots, N\}$ from the nodes.
5: $a_t = -\frac{1}{N}\sum_{n=1}^{N} k_{t,n}$
6: Send a_t to the network operator.
7: **end while**

TABLE III
SPACE AND TIME COMPLEXITY OF THE PROPOSED PROCEDURE AT THE AUXILIARY NODE

	Space	Time
Online	$O(1)$	$O(N)$

operator or any other node. The proposed procedure at the auxiliary node is summarized in Algorithm 2. The auxiliary node is employed only in the online phase. The space and time complexity of Algorithm 2 is provided in Table III.

After receiving $\{z_{t,n}, n = 1, 2, \ldots, N\}$ from the nodes, and a_t from the auxiliary node, the network operator takes the average of the node messages, then sums the average with a_t, and obtains y_t, see Eq. (2):

$$y_t = a_t + \frac{1}{N}\sum_{n=1}^{N} z_{t,n} = a_t + \frac{1}{N}\sum_{n=1}^{N}(\tilde{p}_{t,n} + k_{t,n})$$
$$= \underbrace{a_t + \frac{1}{N}\sum_{n=1}^{N} k_{t,n}}_{0} + \frac{1}{N}\sum_{n=1}^{N}\tilde{p}_{t,n} = \frac{1}{N}\sum_{n=1}^{N}\tilde{p}_{t,n}.$$

Note that we assume ideal (i.e., error-free) communications within the network (i.e., between the nodes, the auxiliary node, and the network operator).

Remark 1: The sole aim of the cryptographic communication is that the network operator learns only an aggregate statistic over the network but nothing else about the individual nodes. This is useful to achieve distributed DP. The details of the key management and the security analysis of the cryptographic communication protocol are out of the scope of our paper. We refer interested readers to [12], [26], [44]–[48] for details.

Remark 2: An alternative cryptographic communication protocol can be designed with secret key sharing without needing the auxiliary node [26]. However, it requires coordination and peer-to-peer communication between nodes. Furthermore, it is not robust to node failures and dynamic networks, since it needs a redesign when any node joins or leaves.

C. Online Network Anomaly Detection

We first analyze the distribution of the aggregated statistic y_t at the network operator in both the nominal and anomaly cases. Next, we present the proposed online network anomaly detection method.

1) Distribution of y_t: The information aggregated at the network operator at time t can be rewritten as

$$y_t = \frac{1}{N}\sum_{i=1}^{N}\tilde{p}_{t,n} = \frac{1}{N}\sum_{i=1}^{N}(p_{t,n} + v_{t,n})$$
$$= \underbrace{\frac{1}{N}\sum_{i=1}^{N} p_{t,n}}_{\bar{p}_t} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} v_{t,n}}_{\bar{v}_t}$$
$$= \bar{p}_t + \bar{v}_t, \tag{3}$$

where $\bar{v}_t \sim \mathcal{N}(0, \sigma^2/N)$.

Recalling that $p_{t,n} \sim \mathcal{U}[0,1], \forall n \in \{1, 2, \ldots, N\}$ for $t < \tau$ and assuming $p_{t,n}$ is i.i.d. over time and space,[2] the central limit theorem yields, asymptotically

$$\bar{p}_t \sim \mathcal{N}\left(0.5, \frac{1}{12N}\right), \quad t < \tau. \tag{4}$$

Then, since $\bar{p}_t$ and $\bar{v}_t$ are independent, we can write

$$y_t \sim \mathcal{N}\left(0.5, \frac{\sigma^2 + 1/12}{N}\right), \quad t < \tau. \tag{5}$$

Further, if $\sigma^2 \gg 1/12$, it holds that approximately

$$y_t \sim \mathcal{N}(0.5, \sigma^2/N), \quad t < \tau. \tag{6}$$

In case of a network anomaly (i.e., for $t \geq \tau$), it is expected that anomalous nodes observe more frequent outliers, correspondingly smaller p-values, which leads to a decrease in the mean of y_t. Hence, we can argue that the mean of y_t is $0.5 - \gamma_t$ for $t \geq \tau$, where $\gamma_t \geq 0$ denotes the unknown and possibly time-varying mean decrease. Moreover, in case of an anomaly, we no longer have $p_{t,n} \sim \mathcal{U}[0,1]$. In fact, the pdf of $p_{t,n}$ is unknown for $t \geq \tau$. Nevertheless, $p_{t,n}$ is always between 0 and 1 and for a random variable taking values in this range, its variance can at most[3] be $1/4$. Hence, for $t \geq \tau$, $p_{t,n}$ has a mean $\mu_{t,n} \in [0, 0.5]$ and a variance $\sigma^2_{t,n} \in [0, 1/4]$, $\forall n \in \{1, 2, \ldots, N\}$.

The central limit theorem has variants, ensuring convergence to the normal distribution for non-identical or dependent distributions under certain conditions. In large-scale networks (i.e., large N), for $t \geq \tau$, we can consider $p_{t,n}$ as nearly independent but non-identically distributed across the nodes. Then, we can use the Lindeberg central limit theorem [49, p. 369] stating that given

$$s^2_{t,N} \triangleq \sum_{n=1}^{N}\sigma^2_{t,n},$$

[2]An i.i.d. $p_{t,n}$ stream can be achieved through local data processing. For example, in the PCA-based method, if the linear subspace well represents the local data, the residual term mostly corresponds to noise, that can be assumed i.i.d. over time and space. Furthermore, in large-scale networks, data of any single node can be assumed nearly independent from the vast majority of network nodes, except the node's immediate neighborhood.

[3]For a random variable $x \in [0,1]$, its variance can be written as $\sigma^2_x \triangleq \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \leq \mathbb{E}[x] - (\mathbb{E}[x])^2$, where the inequality is because $x^2 \leq x$ for $x \in [0,1]$. Denoting $m \triangleq \mathbb{E}[x]$, we have $\sigma^2_x \leq f(m) \triangleq m - m^2$ where $m \in [0,1]$. Since the maximum value of the function $f(m)$ is $1/4$ at $m = 1/2$, we have $\sigma^2_x \leq 1/4$.

if for every $\varepsilon > 0$, the condition

$$\lim_{N\to\infty} \frac{1}{s_{t,N}^2} \sum_{n=1}^{N} \mathrm{E}[(p_{t,n}-\mu_{t,n})^2 \, \mathbb{1}\{|p_{t,n}-\mu_{t,n}| > \varepsilon \, s_{t,N}\}] = 0 \tag{7}$$

is satisfied, then

$$\frac{1}{s_{t,N}} \sum_{n=1}^{N} (p_{t,n} - \mu_{t,n})$$

converges to the standard normal distribution $\mathcal{N}(0,1)$. Equivalently stating, under the condition above, it asymptotically holds that

$$\bar{p}_t = \frac{1}{N} \sum_{n=1}^{N} p_{t,n} \sim \mathcal{N}\left(\frac{\sum_{n=1}^{N} \mu_{t,n}}{N}, \frac{s_{t,N}^2}{N^2} \right).$$

Noticing further that

$$0 \le \sum_{n=1}^{N} \mu_{t,n} \le N/2,$$

and

$$0 \le s_{t,N}^2 \le N/4,$$

the mean of $\bar{p}_t$ is between 0 and 0.5 and the variance of $\bar{p}_t$ is between 0 and $\frac{1}{4N}$. Then, if $\sigma^2 \gg 1/4$, it approximately holds that, see Eq. (3),

$$y_t = \bar{p}_t + \bar{v}_t \sim \mathcal{N}(0.5 - \gamma_t, \sigma^2/N), \quad t \ge \tau. \tag{8}$$

Note that the condition in Eq. (7) is satisfied in our case, as the indicator in Eq. (7) tends to 0 as $N \to \infty$. This is because the term $|p_{t,n} - \mu_{t,n}|$ is bounded because of $p_{t,n}, \mu_{t,n} \in [0,1]$, whereas $s_{t,N} \to \infty$ as $N \to \infty$.

Finally, if $\sigma^2 \gg 1/4$ we can write, see Eq. (6) and Eq. (8),

$$y_t \sim \begin{cases} \mathcal{N}(0.5, \sigma^2/N), & \text{if } t < \tau, \\ \mathcal{N}(0.5 - \gamma_t, \sigma^2/N), & \text{if } t \ge \tau. \end{cases} \tag{9}$$

Notice that the high-dimensional network anomaly detection problem reduces to the sequential detection of a mean decrease over a univariate Gaussian data stream, see Eq. (9). Hence, at this point we can make a transition from the originally nonparametric setting to a parametric setting, as detailed next.

2) Online Anomaly Detection: Denoting the nominal and anomalous pdfs of y_t by f_0^y and f_1^y, respectively, and defining $\theta \triangleq \sigma/\sqrt{N}$, we have $f_0^y \sim \mathcal{N}(0.5, \theta^2)$ and $f_1^y \sim \mathcal{N}(0.5 - \gamma_t, \theta^2)$, where f_1^y has an unknown and possibly time-varying parameter γ_t, see Eq. (9). Then, we propose the following generalized CUSUM algorithm for online anomaly detection at the network operator:

$$\Gamma = \inf \Bigg\{ m \in \mathbb{N} : \underbrace{\max_{1\le j\le m} \sum_{t=j}^{m} \underbrace{\log \frac{\sup_{\gamma_t \ge \eta} f_1^y(y_t \,|\, \gamma_t)}{f_0^y(y_t)}}_{\beta_t}}_{g_m} \ge h \Bigg\}, \tag{10}$$

Algorithm 3 Procedure at the Network Operator

1: Initialization: $t \leftarrow 0$, $g_0 \leftarrow 0$
2: **while** $g_t < h$ **do**
3: $t \leftarrow t + 1$
4: Receive $\{z_{t,n}, n = 1, 2, \ldots, N\}$ from the nodes, and a_t from the auxiliary node.
5: $y_t = a_t + \frac{1}{N} \sum_{n=1}^{N} z_{t,n}$
6: Compute the GLLR β_t using Eq. (12).
7: $g_t \leftarrow (g_{t-1} + \beta_t)^+$
8: **end while**
9: $\Gamma \leftarrow t$, declare a network anomaly.

TABLE IV
SPACE AND TIME COMPLEXITY OF THE PROPOSED PROCEDURE AT THE NETWORK OPERATOR

	Space	Time
Online	$O(1)$	$O(N)$

where η denotes the minimum change of interest, indicating the detector sensitivity, and h denotes the test threshold. Further, β_t and g_t denote the generalized log-likelihood ratio (GLLR) and the decision statistic at time t, respectively, where the decision statistic can be written in the following recursive form, see [1, Sec. 2.2]:

$$g_t = (g_{t-1} + \beta_t)^+. \tag{11}$$

Moreover, the GLLR β_t can be computed as follows:

$$\begin{aligned} \beta_t &= \frac{1}{2\theta^2} \sup_{\gamma_t \ge \eta} (1 - 2\, y_t)\gamma_t - \gamma_t^2 \\ &= \begin{cases} \dfrac{1}{2\theta^2}(0.5 - y_t)^2, & \text{if } y_t \le 0.5 - \eta, \\ \dfrac{1}{2\theta^2}(1 - 2\, y_t)\eta - \dfrac{\eta^2}{2\theta^2}, & \text{if } y_t > 0.5 - \eta. \end{cases} \end{aligned} \tag{12}$$

Note that η is only a detector parameter and not part of an anomaly model. Between the conventional CUSUM algorithm [1, Sec. 2.2] and the proposed algorithm, the only difference is the online estimation of γ_t. We summarize the proposed procedure at the network operator in Algorithm 3. The network operator is employed only in the online phase. The space and time complexity of Algorithm 3 is provided in Table IV.

Remark 3: The proposed generalized CUSUM algorithm is different from the well-known GLR test [1, Sec. 5.3]. In the GLR test, the unknown pdf parameters are assumed to be fixed over time and the GLR test cannot be written in a recursive form. In the proposed algorithm, the unknown pdf parameter γ_t is assumed to be time-varying because of unknown anomaly and it is estimated at each time t separately. The proposed algorithm can thus be written in a recursive form.

IV. ANALYSIS

In this section, we first analyze the DP of the proposed algorithm. We next analyze the anomaly detection performance, particularly the FAP and the ADD. Finally, we characterize the effect of DP constraints on the anomaly detection performance.

A. Differential Privacy (DP)

First, we state the definition of (ϵ, δ)-DP below.

Definition: Let $\epsilon > 0$, $\delta < 1$, and $\phi(\cdot)$ be a randomized function taking a dataset as its input. Moreover, let $\text{im}(\phi)$ denote the image of ϕ, that is, the set of all output values it can take. The function $\phi(\cdot)$ is (ϵ, δ)-differentially private if

$$\mathbb{P}(\phi(D_1) \in \mathcal{S}) \leq e^{\epsilon}\, \mathbb{P}(\phi(D_2) \in \mathcal{S}) + \delta$$

for all datasets D_1 and D_2 differing in only a single entry and over all subsets $\mathcal{S}$ of $\text{im}(\phi)$.

The definition above mainly states that the outcome of a differentially private function does not vary significantly if any single entry in the database is changed, where the amount of significance is captured with ϵ and δ parameters. Here, ϵ and δ represent the worst-case privacy loss where lower values of them imply stronger privacy guarantees. Furthermore, if $\delta = 0$, it is called ϵ-DP.

Next, we state the definition of sensitivity of a function below.

Definition: Let D be a collection of datasets and d be a positive integer. The sensitivity of a function $\phi : D \to \mathbb{R}^d$ is defined by

$$\Delta\phi \triangleq \max_{\text{dist}(D_1, D_2)=1} \|\phi(D_1) - \phi(D_2)\|_1,$$

over all datasets D_1 and D_2 in D differing in a single entry, denoted by $\text{dist}(D_1, D_2) = 1$.

Computing the sensitivity of a function enables easy calibration of the level of randomization to achieve the desired privacy level. A common randomization technique is perturbation via additive noise. Particularly, the Gaussian mechanism achieves the DP by adding Gaussian noise to the output of a function operating on a database. The following lemma is useful to calibrate the amount of Gaussian perturbation noise to obtain (ϵ, δ)-DP [3]:

Lemma 1: Let the information released from a database D be

$$\tilde{\phi}(D) = \phi(D) + \omega_t,$$

where ω_t is the perturbation noise. If

$$\omega_t \sim \mathcal{N}\left(0, \frac{\Delta^2\phi\; 2\log(1.25/\delta)}{\epsilon^2}\right),$$

then (ϵ, δ)-DP is achieved.

As we observe through Lemma 1, a stronger privacy level, equivalently lower ϵ and/or δ, requires a higher level of perturbation noise. This is intuitive as the noise variance increases uncertainty of the released information from the database. The following theorem specifies the variance of the local perturbation noise at each node so that the proposed online network anomaly detection scheme achieves the (ϵ, δ)-DP.

Theorem 1: If at each node $n \in \{1, 2, \ldots, N\}$, the variance of the perturbation noise $v_{t,n}$ is set as

$$\sigma^2 = \frac{2\log(1.25/\delta)}{N\epsilon^2}, \tag{13}$$

then the proposed anomaly detection scheme is (ϵ, δ)-differentially private.

Proof: The proof of Theorem 1 is based on the parallel composition rule [50] and the post-processing invariance rule of DP [4]. See Appendix A. □

Theorem 1 gives the variance of local perturbation noise to achieve the desired DP level. It further shows the relationship between the network size N and the required noise level. Particularly, for a smaller network (smaller N), to achieve the same level of DP, we need to add a higher level of noise (higher σ^2). This is intuitive as for a small network consisting of a few nodes, it is usually more difficult to mask individual node information via an aggregate statistic, compared to masking in a large-scale network. Finally, since the desired DP level can be achieved by calibrating σ^2, we hereafter refer to σ^2 (equivalently θ^2) as the DP parameter.

Remark 4: In addition to being task-oriented (see Sec. III-A), another advantage of using the p-value as the output of the local data processing module is that the p-value is always bounded in the range of $[0, 1]$. We use this property to prove the DP of the proposed scheme (see Appendix A).

B. Anomaly Detection Performance

The average run length (ARL), denoted by $\mathbb{E}_\tau[\Gamma]$, is the first time, on average, that the decision statistic g_t exceeds the decision threshold h of the proposed detector (see Algorithm 3). The ARL can be used for both FAP and ADD computations. In particular, if there is no anomaly at all ($\tau = \infty$), the ARL corresponds to the FAP, that is, $\mathbb{E}_\infty[\Gamma]$. Furthermore, setting $\tau = 1$, the ARL corresponds to the worst-case ADD, $\mathbb{E}_1[\Gamma]$. This is because the proposed detector is a CUSUM-type detector for which the decision statistic being zero just before the change-point essentially describes the worst-case scenario in terms of detection delay. In other words, for any $z \geq 0$ with $g_{\tau-1} = z$, and expressing the ADD as a function of z, we have $\text{ADD}(z) \leq \text{ADD}(0)$ for the proposed detector. Note that for $\tau = 1$, we always have $g_{\tau-1} = 0$ since $g_0 = 0$.

In the following, we derive the Wald's approximations for both the FAP and the worst-case ADD of the proposed detector as well as performance bounds, particularly a lower bound on the FAP and an upper bound on the ADD to obtain performance guarantees. Note that our approximations and bounds are asymptotic as we use both the central limit theorems and the Glivenko-Cantelli theorem to asymptotically characterize the distribution of the aggregated statistic at the network operator, that is, $\{y_t\}_t$, (see Sec. III-C.1).

1) Average False Alarm Period (FAP): For reliable anomaly detection, false alarm events should be infrequent, equivalently the FAP should be large. For a given set of system and algorithm parameters, the following theorem gives the Wald's approximation for the FAP of the proposed detector. This approximation is useful to choose the system and algorithm parameters to control the false alarm rate of the proposed detector.

Theorem 2: Let $\rho \triangleq \eta/\theta$. If $\rho > 0.61$, then the Wald's approximation to the FAP is given by

$$\mathbb{E}_\infty[\Gamma] \approx \frac{2h + 2(e^{-w_0 h} - 1)/w_0}{Q(\rho) - \rho^2\, Q(-\rho)}$$

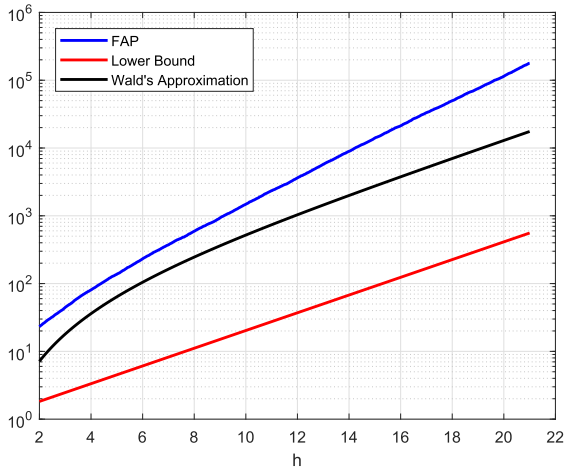


Fig. 4. The FAP of the proposed detector, the theoretical lower bound on the FAP, and the Wald's approximation to the FAP for various test thresholds h.

where $Q(\cdot)$ denotes the Q-function, that is, $Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2}\, du$, and $-1 < w_0 < 0$ is the unique solution to

$$f(w_0) \triangleq \frac{1}{\sqrt{w_0+1}} Q(\rho) + Q(-\rho)\, e^{0.5\,\rho^2 (w_0 + w_0^2)} = 1. \quad (14)$$

Proof: See Appendix A. □

In Theorem 2, w_0 is computed by solving Eq. (14) numerically [51], which is straightforward since we look for a unique w_0 in the range of $(-1, 0)$ for which $f(w_0) = 1$. Next, the following theorem provides a lower bound on the FAP of the proposed detector. A lower bound on the FAP is equivalent to an upper bound on the false alarm rate. Hence, Theorem 3 can be used to choose the system and algorithm parameters to limit the false alarm rate of the proposed detector with a desired value.

Theorem 3: If $\rho > 0.61$, a lower bound for the FAP is given by

$$\mathbb{E}_\infty[\Gamma] \geq e^{-w_0 h},$$

where w_0 is as given in Eq. (14).

Proof: See Appendix B. □

Next, to illustrate Theorem 2 and Theorem 3, we assume $\tau = \infty$, $y_t \sim \mathcal{N}(0.5, \theta^2), t \geq 1$ and plot in Fig. 4 the actual FAP of the proposed detector as well as the derived lower bound and the Wald's approximation as the test threshold h varies. In this experiment, we choose $\eta = 0.06$ and $\theta = 0.08$. We observe through Fig. 4 that the Wald's approximation underestimates the FAP. Note that this is because the Wald's approximation ignores the excess over the boundary for the random walk driven by the GLLR β_t, see Eq. (11), and in the nominal case usually the lower threshold 0 is exceeded frequently during the random walk. Nevertheless, the derived approximation can still be useful, at least to achieve a performance guarantee (i.e., a lower bound on the FAP) in most cases.

2) Average Detection Delay (ADD): For quick anomaly detection, detection delay should be small. Hence, the ADD analysis is useful to evaluate the expected performance of a detector. However, since anomalies can happen because of many reasons, it is practically hard to fully specify the anomaly. Specifically, for our detector, it is difficult to characterize the mean decrease γ_t for $t \geq \tau$. But, in a special case where the mean decrease is constant after the change-point, that is,

$$\gamma_t = \gamma, \quad t \geq \tau,$$

we can provide an analysis for the worst-case ADD of our detector. The following theorem states the Wald's approximation to the worst-case ADD of the proposed detector under this special condition.

Theorem 4: The Wald's approximation to the worst-case ADD is given by

$$\mathbb{E}_1[\Gamma] \approx \frac{h + \frac{e^{-w_1 h} - 1}{w_1}}{\frac{\gamma^2+\theta^2}{2\theta^2} Q\left(\frac{\eta-\gamma}{\theta}\right) + \frac{2\eta\gamma-\eta^2}{2\theta^2} Q\left(\frac{\gamma-\eta}{\theta}\right)},$$

where $w_1 > 0$ is the unique solution to

$$g(w_1) \triangleq Q\left(\frac{\eta-\gamma}{\theta}\right) \frac{e^{\frac{-w_1\gamma^2}{2\theta^2(w_1+1)}}}{\sqrt{w_1+1}} + Q\left(\frac{\gamma-\eta}{\theta}\right) e^{\frac{(\gamma^2-2\gamma\eta)w_1 + \gamma^2 w_1^2}{2\theta^2}} = 1. \quad (15)$$

Proof: See Appendix C. □

As before, we can compute w_1 by solving Eq. (15) numerically. An upper bound on the ADD provides a performance guarantee. Next, given the system and algorithm parameters, the following theorem provides a theoretical upper bound on the ADD of the proposed detector.

Theorem 5: If $\gamma > \eta/2$, an upper bound on ADD is given by

$$\mathbb{E}_1[\Gamma] \leq \frac{h + Q\left(\frac{\eta-\gamma}{\theta}\right) \frac{\gamma^2+\theta^2}{2\theta^2} + Q\left(\frac{\gamma-\eta}{\theta}\right) \psi(a,b)}{\frac{\gamma^2+\theta^2}{2\theta^2} Q\left(\frac{\eta-\gamma}{\theta}\right) + \frac{2\eta\gamma-\eta^2}{2\theta^2} Q\left(\frac{\gamma-\eta}{\theta}\right)},$$

where

$$\psi(a,b) \triangleq a + \frac{\sqrt{b}\, e^{\frac{-a^2}{2b}}}{\sqrt{2\pi}\, Q(-a/\sqrt{b})},$$

$a \triangleq (2\gamma\eta - \eta^2)/(2\theta^2)$, and $b \triangleq \eta^2/\theta^2$.

Proof: See Appendix D. □

Recall that our derivation for the theoretical upper bound on the ADD is based on the assumption that the post-change mean decrease satisfies $\gamma_t = \gamma$ for $t \geq \tau$. Using the result in Theorem 5, we can also derive the worst possible upper bound on the ADD without needing this assumption. Particularly, we can use the fact that the Kullback-Leibler (KL) divergence is a measure of separability between the pre- and post-change data distributions and hence a measure of detectability in the sequential detection theory such that as the KL divergence between the pre- and post-change pdfs decreases, then the ADD increases. Hence, if $\gamma_t, t \geq \tau$ takes values minimizing the KL divergence between the pre- and post-change pdfs, then the ADD is maximized. Given that our generalized CUSUM detector is designed with the detector sensitivity parameter η such that $\gamma_t \geq \eta, t \geq \tau$, see Eq. (10), then $\gamma_t = \eta, t \geq \tau$, minimizes the KL divergence between pre- and post-change pdfs, see Eq. (9), and hence maximizes the ADD of

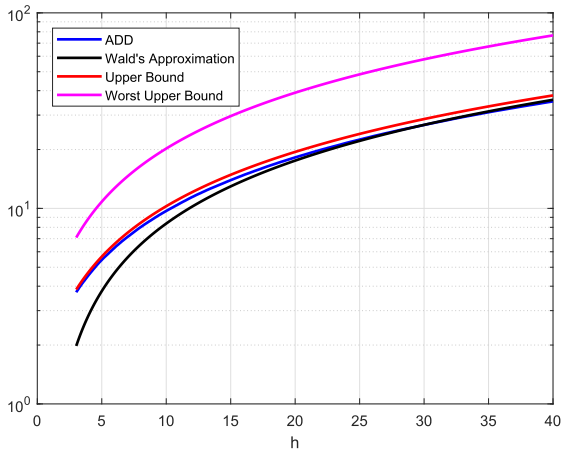


Fig. 5. The worst-case ADD of the proposed detector, the Wald's approximation to the ADD, the theoretical upper bound on the ADD, and the worst-case upper bound on the ADD for various levels of test threshold h.

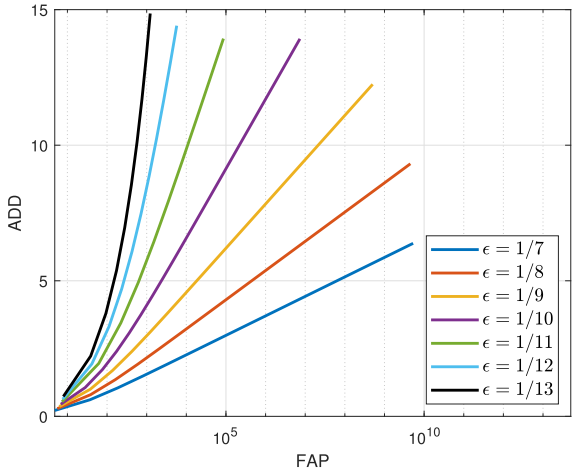


Fig. 6. The Wald's approximation to the worst-case ADD vs. the Wald's approximation to the FAP of the proposed detector as the test threshold h varies, for various DP levels.

our detector. Then, we derive the worst-case upper bound on the ADD irrespective of the unknown and time-varying γ_t by replacing γ in Theorem 5 with η, as presented in the Corollary 1 below.

Corollary 1: The worst-case upper bound on ADD is given by

$$\mathbb{E}_1[\Gamma] \leq \frac{2h + a^* + 0.5 + \psi(a^*, b^*)}{b^* + 0.5},$$

where $a^* \triangleq \eta^2/(2\theta^2)$, and $b^* \triangleq \eta^2/\theta^2$.

To illustrate Theorem 4, Theorem 5, and Corollary 1, we assume $\tau = 1$, $y_t \sim \mathcal{N}(0.5 - \gamma, \theta^2), t \geq \tau$, $\gamma = 0.1$, $\theta = 0.08$, and $\eta = 0.06$, and we plot in Fig. 5 the ADD of the proposed detector as well as the presented upper bounds and the approximation for the ADD, as the test threshold h varies. As we observe through Fig. 5, the Wald's approximation well approximates the ADD as the excess over the lower boundary 0 does not frequently happen in case of an anomaly, unlike the nominal case leading to an underestimation of the FAP, as discussed in Sec. IV-B.1.

Finally, note that there is an inherent tradeoff between the ADD and the FAP and this is controlled by the decision threshold h and the detector sensitivity parameter η. Particularly, increasing h and/or η leads to a larger FAP (i.e., lower false alarm rate) but also a larger ADD, and vice versa.

C. Analytical Privacy-Anomaly Detection Tradeoff

We next present the tradeoff between the DP level and the anomaly detection performance based on the theoretical results derived in Sec. IV-A and Sec. IV-B. If the proposed anomaly detection scheme is employed in security applications such as attack detection or fraud detection over safety-critical networks, this tradeoff can also be termed as the privacy-security tradeoff. Recall that for any chosen system and algorithm parameters including the DP parameters ϵ and δ, we provide theoretical approximations to both the FAP and ADD of our network anomaly detector. We can hence obtain the analytical privacy-anomaly detection tradeoff based on the Wald's approximations to the FAP and ADD (see Theorem 2 and Theorem 4).

For an illustration, consider a simple example where

$$y_t \sim \begin{cases} \mathcal{N}(0.5, \theta^2), & \text{if } t < \tau, \\ \mathcal{N}(0.5 - \gamma, \theta^2), & \text{if } t \geq \tau, \end{cases}$$

$\gamma = 0.2$, and $\eta = 0.08$. We plot in Fig. 6 the Wald's approximations to the ADD and the FAP, as the test threshold h varies and for various DP levels. We assume $N = 300$ and for a linearly spaced set of θ values in the range of $[0.07, 0.13]$, we obtain various DP levels as $\epsilon \approx 1/(100\,\theta)$ assuming $\delta = 0.0139$ and recalling that $\theta = \sigma/\sqrt{N}$, see Eq. (13). Fig. 6 illustrates that at the same levels of FAP, we obtain larger ADDs as ϵ decreases. This implies that the anomaly detection performance degrades for stronger DP guarantees, clearly showing the privacy-anomaly detection tradeoff.

V. Performance Evaluation

In this section, we evaluate the proposed privacy-preserving anomaly detection scheme against botnet attacks over a real IoT network consisting of 9 nodes: a thermostat, a baby monitor, a webcam, two doorbells, and four security cameras. We use the network-based detection of IoT botnet attacks (N-BaIoT) dataset [52] in the UCI Machine Learning Repository [53], where the data dimensionality at each node n is 115, that is, $\mathbf{x}_{t,n} \in \mathbb{R}^{115}$. The data are the network traffic statistics of each node, particularly, the number of data packets received and sent, time intervals between packet arrivals, packet sizes, and so forth. The dataset contains both nominal data and anomalous data obtained under IoT botnet attacks. We employ the PCA-based local data processing (see Algorithm 1) at each node since the local nominal data at each node can be well represented in a linear subspace. Particularly, almost all the data variance is retained in the 5-dimensional principal subspace, that is $r_n = 5$, for the local nominal data of each node n.

We consider the UDP flooding attacks and the spam attacks by the BASHLITE botnet against the IoT network [52]. As the

performance metrics, we use the worst-case ADD, $\mathbb{E}_1\left[(\Gamma-\tau)^+\right]$, the FAP, $\mathbb{E}_\infty[\Gamma]$, and the false alarm rate (FAR), which is the reciprocal of the FAP:

$$\text{FAR} \triangleq \frac{1}{\mathbb{E}_\infty[\Gamma]}.$$

We present the ADD vs. FAR curves of the proposed detector for various levels of the DP to illustrate the privacy-anomaly detection tradeoff. Moreover, we compare our theoretical lower bound and approximation to the FAP with the actual FAP of the proposed detector obtained with the real-world data. Furthermore, we compare our theoretical worst-case upper bound on the ADD with the actual ADD of the proposed detector. Here, we note that we compare the ADD only with the worst-case upper bound since in this experiment, $\gamma_t, t \geq \tau$ are unknown and possibly time-varying. Finally, we present the nonparametric sliding-window chi-squared test as a benchmark detector and compare its performance with the proposed detector.

In our experiments, since we set the DP parameter σ^2 to values comparable to $1/12$, we use $\theta^2 = (\sigma^2 + 1/12)/N$, that is, we do not ignore the $1/12$ term here (see Sec. III-C.1). This is mainly because of the small network size ($N = 9$). Moreover, although the maximum variance of each of the $p_{t,n}$ is $1/4$ in the post-change regime as argued before, their actual variance can be much smaller than the maximum value, and hence we consider the same variance ($\theta^2 = (\sigma^2 + 1/12)/N$) for both the nominal and anomaly cases in our experiments.

As an example, we choose the detector sensitivity parameter as $\eta = 0.08$. For the reliability of the proposed detector as well as the presented FAP lower bound and the approximation to be valid, we need $\rho > 0.61$ (see Theorem 2), or equivalently, $0.1311 > \theta$ and

$$0.1311^2 > \frac{\sigma^2 + 1/12}{N},$$

that finally leads to $\sigma^2 < 0.0715$. Hence, we vary σ^2 in this range to obtain several different DP levels. Notice that the range of possible σ^2 values is, in fact, comparable to $1/12$.

First, we plot in Fig. 7 the FAP of the proposed anomaly detection scheme as well as the presented lower bound and the approximation. Here, we set the DP parameter as $\sigma^2 = 1/81$. Then, in case of the spam and the UDP flooding attacks over the network, we present ADD vs. FAR curves for various DP levels in Fig. 8 and Fig. 9, respectively. Particularly, we choose the local noise variance σ^2 from the set $\{0, 1/81, 1/36, 1/24, 1/16, 1/12, 1/9\}$. Then, assuming $\delta = 0.0139$, we obtain various DP levels as $\epsilon \approx 1/\sigma$, see Eq. (13). The figures illustrate that for the same level of FARs, the ADDs are larger for stronger DP levels (lower ϵ). The implication is that stronger DP guarantees worsen the anomaly detection performance, in compliance with the analytical tradeoff discussed in Sec. IV-C.

In case of a spam attack and with the chosen DP parameter $\sigma^2 = 1/16$, we compare in Fig. 10 the ADD of the proposed detector with its worst-case upper bound presented in Corollary 1, as the test threshold h varies. Fig. 10 validates that even if the anomaly cannot be specified in this experiment, we are still able to provide a theoretical guarantee in terms of the worst ADD of the proposed detector.

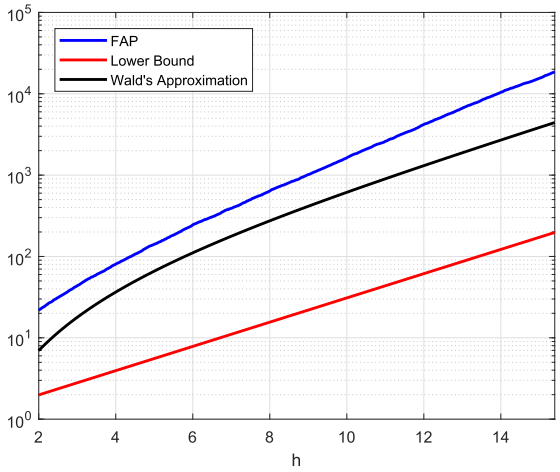


Fig. 7. The FAP of the proposed network anomaly detection scheme over the N-BaIoT dataset, the analytical lower bound, and the approximation for the FAP.

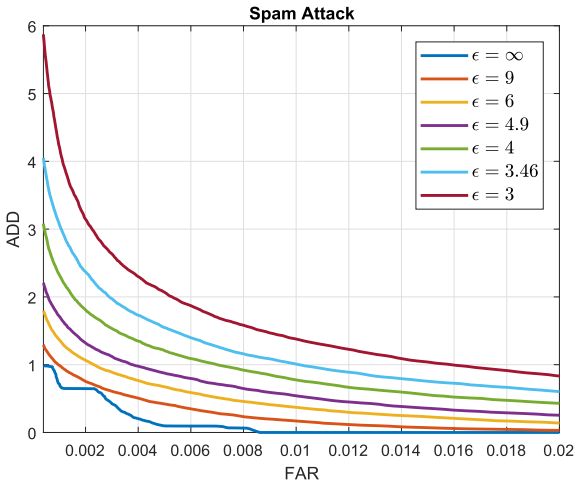


Fig. 8. ADD vs. FAR of the proposed network anomaly detection scheme in case of a spam attack over the network, for various DP levels.

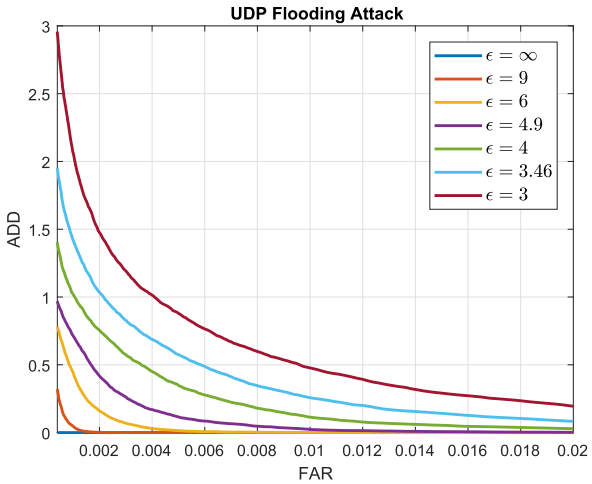


Fig. 9. ADD vs. FAR of the proposed network anomaly detection scheme in case of a UDP flooding attack over the network, for various DP levels.

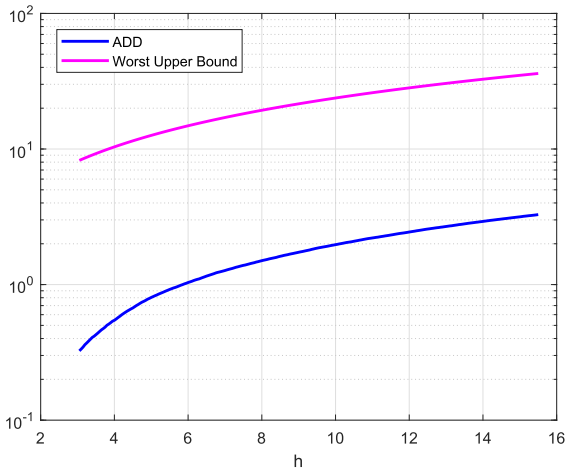


Fig. 10. The ADD of the proposed network anomaly detection scheme over the N-BaIoT dataset in case of a spam attack and the theoretical worst-case upper bound on the ADD.

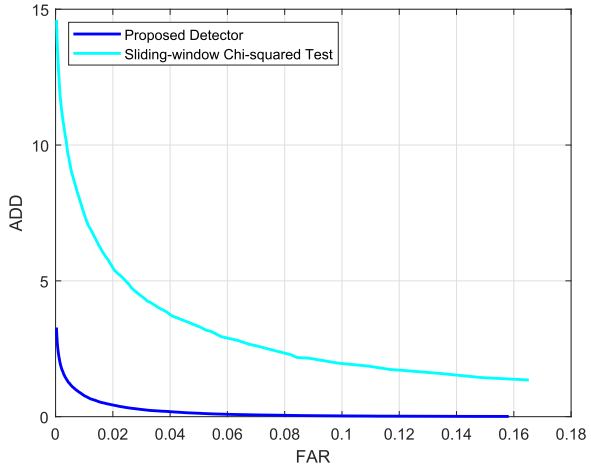


Fig. 11. ADD vs. FAR of the proposed generalized CUSUM detector and the sliding-window chi-squared test in case of a spam attack over the network.

We next present the sliding-window chi-squared test as a benchmark. For this test, we first obtain a nominal statistic that is independent of the system and algorithm parameters so as to use the same test regardless of the parameter values. By inspecting the pre-change model of y_t given in Eq. (5), we can write

$$q_t \triangleq \frac{(y_t - 0.5)^2}{\frac{\sigma^2+1/12}{N}} \sim \chi(1), \quad t < \tau, \tag{16}$$

where $\chi(1)$ denotes a chi-squared random variable with 1 degree of freedom. Notice that $q_t \sim \chi(1)$ is true irrespective of the network size and the DP parameter. Then, using Eq. (16), we can evaluate whether the observed sequence of $\{q_t, t = 1, 2, \dots\}$ fits to its nominal model. For this purpose, goodness-of-fit tests such as the Kolmogorov-Smirnov test and the Anderson-Darling test can be used [54]. We propose to use an online version of the Pearson's chi-squared test, as in [55]. Particularly, we divide the range $[0, \infty)$ of q_t into L disjoint and mutually exclusive intervals $I_1, I_2, \dots, I_L$, and based on the density of $\chi(1)$, we compute the probabilities $p_1 = \mathbb{P}(q_t \in I_1)$, $p_2 = \mathbb{P}(q_t \in I_2)$, ..., $p_L = \mathbb{P}(q_t \in I_L)$, where $\sum_{i=1}^{L} p_i = 1$. Let $\mathbf{W}_t \triangleq [q_{t-K+1}, q_{t-K+2}, \dots, q_t]$ be the online sliding window of size K. The expected number of window entries in each interval is then Kp_1, Kp_2, ..., Kp_L, respectively. Hence, we have a multinomial distribution with the expected number of samples in the disjoint intervals as Kp_1, Kp_2, ..., Kp_L. For the observed sliding window $\mathbf{W}_t$ at time t, we then count how many of its entries reside in each interval. Let the number of entries of $\mathbf{W}_t$ residing in each interval be $N_{1,t}$, $N_{2,t}$, ..., $N_{L,t}$ at time t. The Pearson's chi-squared test is then given as follows:

$$\Gamma = \inf \left\{ t : d_t \triangleq \sum_{i=1}^{L} \frac{(N_{i,t} - Kp_i)^2}{Kp_i} \geq \varphi \right\},$$

where φ is the test threshold that controls the false alarm rate. Here, the decision statistic d_t is asymptotically (as $K \to \infty$) a chi-squared random variable with $L-1$ degrees of freedom under the null hypothesis (no anomaly). Then, the decision threshold φ can be determined using the cdf of the chi-squared random variable with $L-1$ degrees of freedom in order to achieve the desired false alarm rate.

Fig. 11 compares the proposed detector with the sliding-window chi-squared test in case of a spam attack where the DP parameter is chosen as $\sigma^2 = 1/16$. For the chi-squared test, we choose $L = 8$, $p_1 = p_2 = \cdots = p_8 = 1/8$, and the window size as $K = 96$. Fig. 11 illustrates that the proposed detector outperforms the chi-squared test as it achieves lower ADD at the same levels of FAR.

VI. Concluding Remarks

We have studied online privacy-preserving data-driven network anomaly detection. We have proposed a distributed differentially private generalized CUSUM detector that infers network anomalies based on the perturbed and encrypted messages received from nodes. We have analyzed the anomaly detection performance of the proposed scheme in terms of system and algorithm parameters including the differential privacy parameter. In particular, we have derived a lower bound and an approximation for the average false alarm period (FAP) as well as an upper bound and an approximation for the average detection delay (ADD) of the proposed detector. Furthermore, we have used the derived FAP and ADD approximations to illustrate the analytical privacy-anomaly detection trade-off in the network anomaly detection problem. Our experiments over a real-world IoT dataset support our theoretical findings.

In this work, we assume that the anomaly is persistent, which means that once an anomaly happens, it lasts for a long time period. However, if the anomaly is non-persistent (e.g., after a short-term anomaly, the system returns back to normal operating conditions, and then a new short-term anomaly happens, and so on), the proposed detector might loose its effectiveness. Smart attackers can design such stealthy attacks against CUSUM-type detectors [55]. Analysis of possible privacy-preserving countermeasures against stealthy attacks can be studied in a future work.

APPENDIX

A. Proof of Theorem 1

Proof: In the proposed procedure, the information released to the network operator at time t can be written by, see Eq. (3),

$$y_t = \frac{1}{N}\sum_{i=1}^{N} p_{t,n} + \bar{v}_t,$$

where $\bar{v}_t \sim \mathcal{N}(0, \sigma^2/N)$ corresponds to the average perturbation noise over nodes. Defining a mean function

$$\phi(\{p_{t,n}\}_n) \triangleq \frac{1}{N}\sum_{i=1}^{N} p_{t,n},$$

we have

$$y_t = \phi(\{p_{t,n}\}_n) + \bar{v}_t.$$

Recalling that the p-value estimate $p_{t,n}$ takes values in the range of $[0, 1]$ irrespective of the nominal or anomaly cases, any single node n can change $\phi(\cdot)$ by at most $1/N$, for example, considering the change from $p_{t,n} = 0$ to $p_{t,n} = 1$. Hence, the sensitivity of the function $\phi(\cdot)$ is

$$\Delta\phi = \frac{1}{N}.$$

Since $\bar{v}_t$ is zero-mean AWGN, we can use Lemma 1 to decide the required noise variance for $\bar{v}_t$ to achieve (ϵ, δ)-DP at time t, as follows:

$$\frac{\sigma^2}{N} = \frac{2\log(1.25/\delta)}{N^2\epsilon^2},$$

and hence

$$\sigma^2 = \frac{2\log(1.25/\delta)}{N\epsilon^2}.$$

If every node n perturbs its output $p_{t,n}$ via zero-mean AWGN with the given variance of σ^2 above, we obtain a (ϵ, δ)-differentially private aggregation at the network operator at time t. Moreover, since we consider a data stream, at each time t, the incoming local data are processed and the corresponding $\{p_{t,n}\}_n$ are used once, and then never used again. This, in fact, corresponds to a scheme where at each time, a disjoint subset of the dataset is used, considering that the data obtained over all nodes and at all times form the database. The parallel composition rule of the DP [50] states that if ϵ_i-differentially private mechanisms are employed over disjoint subsets of a database, then the overall mechanism achieves $\max_i \epsilon_i$-DP. Then, by invoking the parallel composition property and since at each time t we employ a (ϵ, δ)-differentially private mechanism over a disjoint subset of the entire database, the overall stream aggregation at the network operator achieves the (ϵ, δ)-DP.

The network operator employs the generalized CUSUM algorithm over the privately aggregated stream of $\{y_t\}_t$. The post-processing invariance rule of the DP [4] states that for an output v of an (ϵ, δ)-differentially private algorithm, any non-private function $\psi(v)$ of the output also achieves (ϵ, δ)-DP, as long as the post-processing does not use the original data. Then, since the generalized CUSUM algorithm can be considered as a non-private function, overall the proposed online anomaly detection scheme is (ϵ, δ)-differentially private. □

B. Proof of Theorem 2

Proof: For the CUSUM-type detectors in the form of

$$\Gamma = \inf\{t : g_t \geq h\},$$
$$g_t = (g_{t-1} + \beta_t)^+, \tag{17}$$

where $g_0 = 0$, the Wald's approximation to the ARL is given by [1, Sec. 5.2.2]:

$$\mathbb{E}_\tau[\Gamma] \approx \frac{1}{\mathbb{E}[\beta_t]}\left(h + \frac{e^{-w_0 h} - 1}{w_0}\right), \tag{18}$$

where the equation

$$\mathbb{E}[e^{-w_0\beta_t}] = 1 \tag{19}$$

has only one nonzero root w_0 such that

$$\begin{cases} w_0 > 0, & \text{if } \mathbb{E}[\beta_t] > 0, \\ w_0 < 0, & \text{if } \mathbb{E}[\beta_t] < 0. \end{cases}$$

The proposed generalized CUSUM detector can be expressed in the form of Eq. (17), see Eq. (10) and Eq. (11). Then, to derive the Wald's approximation for the ARL, we need to compute $\mathbb{E}[\beta_t]$ and also w_0 from Eq. (19). To this end, we first compute the pdf of β_t for the nominal case ($t < \tau$) since in the FAP computations, the assumption is that no anomaly happens at all, that is, $\tau = \infty$.

Let $\mathrm{E}_1 \triangleq \{y_t \leq 0.5 - \eta\}$ and $\mathrm{E}_2 \triangleq \{y_t > 0.5 - \eta\}$ be two complementary events. For $t < \tau$, we have, see Eq. (9),

$$\begin{aligned}\mathbb{P}(\mathrm{E}_1) &= \mathbb{P}(y_t - 0.5 \leq -\eta) \\ &= \mathbb{P}\left(\frac{y_t - 0.5}{\theta} \leq \frac{-\eta}{\theta}\right) = Q(\eta/\theta)\end{aligned}$$

and hence $\mathbb{P}(\mathrm{E}_2) = Q(-\eta/\theta)$. Moreover, if E_1 is true, we have, see Eq. (12),

$$\begin{aligned}\beta_t &= \frac{1}{2}\left(\frac{y - 0.5}{\theta}\right)^2 \\ &\sim \frac{1}{2}\chi(1),\end{aligned}$$

where $\chi(1)$ denotes the chi-squared random variable with 1 degrees of freedom. Furthermore, if E_2 is true, we have, see Eq. (12),

$$\begin{aligned}\beta_t &= \frac{-\eta}{\theta}\left(\frac{y - 0.5}{\theta}\right) - \frac{\eta^2}{2\theta^2} \\ &\sim \mathcal{N}\left(-\frac{\eta^2}{2\theta^2}, \frac{\eta^2}{\theta^2}\right).\end{aligned}$$

In summary, for $t < \tau$, we have

$$\beta_t \sim \begin{cases} \chi(1)/2, & \text{w.p. } Q(\eta/\theta) \\ \mathcal{N}\left(-\dfrac{\eta^2}{2\theta^2}, \dfrac{\eta^2}{\theta^2}\right), & \text{w.p. } Q(-\eta/\theta), \end{cases} \tag{20}$$

where w.p. denotes "with probability". Then, using the linearity of the expectation, we can write

$$\mathbb{E}[\beta_t] = \frac{1}{2}Q(\eta/\theta) - \frac{\eta^2}{2\theta^2}Q(-\eta/\theta). \tag{21}$$

The Wald's approximation to the FAP requires $\mathbb{E}[\beta_t] < 0$, equivalently, after defining $\rho \triangleq \eta/\theta$, we need, see Eq. (21),

$$g(\rho) \triangleq Q(\rho) - \rho^2\, Q(-\rho) < 0.$$

Notice that since the Q-function is monotonically decreasing, the function $g(\rho)$ is monotonically decreasing in ρ and it takes the value of zero when $\rho \approx 0.61$. Hence, for $\rho > 0.61$, we have $\mathbb{E}[\beta_t] < 0$.

Next, we solve Eq. (19) to find w_0. Firstly, from Eq. (20), under E_1, $\beta_t \sim \chi(1)/2$ and hence the pdf of β_t in this case can be written as follows:

$$f(\beta) = \frac{1}{\sqrt{\pi\beta}} e^{-\beta}, \quad \beta \geq 0. \tag{22}$$

Then, using Eq. (20) and Eq. (22), we can rewrite Eq. (19) as

$$\begin{aligned} &Q(\rho) \underbrace{\int_0^\infty e^{-w_0\beta} \frac{1}{\sqrt{\pi\beta}} e^{-\beta} d\beta}_{A_1} \\ &\quad + Q(-\rho) \underbrace{\int_{-\infty}^\infty e^{-w_0\beta} \frac{1}{\sqrt{2\pi\rho^2}} e^{-\frac{1}{2\rho^2}(\beta+0.5\,\rho^2)^2} d\beta}_{A_2} = 1, \end{aligned} \tag{23}$$

where

$$\begin{aligned} A_1 &\triangleq \int_0^\infty e^{-w_0\beta} \frac{1}{\sqrt{\pi\beta}} e^{-\beta} d\beta \\ &= \int_0^\infty \frac{1}{\sqrt{\pi\beta}} e^{-(w_0+1)\beta} d\beta \end{aligned}$$

and letting $x \triangleq (w_0+1)\beta$, we can write

$$\begin{aligned} A_1 &= \frac{1}{\sqrt{w_0+1}} \underbrace{\int_0^\infty \frac{1}{\sqrt{\pi x}} e^{-x} dx}_{1} \\ &= \frac{1}{\sqrt{w_0+1}}, \end{aligned} \tag{24}$$

provided that $w_0 + 1 > 0$, or equivalently $w_0 > -1$. In the equation above, we use the fact that $\frac{1}{\sqrt{\pi x}} e^{-x}, x \geq 0$ represents a pdf (particularly, the pdf of $\chi(1)/2$).

Furthermore, we have

$$\begin{aligned} A_2 &= \int_{-\infty}^\infty e^{-w_0\beta} \frac{1}{\sqrt{2\pi\rho^2}} e^{-\frac{1}{2\rho^2}(\beta+0.5\,\rho^2)^2} d\beta \\ &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\rho^2}} e^{-\frac{1}{2\rho^2}(\beta^2+2(0.5\,\rho^2+\rho^2 w_0)\beta+\rho^4/4)} d\beta \\ &= e^{0.5\,\rho^2(w_0+w_0^2)} \underbrace{\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\rho^2}} e^{-\frac{1}{2\rho^2}(\beta-(-0.5\,\rho^2-\rho^2 w_0))^2} d\beta}_{1} \\ &= e^{0.5\,\rho^2(w_0+w_0^2)}, \end{aligned} \tag{25}$$

where by completing the square, we obtain a Gaussian pdf, whose under area curve is equal to 1.

Hence, we can rewrite Eq. (19) based on Eq. (23), Eq. (24), and Eq. (25) as follows:

$$f(w_0) \triangleq Q(\rho) \frac{1}{\sqrt{w_0+1}} + Q(-\rho)\, e^{0.5\,\rho^2(w_0+w_0^2)} = 1, \tag{26}$$

where there exists a unique $-1 < w_0 < 0$ solving the equation above. □

C. Proof of Theorem 3

Proof: For the CUSUM-type detectors given in Eq. (17) and if $\mathbb{E}[\beta_t] < 0$, we have the following lower bound on the ARL [1, Sec. 5.2.2]:

$$\mathbb{E}_\infty[\Gamma] \geq e^{-w_0 h},$$

where $w_0 < 0$ is obtained from Eq. (19) and hence from Eq. (26). □

D. Proof of Theorem 4

Proof: We can use the Wald's approximation to the ARL given in Eq. (18) to derive an approximation for the worst-case ADD of the proposed algorithm provided that $\mathbb{E}[\beta_t] > 0$. For this approximation, similar to Appendix VI-B, we need to compute $\mathbb{E}[\beta_t]$ and w_1 (for the ADD calculations, we use w_1 instead of w_0). To this end, we next determine the pdf of β_t for the post-change case, that is, for $t \geq \tau$.

Firstly, using the same event definitions E_1 and E_2 in Appendix VI-B and assuming $\gamma_t = \gamma$ for $t \geq \tau$, we have, see Eq. (9),

$$\begin{aligned} \mathbb{P}(\mathrm{E}_1) &= \mathbb{P}(y_t - 0.5 + \gamma \leq \gamma - \eta) \\ &= \mathbb{P}\left(\frac{y_t - 0.5 + \gamma}{\theta} \leq \frac{\gamma - \eta}{\theta}\right) = Q\left(\frac{\eta - \gamma}{\theta}\right) \end{aligned}$$

and hence $\mathbb{P}(\mathrm{E}_2) = Q(\frac{\gamma-\eta}{\theta})$. Moreover, if E_1 is true, we have, see Eq. (12),

$$\beta_t = \tfrac{1}{2}\left(\tfrac{y-0.5}{\theta}\right)^2 = \tfrac{1}{2}\,x^2,$$

where $x \sim \mathcal{N}(-\gamma/\theta, 1)$. Further, if E_2 is true, we have, see Eq. (12),

$$\begin{aligned} \beta_t &= \frac{-(y-0.5)\eta}{\theta^2} - \frac{\eta^2}{2\theta^2} \\ &= \frac{-\eta}{\theta}\left(\frac{y - 0.5 + \gamma}{\theta}\right) + \frac{2\eta\gamma - \eta^2}{2\theta^2} \\ &\sim \mathcal{N}\left(\frac{2\eta\gamma - \eta^2}{2\theta^2}, \frac{\eta^2}{\theta^2}\right) \end{aligned}$$

In summary, for $t \geq \tau$, we can write

$$\beta_t \sim \begin{cases} \frac{1}{2}\left(\mathcal{N}(-\gamma/\theta, 1)\right)^2, & \text{w.p. } Q\left(\frac{\eta-\gamma}{\theta}\right) \\ \mathcal{N}\left(\frac{2\eta\gamma-\eta^2}{2\theta^2}, \frac{\eta^2}{\theta^2}\right), & \text{w.p. } Q\left(\frac{\gamma-\eta}{\theta}\right). \end{cases} \tag{27}$$

Then, we have

$$\mathbb{E}[\beta_t] = \frac{\gamma^2+\theta^2}{2\theta^2} Q\left(\frac{\eta-\gamma}{\theta}\right) + \frac{2\eta\gamma-\eta^2}{2\theta^2} Q\left(\frac{\gamma-\eta}{\theta}\right). \tag{28}$$

To use the Wald's approximation for the ADD, we need $\mathbb{E}[\beta_t] > 0$, for which a sufficient condition is

$$2\eta\gamma - \eta^2 > 0,$$

equivalently $\gamma > \eta/2$. This is because all the other terms in Eq. (28) are nonnegative. Moreover, since $\gamma \geq \eta$ by the definition of the proposed detector, see Eq. (10), the sufficient condition is satisfied.

Next, we solve Eq. (19) to determine w_1. We write Eq. (19) based on Eq. (27) as follows:

$$Q\left(\frac{\eta-\gamma}{\theta}\right)\underbrace{\int_{-\infty}^{\infty} e^{-w_1x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(x+\gamma/\theta)^2/2}dx}_{B_1} + Q\left(\frac{\gamma-\eta}{\theta}\right) \times \underbrace{\int_{-\infty}^{\infty} e^{-w_1\beta}\frac{1}{\sqrt{2\pi\eta^2/\theta^2}}e^{-\frac{1}{2\eta^2/\theta^2}\left(\beta-\frac{2\gamma\eta-\eta^2}{2\theta^2}\right)^2}d\beta}_{B_2} = 1, \tag{29}$$

where for the first term in the summation, we use $\beta = x^2/2$ where $x \sim \mathcal{N}(-\gamma/\theta, 1)$. Then, we have

$$\begin{aligned} B_1 &\triangleq \int_{-\infty}^{\infty} e^{-w_1x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(x+\gamma/\theta)^2/2}dx \\ &= e^{-\frac{w_1\gamma^2}{2\theta^2(w_1+1)}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\sqrt{w_1+1}\,x+\frac{\gamma}{\theta\sqrt{w_1+1}}\right)^2}dx \\ &= \frac{1}{\sqrt{w_1+1}}e^{-\frac{w_1\gamma^2}{2\theta^2(w_1+1)}}\underbrace{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(y+\frac{\gamma}{\sqrt{w_1+1}}\right)^2}dy}_{1} \\ &= \frac{1}{\sqrt{w_1+1}}e^{\frac{-w_1\gamma^2}{2\theta^2(w_1+1)}}, \end{aligned} \tag{30}$$

where we again use the method of completing the square and $y \triangleq \sqrt{w_1+1}\,x$. Further, we have

$$B_2 \triangleq \textstyle\int_{-\infty}^{\infty} e^{-w_1\beta}\frac{1}{\sqrt{2\pi\eta^2/\theta^2}}e^{-\frac{1}{2\eta^2/\theta^2}\left(\beta-\frac{2\gamma\eta-\eta^2}{2\theta^2}\right)^2}d\beta.$$

We determine B_2 by following the same methodology to find the A_2 in Appendix B, that is, completing the square. Then, we obtain

$$B_2 = e^{\frac{(\gamma^2-2\gamma\eta)w_1+\gamma^2w_1^2}{2\theta^2}} \tag{31}$$

Finally, based on Eq. (29), Eq. (30), and Eq. (31), we have

$$g(w_1) \triangleq Q\left(\frac{\eta-\gamma}{\theta}\right)\frac{e^{\frac{-w_1\gamma^2}{2\theta^2(w_1+1)}}}{\sqrt{w_1+1}} + Q\left(\frac{\gamma-\eta}{\theta}\right)e^{\frac{(\gamma^2-2\gamma\eta)w_1+\gamma^2w_1^2}{2\theta^2}} = 1,$$

where there exists a unique $w_1 > 0$ satisfying the equation. □

E. Proof of Theorem 5

Proof: For the CUSUM-type detectors given in the general form of Eq. (17), if $\mathbb{E}[\beta_t] > 0$ and the observation sequence is Gaussian, we have the following upper bound on the ARL [1, Sec. 5.2.2]:

$$\mathbb{E}_1[\Gamma] \le \frac{1}{\mathbb{E}[\beta_t]}\left(h + \mathbb{E}[\beta_t|\beta_t > 0]\right). \tag{32}$$

Since the proposed detector fits to Eq. (17) and y_t is Gaussian, we can derive an upper bound on the worst-case ADD of the proposed detector using Eq. (32). To this end, we next compute $\mathbb{E}[\beta_t|\beta_t > 0]$. Notice that $\mathbb{E}[\beta_t]$ is already computed and given in Eq. (28).

Firstly, based on Eq. (27), under E_1, since $\beta_t \ge 0$ is always true, we can easily compute

$$\mathbb{E}[\beta_t|\beta_t > 0, \mathrm{E}_1] = \frac{\gamma^2+\theta^2}{2\theta^2}$$

Further, under E_2, we have $\beta_t \sim \mathcal{N}(a, b)$ where $a \triangleq \frac{2\eta\gamma-\eta^2}{2\theta^2}$ and $b \triangleq \frac{\eta^2}{\theta^2}$. Then,

$$\begin{aligned} \mathbb{E}[\beta_t|\beta_t > 0, \mathrm{E}_2] &= \frac{\mathbb{E}[\beta_t, \beta_t > 0|\mathrm{E}_2]}{\mathbb{P}(\beta_t > 0|\mathrm{E}_2)} \\ &= \frac{\mathbb{E}[\beta_t, \beta_t > 0|\mathrm{E}_2]}{Q(-a/\sqrt{b})}, \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}[\beta_t, \beta_t > 0|\mathrm{E}_2] &= \int_0^{\infty}\beta\frac{1}{\sqrt{2\pi b}}e^{-\frac{1}{2b}(\beta-a)^2}d\beta \\ &= \underbrace{\int_0^{\infty}(\beta-a)\frac{1}{\sqrt{2\pi b}}e^{-\frac{1}{2b}(\beta-a)^2}d\beta}_{C_1} \\ &\quad + \underbrace{\int_0^{\infty}a\frac{1}{\sqrt{2\pi b}}e^{-\frac{1}{2b}(\beta-a)^2}d\beta}_{C_2} \end{aligned}$$

Let $u \triangleq (\beta-a)^2$. Then,

$$\begin{aligned} C_1 &\triangleq \int_0^{\infty}(\beta-a)\frac{1}{\sqrt{2\pi b}}e^{-\frac{1}{2b}(\beta-a)^2}d\beta \\ &= \int_{a^2}^{\infty}\frac{1}{2\sqrt{2\pi b}}e^{-\frac{u}{2b}}du \\ &= \frac{\sqrt{b}}{\sqrt{2\pi}}e^{-\frac{a^2}{2b}}. \end{aligned}$$

Moreover, letting $y \triangleq (x-a)/\sqrt{b}$, we have

$$\begin{aligned} C_2 &\triangleq \int_0^{\infty}a\frac{1}{\sqrt{2\pi b}}e^{-\frac{1}{2b}(\beta-a)^2}d\beta \\ &= a\int_{-\frac{a}{\sqrt{b}}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy \\ &= a\,Q\left(-\frac{a}{\sqrt{b}}\right) \end{aligned}$$

Then, we obtain the following:

$$\begin{aligned} \mathbb{E}[\beta_t|\beta_t > 0, \mathrm{E}_2] &= \frac{\frac{\sqrt{b}}{\sqrt{2\pi}}e^{-\frac{a^2}{2b}} + a\,Q\left(-\frac{a}{\sqrt{b}}\right)}{Q(-a/\sqrt{b})} \\ &= a + \frac{\sqrt{b}\,e^{-\frac{a^2}{2b}}}{\sqrt{2\pi}\,Q(-a/\sqrt{b})} \\ &\triangleq \psi(a,b) \end{aligned}$$

Finally, we have

$$\begin{aligned} \mathbb{E}[\beta_t|\beta_t > 0] &= \mathbb{P}(\mathrm{E}_1)\,\mathbb{E}[\beta_t|\beta_t > 0, \mathrm{E}_1] \\ &\quad + \mathbb{P}(\mathrm{E}_2)\,\mathbb{E}[\beta_t|\beta_t > 0, \mathrm{E}_2] \\ &= Q\left(\frac{\eta-\gamma}{\theta}\right)\frac{\gamma^2+\theta^2}{2\theta^2} + Q\left(\frac{\gamma-\eta}{\theta}\right)\psi(a,b) \end{aligned} \tag{33}$$

Then, in Eq. (32), by replacing $\mathbb{E}[\beta_t]$ and $\mathbb{E}[\beta_t|\beta_t > 0]$ with Eq. (28) and Eq. (33), respectively, we obtain an upper bound on the ADD. □

REFERENCES

[1] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

[2] M. N. Kurt, Y. Yilmaz, and X. Wang, "Distributed quickest detection of cyber-attacks in smart grid," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 8, pp. 2015–2030, Aug. 2018.

[3] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[4] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE Signal Process. Mag.*, vol. 30, no. 5, pp. 86–94, Sep. 2013.

[5] M. N. Kurt, Y. Ylmaz, and X. Wang, "Real-time nonparametric anomaly detection in high-dimensional settings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2463–2479, Jul. 2021.

[6] M. N. Kurt, Y. Yilmaz, and X. Wang, "Sequential model-free anomaly detection for big data streams," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2019, pp. 421–425.

[7] D. Eckhoff and C. Sommer, "Driving for big data? Privacy concerns in vehicular networking," *IEEE Security Privacy*, vol. 12, no. 1, pp. 77–79, Jan. 2014.

[8] S. Finster and I. Baumgart, "Privacy-aware smart metering: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1732–1745, 3rd Quart., 2014.

[9] S. Tonyali, K. Akkaya, N. Saputro, A. S. Uluagac, and M. Nojoumian, "Privacy-preserving protocols for secure and reliable data aggregation in IoT-enabled smart metering systems," *Future Gener. Comput. Syst.*, vol. 78, pp. 547–557, Jan. 2018.

[10] O. Gornerup, N. Dokoohaki, and A. Hess, "Privacy-preserving mining of frequent routes in cellular network data," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, Aug. 2015, pp. 581–587.

[11] X. Liu, P. Zhu, Y. Zhang, and K. Chen, "A collaborative intrusion detection mechanism against false data injection attack in advanced metering infrastructure," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2435–2443, Sep. 2015.

[12] E. Shi, H. Chan, E. Rieffel, R. Chow, and D. Song, "Privacy-preserving aggregation of time-series data," in *Proc. Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*. Princeton, NJ, USA: Citeseer, 2011, pp. 1–17.

[13] C. Castelluccia, A. C. Chan, E. Mykletun, and G. Tsudik, "Efficient and provably secure aggregation of encrypted data in wireless sensor networks," *ACM Trans. Sensor Netw.*, vol. 5, no. 3, p. 20, 2009.

[14] C. Zhao *et al.*, "Secure multi-party computation: Theory, practice and applications," *Inf. Sci.*, vol. 476, pp. 357–372, Feb. 2019.

[15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016, *arXiv:1602.05629*.

[16] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[17] J. L. Ny and G. J. Pappas, "Differentially private filtering," *IEEE Trans. Autom. Control*, vol. 59, no. 2, pp. 341–354, Feb. 2014.

[18] P. Jain, P. Kothari, and A. Thakurta, "Differentially private online learning," in *Proc. Conf. Learn. Theory*, 2012, pp. 1–24.

[19] T. Li, J. Li, Z. Liu, P. Li, and C. Jia, "Differentially private Naive Bayes learning over multiple data sources," *Inf. Sci.*, vol. 444, pp. 89–104, May 2018.

[20] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "CpSGD: Communication-efficient and differentially-private distributed SGD," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7564–7575.

[21] R. Cummings, S. Krehbiel, Y. Mei, R. Tuo, and W. Zhang, "Differentially private change-point detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10825–10834.

[22] C. L. Canonne, G. Kamath, A. Mcmillan, A. Smith, and J. Ullman, "The structure of optimal private tests for simple hypotheses," in *Proc. 51st Annu. ACM SIGACT Symp. Theory Comput.*, Jun. 2019, pp. 310–321.

[23] K. H. Degue and J. L. Ny, "On differentially private Gaussian hypothesis testing," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 842–847.

[24] T. Zhang and Q. Zhu, "Distributed privacy-preserving collaborative intrusion detection systems for VANETs," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 1, pp. 148–161, Mar. 2018.

[25] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn.* Berlin, Germany: Springer, 2006, pp. 486–503, doi: 10.1007/11761679_29.

[26] I. Leontiadis, K. Elkhiyaoui, and R. Molva, "Private and dynamic time-series data aggregation with trust relaxation," in *Proc. Int. Conf. Cryptol. Netw. Secur.* Cham, Switzerland: Springer, 2014, pp. 305–320, doi: 10.1007/978-3-319-12280-9_20.

[27] M. Keshk, E. Sitnikova, N. Moustafa, J. Hu, and I. Khalil, "An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems," *IEEE Trans. Sustain. Comput.*, vol. 6, no. 1, pp. 66–79, Jan. 2021.

[28] V. Mothukuri, P. Khare, R. M. Parizi, S. Pouriyeh, A. Dehghantanha, and G. Srivastava, "Federated learning-based anomaly detection for IoT security attacks," *IEEE Internet Things J.*, early access, May 5, 2021, doi: 10.1109/JIOT.2021.3077803.

[29] Y. Zhao, J. Chen, Q. Guo, J. Teng, and D. Wu, "Network anomaly detection using federated learning and transfer learning," in *Proc. Int. Conf. Secur. Privacy Digit. Economy*. Singapore: Springer, 2020, pp. 219–231.

[30] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "DÏoT: A federated self-learning anomaly detection system for IoT," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 756–767.

[31] Y. Liu *et al.*, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6348–6358, Apr. 2021.

[32] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, "DeepFed: Federated deep learning for intrusion detection in industrial cyber–physical systems," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5615–5624, Aug. 2021.

[33] Y. Qin and M. Kondo, "Federated learning-based network intrusion detection with a feature selection approach," in *Proc. Int. Conf. Electr., Commun., Comput. Eng. (ICECCE)*, Jun. 2021, pp. 1–6.

[34] B. Cetin, A. Lazar, J. Kim, A. Sim, and K. Wu, "Federated wireless network intrusion detection," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 6004–6006.

[35] R. A. Sater and A. B. Hamza, "A federated learning approach to anomaly detection in smart buildings," *ACM Trans. Internet Things*, vol. 2, no. 4, pp. 1–23, Nov. 2021.

[36] H. V. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[37] M. N. Kurt and X. Wang, "Multisensor sequential change detection with unknown change propagation pattern," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 3, pp. 1498–1518, Jun. 2019.

[38] G. Lorden, "Procedures for reacting to a change in distribution," *Ann. Math. Stat.*, vol. 42, pp. 1897–1908, Dec. 1971.

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

[40] M. N. Kurt, O. Ogundijo, C. Li, and X. Wang, "Online cyber-attack detection in smart grid: A reinforcement learning approach," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5174–5185, Sep. 2019.

[41] M. N. Kurt, Y. Yilmaz, and X. Wang, "Secure distributed dynamic state estimation in wide-area smart grids," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 800–815, 2020.

[42] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer, 2006.

[43] A. W. Van der Vaart, *Asymptotic Statistics*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[44] T.-H. H. Chan, E. Shi, and D. Song, "Privacy-preserving stream aggregation with fault tolerance," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.* Berlin, Germany: Springer, 2012, pp. 200–214.

[45] F. Benhamouda, M. Joye, and B. Libert, "A new framework for privacy-preserving aggregation of time-series data," *ACM Trans. Inf. Syst. Secur.*, vol. 18, no. 3, pp. 1–21, Apr. 2016.

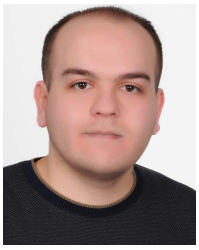
[46] V. Rastogi and S. Nath, "Differentially private aggregation of distributed time-series with transformation and encryption," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2010, pp. 735–746.

[47] F. Valovich, "Aggregation of time-series data under differential privacy," in *Progress in Cryptology—LATINCRYPT 2017*, T. Lange and O. Dunkelman, Eds. Cham, Switzerland: Springer, 2019, pp. 249–270.

[48] D. Becker, J. Guajardo, and K.-H. Zimmermann, "Revisiting private stream aggregation: Lattice-based PSA," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–17.

[49] P. Billingsley, "Probability and measure," in *Wiley Series in Probability and Mathematical Statistics*. Hoboken, NJ, USA: Wiley, 1986.

[50] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2009, pp. 19–30.
[51] *MathWorks, MATLAB, Release R2021a*, MathWorks, Natick, MA, USA, Mar. 2021.
[52] Y. Meidan *et al.*, "N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders," 2018, *arXiv:1805.03409*.
[53] D. Dheeru and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml
[54] M. A. Stephens, "EDF statistics for goodness of fit and some comparisons," *J. Amer. Stat. Assoc.*, vol. 69, no. 347, pp. 730–737, Sep. 1974.
[55] M. N. Kurt, Y. Yılmaz, and X. Wang, "Real-time detection of hybrid and stealthy cyber-attacks in smart grid," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 2, pp. 498–513, Feb. 2019.



Mehmet Necip Kurt (Member, IEEE) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2014 and 2016, respectively, and the Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA, in 2020. He is currently an AI Research Engineer with Samsung Research America, Berkeley Heights, NJ, USA. His research interests include sequential analysis, statistical signal processing, and machine learning with applications to cybersecurity, cyber-physical systems, and networks. He received the Eli Jury Award from Columbia University in 2020 for his doctoral studies.

Yasin Yılmaz (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Columbia University, New York, NY, USA, in 2014. He is currently an Assistant Professor of electrical engineering with the University of South Florida, Tampa. His research interests include machine learning, statistical signal processing, and their applications to computer vision, cybersecurity, energy systems, transportation systems, communications systems, and socioeconomic systems.

Xiaodong Wang (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Princeton University. He is currently a Professor of electrical engineering with Columbia University, New York. His research interests fall in the general areas of computing, signal processing, and communications, and he has published extensively in these areas. Among his publications is a book entitled *Wireless Communication Systems: Advanced Techniques for Signal Reception* (Prentice Hall, 2003). His current research interests include wireless communications, statistical signal processing, and genomic signal processing. He is listed as the ISI Highly-Cited Author. He received the 1999 NSF CAREER Award, the 2001 IEEE Communications Society and Information Theory Society Joint Paper Award, and the 2011 IEEE Communication Society Award for Outstanding Paper on New Communication Topics. He has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE TRANSACTIONS ON INFORMATION THEORY.

Pieter J. Mosterman (Member, IEEE) received the M.Sc. degree in electrical engineering from the University of Twente, The Netherlands, and the Ph.D. degree in electrical and computer engineering from Vanderbilt University, Nashville, TN, USA. He is currently the Chief Research Scientist and the Director of the MathWorks Advanced Research and Technology Office, Natick, MA, USA, where he works on computational methodologies and technologies. From 2009 to 2017, he held an Adjunct Professor position with the School of Computer Science, McGill University. Prior to joining MathWorks, he was a Research Associate with the German Aerospace Center (DLR), Oberpfaffenhofen. He has published over 100 peer-reviewed articles, and is the inventor of over 100 awarded patents. His primary research interests are in computer automated multiparadigm modeling (CAMPaM) with principal applications in design automation, training systems, and fault detection; isolation; and reconfiguration. He designed the Electronics Laboratory Simulator that was nominated for the Computerworld Smithsonian Award by Microsoft Corporation in 1994. In 2003, he was awarded the IMechE Donald Julius Groen Prize for his paper on the hybrid bond graph modeling and simulation environment HyBrSim. In 2009, he received the Distinguished Service Award of the Society for Modeling and Simulation International (SCS) for his services as the Editor-in-Chief of *SIMULATION: Transactions of SCS*. He was a Guest Editor of special issues on CAMPaM of *SIMULATION*, IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY, and *ACM Transactions on Modeling and Computer Simulation*. He has chaired over 30 scientific events and served on more than 100 international program committees.