Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Intelligent Cybersecurity Framework for IoT-Based Smart Grid Systems

Xiangsen Xu

xuxiangsen2026@yeah.net

State Grid Jibei Information & Telecommunication Company

Jing Zeng

State Grid Jibei Information & Telecommunication Company

Pingzhou Li

State Grid Jibei Electric Power Company Limited

Research Article

Keywords: Smart grid security, Internet of Things, Anomaly detection, Sensor compromise, False data injection, Certified robustness, Industrial control systems

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11232555/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Intelligent Cybersecurity Framework for IoT-Based Smart Grid Systems

Xiangsen Xu[1*], Jing Zeng[1], Pingzhou Li[2]

[1] State Grid Jibei Information & Telecommunication Company, Beijing 100032, China

[2] State Grid Jibei Electric Power Company Limited, Beijing 100032, China

Jing Zeng: 760044312@qq.com

Pingzhou Li: li.pingzhou@jibei.sgcc.com.cn

[*] Corresponding author: Xiangsen Xu: xuxiangsen2026@yeah.net

Abstract

Learned anomaly detectors for smart grid telemetry read every sensing node through one model, so an attacker who controls a few smart meters, relays or process sensors can hide a real event, raise false alarms, or inject false data without being pointed at. This paper presents TALLY, a three-layer detection framework in which the number of compromised nodes that can be tolerated is a design parameter with a proof. Every channel is scored by an ensemble of small predictors that each read one block of other nodes, plus a predictor of the channel's own past; the blocks partition the other nodes, so a compromised node can corrupt at most one member of every honest channel's ensemble, and an order statistic of the member residuals is provably confined between two honest order statistics. Thresholds calibrated on attack-free data at a guarded rank keep honest nodes from being flagged, a quorum keeps the compromised nodes from raising an alarm on their own, and a per-window certificate states when an alarm survives every compromise of the tolerated size. We evaluate ten detectors with three seeds on a real power generation testbed with labelled attacks, a relay testbed, and a distribution feeder driven by real customer demand with coordinated load surges, against gradient, gradient-free and nominal-value attacks that adapt to the defence. On the feeder the tolerant design is the best detector we measured; on the generation testbed tolerating four compromised tags costs five points of area under the curve against an untolerant predictor; every joint baseline reaches a false-alarm rate near one under a single compromised node while the framework stays within its bound; no certified alarm was removed by any attack; and the cases where the framework is weaker, or where no detector works, are reported with their numbers.

Keywords: Smart grid security, Internet of Things, Anomaly detection, Sensor compromise, False data injection, Certified robustness, Industrial control systems

1 Introduction

A modern power system is monitored by a large population of networked sensing devices: smart meters at customer connections, phasor measurement units and intelligent electronic devices in substations, and process instrumentation in generating plants [1, 2]. Their readings are collected over field networks and read at an aggregation point or a control centre, increasingly by a learned anomaly detector: a model trained on attack-free telemetry that raises an alarm when the readings stop looking like normal operation

[3, 4, 5]. Such a detector needs no attack labels, reads hundreds of channels at once, and catches process manipulations that no single threshold would.

The same population of devices is the weakest part of the system. Sensing nodes are cheap, numerous, physically exposed, and often remotely updatable, so the compromise of some of them is the threat that every survey of smart grid security lists first [6, 7, 8]. Authentication, key agreement and access control protocols for grid devices and home area networks [9, 10, 11, 12, 13] raise the cost of getting in, but a node that has been taken over holds valid credentials and reports whatever its controller wants. Against a learned detector this is decisive: every published detector of this kind reads the whole measurement vector through one map, and a map with unbounded sensitivity to each coordinate is steered by whoever controls a few coordinates. An attacker with a handful of nodes can hide a physical attack that the honest nodes are reporting, raise alarms from normal operation until operators stop trusting the detector, or inject false data, for instance to move the state estimate [14, 15], without being pointed at.

This paper builds a detection framework for IoT-based smart grid systems in which the number of compromised nodes that can be tolerated is a design parameter with a proof attached, and measures what that tolerance costs on real telemetry. The construction has three layers (Figure 1). At the node layer nothing is trusted. At the consistency layer every channel is scored by a small ensemble of learned predictors, each reading one block of other nodes, plus a predictor of the channel's own past; the blocks partition the other nodes, so a compromised node corrupts at most one member of every honest channel's ensemble, and an order statistic of the member residuals is provably sandwiched between two honest order statistics. At the decision layer thresholds are calibrated on attack-free data at the guarded rank the sandwich prescribes, a node is flagged when any of its channels is in excess, and the alarm requires a quorum of q flagged nodes. With q larger than the tolerated number k of compromised nodes, the compromised nodes can neither raise the alarm on their own nor make an honest node look guilty; whether they can silence the alarm is decided window by window, and an alarm is certified against every compromise of k nodes when at least $q + k$ nodes are flagged at the pessimistic rank. Scoring inputs through ablated views and taking a majority comes from certified defences against sparse and patch attacks on classifiers [16, 17], and secure state estimation establishes the same counting limit for linear systems [18, 19]; what is new is the transfer to normal-only detection with per-channel consistency, the two certificates, and an evaluation on grid telemetry with attacks that adapt to the defence.

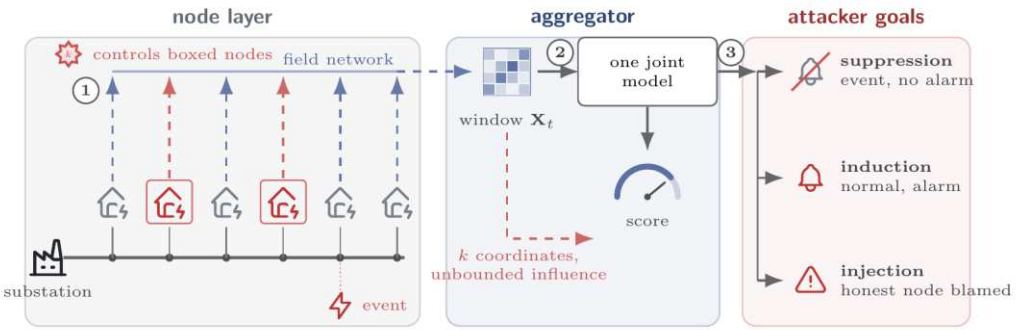


Fig. 1 Node compromise against a learned grid detector.

We evaluate on three deployments that differ in what the telemetry is, how fast it is sampled and what a node is: HAI, a hardware-in-the-loop power generation testbed with real labelled attacks [20, 21]; the

Mississippi State and Oak Ridge power system dataset of four phasor measurement relays under faults and cyber attacks [22]; and an IEEE 33-bus feeder [23, 24] with simulated physics and real demand, whose loads follow the consumption records of real customers [25], a smart meter at every load bus, and coordinated load surges by high-wattage IoT botnets as events [26, 27].

The contributions are the following.

- We formalize node compromise as a threat model for learned grid anomaly detection, with the attacker goals of suppression, alarm induction and false data injection, and show that standard detectors have no tolerance to it.
- We design TALLY, a three-layer framework whose consistency layer scores every channel through a partitioned ensemble of cross-node predictors and an own-history predictor, and whose decision layer combines guarded calibrated thresholds, node flags and a quorum.
- We prove that honest evidence is bounded by honest order statistics under any compromise of k nodes, that the false-alarm rate of honest nodes is controlled, that an alarm can be certified window by window, and that the block size sets the tolerance; we also state what cannot be certified.
- We evaluate ten detectors on three deployments with three seeds against white-box, gradient-free and nominal-value attacks, chart the trade-off between tolerance and clean detection, and report where the framework is weaker than an unprotected detector.

2 Related work

This section places the framework among the security architecture of IoT-based grids, secure estimation, learned anomaly detection for control systems, and certified robustness against sparse attacks.

2.1 Security of IoT-based smart grids

Surveys of smart grid security share the layered picture used here, with device compromise and data integrity attacks at the top of the threat lists [6, 7, 1, 2]. Much of the defensive literature is preventive and works at the device boundary. Identity-based authentication of grid devices with physically unclonable functions [9] and lightweight authenticated key agreement on the same primitive [12] bind a reading to a device that cannot be cloned; access control for IoT-based critical infrastructure in the grid [10] decides which device may talk to which; authentication for the home area network of a smart city [11] covers the customer side of the feeder studied here; and blockchain-backed, quantum-resilient protocols for grid IoT [13] protect the record of what was reported. All of these keep an attacker out of the device population; they are complementary to the present work and are assumed to have failed for the compromised nodes, so that detection at the aggregation point is the layer that must hold. Load-altering attacks by botnets of high-wattage devices show that the demand side is also a physical attack surface [26, 27].

2.2 False data injection and secure estimation

The false data injection attack of Liu, Ning and Reiter builds a measurement error, from knowledge of the network model, that the residual test of state estimation cannot distinguish from a legitimate state [14]; the detection literature that followed is surveyed in [8, 3]. Secure estimation studies the same problem

from the control side: Fawzi, Tabuada and Diggavi show that the number of attacked sensors that can be corrected is a counting quantity [18], Pasqualetti, Dörfler and Bullo give detectability and identifiability conditions [19], and Urbina et al. study the limits of stealth in process control [28]. These results rest on a known linear model. The framework here uses learned predictors instead, which makes it applicable to a boiler, a turbine and a feeder alike, and pays with a weaker but still exact statement: not a recovery of the state, but a bound on what compromised nodes can do to the evidence of honest ones.

2.3 Learned anomaly detection for control systems

Reconstruction and forecasting models are the workhorses of unsupervised detection on industrial telemetry: autoencoders and convolutional or recurrent forecasters [4, 29], adversarially trained autoencoders [30], and graph networks that learn which sensors predict which [5]. Their evaluation has been questioned for flawed benchmarks and point-adjusted scoring [31], so we report point-wise metrics on released labels plus an event-level rate. Their adversarial fragility is documented: Erba et al. conceal attacks from reconstruction-based detectors under realistic constraints [32], and Zizzo et al. attack time-series intrusion detection for control systems [33]; both assume an attacker who perturbs the readings the detector sees, which is what a compromised node does. Self-supervised detection of anomalous IoT nodes reads features the node itself reports [34, 35], so the same question arises there. Training-time attacks, such as poisoning in federated intrusion detection, and the robust aggregation rules that answer them [36, 37, 38] are a different threat: the model below is trained on trusted data and only the inputs at test time are hostile.

2.4 Certified robustness against sparse attacks and robust statistics

For sparse (ℓ_0) attacks on classifiers, Levine and Feizi certify by ablation: each vote reads a random subset of the coordinates, and a coordinate the attacker controls influences only the votes that read it [16]; derandomized smoothing turns the random subsets into a fixed set of blocks and certifies against patches [17]. The consistency layer of TALLY is a derandomized ablation over sensing nodes, adapted to a setting with no labels, a per-channel target instead of a class, and an order statistic instead of a plurality. The median has breakdown point one half [39], and the same counting argument underlies Byzantine-tolerant aggregation in distributed learning [40]; Lemma 1 is its finite-sample form for a fixed rank. Thresholds are set by split conformal calibration, whose finite-sample validity under exchangeability is classical [41] and whose use for outlier testing is analysed in [42].

3 System and threat model

This section fixes the deployment, the attacker, and the quantities that the framework controls and the experiments measure.

3.1 Deployment

The deployment has N sensing nodes. Node n reports the channels $\mathcal{C}_n \subset \{1, \dots, C\}$, and the $\mathcal{C}_n$ partition the C channels. A node may be a single tag (a sensor or actuator of a process), a relay that reports a vector of phasor quantities, or a smart meter that reports active power, reactive power and voltage magnitude. Readings are collected at a fixed interval and standardized channel by channel with statistics of the training data. At step t the aggregator holds the window $\boldsymbol{X}_t \in \mathbb{R}^{L \times C}$ of the last L readings, whose last row

$\boldsymbol{x}_t$ is the current reading. Training data consist of attack-free windows; a further attack-free set is held out for calibration. No attack label is used anywhere.

3.2 Attacker

Definition 1 (Node compromise). *The attacker controls a set $\mathcal{K}$ of at most k nodes. For every channel $c \in \mathcal{C}_{\mathcal{K}} = \bigcup_{n\in\mathcal{K}} \mathcal{C}_n$ and every lag of the window it may replace the reported value by any value inside the range observed for that channel in the training data. It may choose $\mathcal{K}$ after seeing the honest readings and with full knowledge of the detector. It does not modify the readings of any other node.*

The range constraint is a plausibility check any aggregator can apply; without it the results below would only be stronger for the framework and weaker for the baselines. The attacker has three goals.

- *Suppression.* A physical attack or event is in progress and visible in the honest readings. The attacker wants the detector not to alarm.
- *Induction.* Operation is normal. The attacker wants the detector to alarm, or to point at an honest node.
- *Injection.* The attacker reports false values for its own ends, for instance to bias a state estimate or a billing record, and wants not to be pointed at.

We consider a white-box attacker who computes gradients of the detector's statistic, a gradient-free attacker who searches the range of each controlled channel, and a nominal attacker who reports the calibration mean of each controlled channel. All three are evaluated, and the strongest per window is reported.

For a detector with alarm rule $A(\boldsymbol{X}) \in \{0,1\}$, we measure false-alarm and detection rates with and without a compromise, the rate at which honest nodes are flagged when the attacker acts, and, for injection, whether the compromised nodes are the ones flagged. A detector *tolerates* k compromised nodes when no compromise of k nodes can raise the false-alarm rate of honest nodes, and its alarm on a window is *certified* at k when no compromise of k nodes can remove it.

4 The TALLY framework

TALLY has three layers (Figure 2): the node layer reports and is not trusted; the consistency layer turns the window into two evidence values per channel through small learned predictors arranged so that a compromised node touches as little as possible; the decision layer turns evidence into node flags, a system alarm, an attribution and a certificate, with thresholds calibrated on attack-free data at ranks dictated by the tolerance.

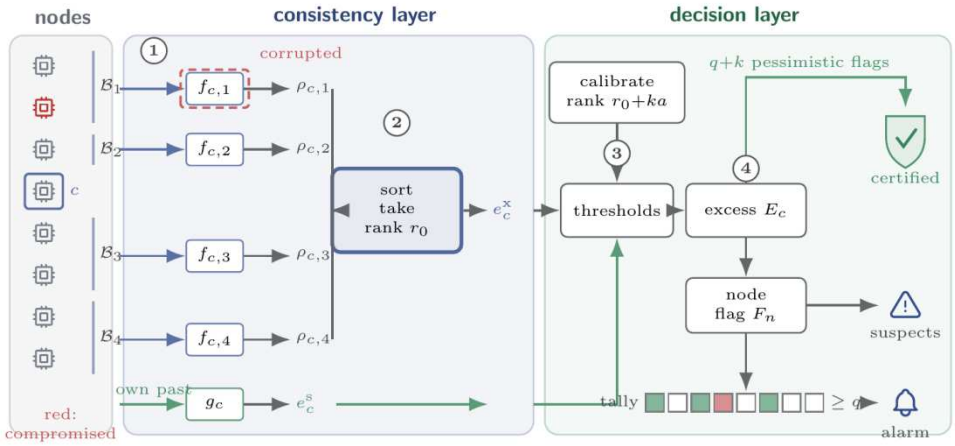


Fig. 2 The three layers of TALLY.

4.1 Consistency layer: block design and evidence bands

Fix a block size m in nodes. For every channel c, the nodes other than the one that owns c are partitioned, in a fixed random order, into M_c blocks $\mathcal{B}_{c,1}, \dots, \mathcal{B}_{c,M_c}$ of m nodes each; when the count does not divide, a remainder smaller than half a block is merged into the last block. Every other node therefore appears in exactly one block of c; we write $a = 1$ for this multiplicity, and keep a in the statements so that overlapping designs are covered. Member j of channel c is a predictor $f_{c,j}$ that reads the window restricted to the channels of $\mathcal{B}_{c,j}$ and outputs a prediction of $x_{c,t}$. Its standardized residual is

$$\rho_{c,j}(\boldsymbol{X}) = \frac{|x_{c,t} - f_{c,j}(\boldsymbol{X}_{\mathcal{B}_{c,j}})|}{s_{c,j}},$$

where $s_{c,j}$ is the root mean square of the member's residual on the calibration windows. Let $\rho_{c,(1)} \leq \cdots \leq \rho_{c,(M_c)}$ be the sorted member residuals. The *cross evidence* of channel c is the order statistic at the median rank,

$$e_c^{\mathrm{x}}(\boldsymbol{X}) = \rho_{c,(r_0)}(\boldsymbol{X}), \qquad r_0 = \lceil M/2 \rceil, \quad M = \min_c M_c.$$

A second predictor g_c reads only the channel's own past values $x_{c,t-L+1}, \dots, x_{c,t-1}$, and its standardized residual is the *self evidence* $e_c^{\mathrm{s}}(\boldsymbol{X}) = |x_{c,t} - g_c(\cdot)|/s_c$. The cross band asks whether the reading agrees with what other nodes report now and stays high for as long as an inconsistency lasts; the self band asks whether it agrees with the channel's own recent trajectory, catches changes at their onset, and is unaffected by any other node. Which band matters more depends on the sampling interval relative to the process dynamics, and Section 7 shows that the two deployments with context differ in exactly this way. All members are trained jointly as one batched model, a two-layer network of width 64 per member, by minimizing the squared residual over the training windows; a linear variant is an ablation. The block size m is the only structural hyperparameter and the tolerance dial of Section 5.

4.2 Decision layer: guarded thresholds, flags, quorum

The decision layer is parameterized by the tolerated number of compromised nodes k, the quorum $q > k$, and a per-channel false-alarm level α. It first fixes thresholds on the calibration windows. For the cross band it does not calibrate the median itself but the *guarded* order statistic $\rho_{c,(r_0+ka)}$: the threshold $\tau_c^{\mathrm{x}}(k)$ is the $\lceil (n+1)(1-\alpha/2) \rceil$-th smallest of the n calibration values of $\rho_{c,(r_0+ka)}$, the split conformal quantile at level $\alpha/2$ [41]. For the self band the threshold τ_c^{s} is the same quantile of the calibration values of e_c^{s}. At test time channel c is *in excess* when

$$E_c(\boldsymbol{X}) = \mathbb{1}[e_c^{\mathrm{x}}(\boldsymbol{X}) > \tau_c^{\mathrm{x}}(k)] \ \vee \ \mathbb{1}[e_c^{\mathrm{s}}(\boldsymbol{X}) > \tau_c^{\mathrm{s}}],$$

node n is *flagged* when any of its channels is in excess, $F_n(\boldsymbol{X}) = \max_{c \in \mathcal{C}_n} E_c(\boldsymbol{X})$, and the system alarm fires when at least q nodes are flagged:

$$A(\boldsymbol{X}) = \mathbb{1}\left[\sum_{n=1}^{N} F_n(\boldsymbol{X}) \geq q\right].$$

The flags are also the attribution, and Section 5 shows that a compromised node can make itself flagged but not an honest node. For ranking the continuous score is the q-th largest node ratio, the ratio of a node being the largest e/τ over its channels and bands. Finally, let $F_n^-(\boldsymbol{X})$ be the flag computed with the cross band at the *pessimistic* rank $r_0 - ka$; the alarm is *certified at* k when $\sum_n F_n^-(\boldsymbol{X}) \geq q + k$, and Section 5 proves that a certified alarm survives every compromise of k nodes. Algorithms 1 and 2 summarize the offline and online steps.

Algorithm 1 TALLY, offline: design, training, calibration

1: Require: nodes $\mathcal{C}_1, \ldots, \mathcal{C}_N$; training windows $\mathcal{T}$; calibration windows $\mathcal{V}$ (n windows); block size m; tolerance k; level α
2: for each channel c do
  3: permute the other nodes with a fixed seed; cut into blocks $\mathcal{B}_{c,1}, \ldots, \mathcal{B}_{c,M_c}$ of m nodes, merging a short remainder into the last block
4: end for
5: $M \leftarrow \min_c M_c$; $r_0 \leftarrow \lceil M/2 \rceil$; require $r_0 - ka \geq 1$ and $r_0 + ka \leq M$ // Proposition 1
6: train all members $f_{c,j}$ and g_c jointly on $\mathcal{T}$ by minimizing $\sum_{c,j} (x_{c,t} - f_{c,j})^2 + \sum_c (x_{c,t} - g_c)^2$, stopping on $\mathcal{V}$
7: $s_{c,j} \leftarrow$ root mean square of $x_{c,t} - f_{c,j}$ on $\mathcal{V}$; $s_c \leftarrow$ the same for g_c
8: for each channel c do
  9: compute $\rho_{c,(r_0+ka)}$ and e_c^{s} on every window of $\mathcal{V}$
  10: $\tau_c^{\mathrm{x}}(k) \leftarrow$ the $\lceil (n+1)(1-\alpha/2) \rceil$-th smallest value of $\rho_{c,(r_0+ka)}$; $\tau_c^{\mathrm{s}} \leftarrow$ the same for e_c^{s}
11: end for
12: Ensure: members, scales, thresholds

Algorithm 2 TALLY, online: evidence, flags, alarm, certificate

1: Require: window $\boldsymbol{X}$; members, scales, thresholds; tolerance k; quorum $q > k$
2: for each channel c do
  3: $\rho_{c,j} \leftarrow |x_{c,t} - f_{c,j}(\boldsymbol{X}_{\mathcal{B}_{c,j}})|/s_{c,j}$ for $j = 1, \ldots, M_c$; sort

4: $e_c^{\mathrm{x}} \leftarrow \rho_{c,(r_0)}$; $e_c^{\mathrm{s}} \leftarrow |x_{c,t} - g_c(x_{c,t-L+1..t-1})|/s_c$
5: $E_c \leftarrow \mathbb{1}[e_c^{\mathrm{x}} > \tau_c^{\mathrm{x}}(k)] \vee \mathbb{1}[e_c^{\mathrm{s}} > \tau_c^{\mathrm{s}}]$; $E_c^- \leftarrow \mathbb{1}\big[\rho_{c,(r_0-ka)} > \tau_c^{\mathrm{x}}(k)\big] \vee \mathbb{1}[e_c^{\mathrm{s}} > \tau_c^{\mathrm{s}}]$
6: end for
7: for each node n do
8: $F_n \leftarrow \max_{c\in\mathcal{C}_n} E_c$; $F_n^- \leftarrow \max_{c\in\mathcal{C}_n} E_c^-$
9: end for
10: alarm $A \leftarrow \mathbb{1}[\sum_n F_n \geq q]$; suspects $\leftarrow \{n: F_n = 1\}$
11: certified $\leftarrow \mathbb{1}[\sum_n F_n^- \geq q + k]$ // Theorem 2
12: Ensure: A, suspects, certified

Figure 3 shows the three steps for one honest channel under a compromise of $k = 2$ nodes: two members read a compromised node and can take any value, the median of all M values stays between the $(r_0 - 2)$-th and the $(r_0 + 2)$-th honest value, the threshold was set on the latter, and the compromised nodes control their own flags and nothing else. The online step is one batched forward pass of all members followed by a sort of M values per channel, so its cost is linear in the number of members and in the window length and needs no attack model, no state estimate and no communication beyond the readings the aggregator already receives; the tolerance k and the quorum q can be changed at run time without retraining, because only the thresholds depend on them.

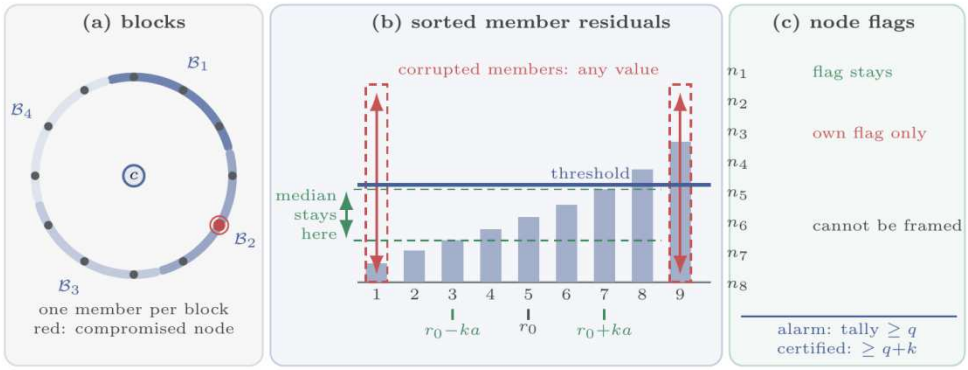


Fig. 3 Block design, order-statistic sandwich and quorum.

5 Guarantees

Throughout, a *compromise* is a choice of $\mathcal{K}$ with $|\mathcal{K}| \leq k$ and arbitrary values on $\mathcal{C}_{\mathcal{K}}$ at every lag, and *clean* quantities are those computed from the true readings. For an honest channel the clean and the compromised self evidence coincide, so only the cross band needs an argument.

Lemma 1 (Order statistics under bounded corruption). Let $v \in \mathbb{R}^M$ and let v' differ from v in at most b coordinates, which may take any values. For every rank r with $b < r \leq M - b$,

$$v_{(r-b)} \;\leq\; v'_{(r)} \;\leq\; v_{(r+b)}.$$

Proof. Let U be the set of at least $M - b$ coordinates on which v' and v agree, and write $w_{U,(i)}$ for the i-th smallest entry of a vector w restricted to U. Adding elements to a multiset can only lower an order statistic of fixed rank, so $v'_{(r)} \leq v'_{U,(r)} = v_{U,(r)}$; and U omits at most b entries of v, so its r-th smallest entry is at most the $(r + b)$-th smallest entry of v. For the lower bound, at most b of the r smallest entries

of v' lie outside U, so at least $r - b$ of them lie in U and $v'_{(r)} \geq v'_{U,(r-b)} = v_{U,(r-b)} \geq v_{(r-b)}$, the last step because an order statistic over a subset is at least the same-rank order statistic over the whole set. ▫

Theorem 1 (Honest evidence under compromise). Consider a compromise of at most k nodes in a design in which every node appears in at most a blocks of every channel, and let c be a channel of an honest node. Then

$$\rho^{\text{clean}}_{c,(r_0-ka)} \;\leq\; e^{\text{x}}_c \;\leq\; \rho^{\text{clean}}_{c,(r_0+ka)}, \qquad e^{\text{s}}_c = e^{\text{s,clean}}_c.$$

If the calibration windows and the benign test window are exchangeable, then under every such compromise $\Pr[E_c = 1] \leq \alpha$ for every honest channel, $\Pr[F_n = 1] \leq |\mathcal{C}_n|\alpha$ for every honest node, and for $q > k$

$$\Pr[A = 1 \mid \text{benign}] \;\leq\; \frac{C\alpha}{q-k}.$$

Proof. Member j of channel c is affected only if $\mathcal{B}_{c,j}$ contains a compromised node, and each of the at most k compromised nodes lies in at most a blocks of c, so at most ka member residuals differ from their clean values; the target $x_{c,t}$ is honest. Lemma 1 with $b = ka$ and $r = r_0$ gives the sandwich; the self band reads only channel c. For the probabilities, $e^{\text{x}}_c \leq \rho^{\text{clean}}_{c,(r_0+ka)}$, and the threshold $\tau^{\text{x}}_c(k)$ is the split conformal quantile of exactly this clean statistic at level $\alpha/2$, so the clean statistic exceeds it with probability at most $\alpha/2$ under exchangeability [41]; the same holds for the self band, and a union bound gives α. A node is flagged when one of its channels is in excess, which gives the second bound. Compromised nodes contribute at most k flags, so an alarm requires at least $q - k$ honest flags, and Markov's inequality applied to the expected number of honest flags, at most $C\alpha$, gives the last bound. ▫

The alarm bound is loose because it assumes no independence across nodes; the measured rates are far below it. Exchangeability is the assumption that fails in practice: when calibration data come from other weeks than the test data, as on HAI, honest channels drift and exceed α for reasons unrelated to the attacker, which is reported.

Theorem 2 (Per-window certificate). *For a window with clean readings $\boldsymbol{X}$, let $F^-_n(\boldsymbol{X})$ be the node flag computed with the cross band at rank $r_0 - ka$. If $\sum_n F^-_n(\boldsymbol{X}) \geq q + k$, then every compromise of at most k nodes leaves the alarm on: $A = 1$ for all admissible replacements of the compromised readings.*

Proof. Among the at least $q + k$ nodes with $F^-_n = 1$, at most k are compromised. Every other such node is honest and has a channel c with $\rho^{\text{clean}}_{c,(r_0-ka)} > \tau^{\text{x}}_c(k)$ or $e^{\text{s}}_c > \tau^{\text{s}}_c$. By Theorem 1, $e^{\text{x}}_c \geq \rho^{\text{clean}}_{c,(r_0-ka)}$ and e^{s}_c is unchanged, so c stays in excess and the node stays flagged. At least q nodes are therefore flagged whatever the compromised nodes report. ▫

The certificate is computed from the clean window: in deployment it is a confidence label on each alarm, and in the evaluation it is checked against the strongest attacks we can mount; an uncertified alarm may still survive a particular attack.

Proposition 1 (Tolerance of the partition design). *With N nodes and blocks of m nodes, every channel has $M_c \geq M = \lfloor (N-1)/m \rfloor$ blocks, up to the merge of a short remainder, and $a = 1$. The guarded and*

pessimistic ranks exist for every $k \le k^\star = \lceil M/2 \rceil - 1 = \lfloor (M-1)/2 \rfloor$*, and for no larger* k*. The self band alone admits every* k*, and the full predictor* $m = N - 1$ *has* $M = 1$ *and* $k^\star = 0$*.*

Proof. The ranks $r_0 - k$ and $r_0 + k$ lie in $\{1, \dots, M\}$ exactly when $k \le r_0 - 1$ and $k \le M - r_0$; with $r_0 = \lceil M/2 \rceil$ the first condition is the binding one and gives $k^\star = \lceil M/2 \rceil - 1$. A design with $a > 1$ divides $k^\star$ by a. The self band reads no other node, so no rank condition arises. ▫

The proposition is the tolerance dial: $k^\star \approx (N-1)/(2m)$, so halving the block size doubles the tolerance and halves what each member may read; members that read less predict less well, and Section 7 measures the price.

Remark 1 (Joint detectors). *A detector whose statistic is one function of the whole window with unbounded sensitivity to each coordinate, such as a reconstruction error or a joint forecast residual, has tolerance zero in the sense above: a single controlled channel with a wide enough range can move the statistic to either side of any threshold. Section 7 shows that this is not only a worst case but the typical outcome of a sixty-step gradient attack on one node.*

Proposition 2 (What is not certified). *(a) With* $q > k$*, an injection confined to at most* k *nodes never raises the system alarm; it is reported through the flags of the injecting nodes, and only if their readings are inconsistent with the honest ones or with their own past. (b) An event whose pessimistic flags cover fewer than* $q + k$ *nodes cannot be certified at* k*, whether or not it is detected. (c) A reading that keeps every member residual and the self residual inside the calibration bulk is undetectable by any statistic built from these residuals.*

Proof. (a) follows from the flag count, (b) from the definition of the certificate, and (c) from the fact that all decisions are functions of the residuals. ▫

Point (c) is the detectability boundary of every consistency detector. Point (a) is a design decision: a system alarm is an operator action, a node flag a maintenance ticket, and an attacker with k nodes can generate at most k tickets, all pointing at itself.

6 Experimental setup

This section describes the three deployments, the detectors, the attacks and the metrics (Figure 4).

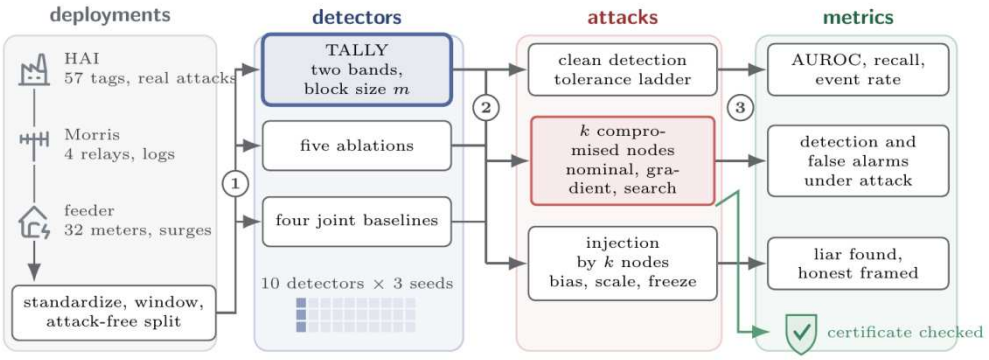


Fig. 4 Experimental pipeline.

6.1 Deployments

Table 1 summarizes the three deployments. They were chosen to differ in what a node is, how fast the telemetry is sampled, and where the physical coupling between nodes comes from.

Table 1 The three deployments.

Deployment	Node	N	C	Interval	L	Train	Calib.	Test	Event steps
HAI	tag	57	57	10 s	10	73,731	18,432	40,205	897
Morris	relay, logs	5	124	sample	1	13,646	2,930	20,886	14,748
Feeder	meter	32	96	15 min	4	35,136	7,680	15,360	842

HAI. The HAI 21.03 release of the HIL-based augmented ICS security testbed records a boiler (P1), a steam turbine (P2), a water treatment loop (P3) and the power simulation that couples them (P4) at one-second resolution, with three attack-free training traces and five test traces containing labelled attacks on setpoints, controllers and actuators [20, 21]. Every tag is a node. We decimate to ten seconds, drop tags that are constant in training, and use ten steps of context. The last fifth of every training trace is held out for calibration, so that calibration spans three operating periods; the test traces come from other weeks. The subsystem labels give three attack groups plus a group of attacks touching several subsystems.

Morris. The power system attack datasets of Mississippi State University and Oak Ridge National Laboratory record four phasor measurement relays on a two-line transmission testbed, twenty-nine quantities per relay plus control-panel, relay and intrusion-detection logs, under thirty-seven scenarios: natural faults and line maintenance, no-event operation, and cyber attacks of three kinds, replayed fault data, remote tripping commands and relay setting changes [22]. A node is a relay, the log columns form a fifth node, and the context is one step because the released samples carry no usable time order. Benign rows of nine of the fifteen files form the training set, those of two further files the calibration set, and all rows of the remaining four files the test set. Every row of an attack scenario counts as an attack step, including those recorded before the attack has any physical effect.

Feeder. The IEEE 33-bus radial feeder [23] is solved with pandapower [24] every fifteen minutes; each of its thirty-two loads follows the consumption record of one real customer of the ElectricityLoadDiagrams20112014 set [25], rescaled to the nominal load of its bus at power factor 0.9, and a smart meter at every load bus reports active power, reactive power and voltage magnitude with class-0.2 metering noise. Training uses the whole of 2012; the first 240 days of 2013 are split day by day at random into calibration (one third) and test (two thirds), so calibration and test are exchangeable by construction. Sixty percent of the test days carry one physical event, a coordinated surge in which two to four buses raise their demand by 50 to 150 percent for one to four hours, the demand-side attack of a high-wattage IoT botnet [26, 27]. The context is four steps.

Channels are standardized with training statistics, windows never cross a trace boundary, and no attack label is used for training, calibration or model selection.

6.2 Detectors

TALLY is run with blocks of six tags on HAI ($M = 9, k^{\star} = 4$), three meters on the feeder ($M = 11$, $k^{\star} = 5$) and one relay on Morris ($M = 4$, $k^{\star} = 1$), with $\alpha = 10^{-3}$ per channel and, unless stated otherwise, $q = 2$ at $k = 0$. Members are two-layer networks of width 64 trained with Adam for at most

thirty epochs with early stopping on the calibration windows. Four ablations isolate the design choices: the cross band alone, the self band alone, a single *mixed* member per block that reads the block and the channel's own past together, and linear members. The *full-block predictor* uses both bands with one member that reads every other node ($M = 1$, $k^\star = 0$), the strongest per-channel predictor the design admits and one with no tolerance. Four joint detectors represent the reconstruction and forecasting families used on industrial telemetry [4, 30, 29]: a dense autoencoder on the flattened window (256 and 32 units), a joint forecaster (256 and 128 units) that predicts every channel of the last step from the earlier steps, the Mahalanobis distance of the last step under a Ledoit–Wolf covariance [43], and an isolation forest [44]; each alarms above the split conformal one-percent quantile of its score on the calibration windows. The forecaster and the self band need context and are absent on Morris. Every detector is trained with three seeds.

6.3 Attacks and metrics

Clean detection is measured point-wise on the test stream, without any point adjustment [31]: the area under the ROC curve of the continuous score, the recall at a matched one-percent false-alarm rate (an oracle threshold on the test stream, the same yardstick for every detector), the recall and false-alarm rate of each detector's own rule, and the fraction of labelled events, defined as maximal runs of attack steps, that receive at least one alarm.

Compromise follows Definition 1: four hundred event windows are attacked for suppression and eight hundred benign windows for induction. The attacker picks its k nodes at random or as an *insider*, the k nodes with the largest evidence (TALLY) or the largest gradient of the score (joint detectors), and runs three attacks on the controlled channels, keeping the best per window: the *nominal* attack reports each channel's calibration mean; the *gradient* attack runs forty steps of projected gradient descent [45] on the attacker's objective, a soft count of flagged nodes for TALLY and the score for a joint detector, inside the box of Definition 1; the *search* attack, which needs no gradient, tries the bottom, middle and top of every controlled channel's range one channel at a time and keeps improvements, and is also the check recommended for defences with piecewise gradients [46]. TALLY is attacked in its certified configuration, tolerance k and quorum $q = k + 1$, and with $q = 2$ held fixed, which Theorem 1 says must fail once $k \geq 2$; detectors with no tolerance keep their clean configuration. We report detection and false-alarm rate under the strongest attack, the honest-node flag rate before and during the attack, the certified fraction of attacked windows, and the number of certified windows on which the strongest attack removed the alarm, which Theorem 2 says must be zero.

Injection takes 1,500 benign windows and k random nodes that report a lie, a bias of 0.5 to 4 standard deviations held over the whole window, a doubling of every deviation from the window mean, or a frozen value replayed from the first lag, clipped to the training range. We report the precision of the k highest-ranked nodes against the true liars, the fraction of lying and of honest nodes flagged, and the system alarm rate, which Proposition 2(a) says must stay at its clean value.

7 Results

The results answer, in order, what tolerance costs when nobody attacks, how the cost scales with the block size, what compromised nodes can do to each detector, and whether injecting nodes are the ones flagged.

7.1 Clean detection

Table 2 gives the clean detection of every detector on every deployment with the tolerance $k^\star$ of its design; the AUROC and the recall at a matched one-percent false-alarm rate compare rankings, and the own rule shows what a deployment would see at the configuration used in the rest of the paper.

HAI. The ordering is the one Proposition 1 predicts: the detector that reads the most is the best, and tolerance is paid for in detection. The full-block predictor reaches an AUROC of 0.961 and a recall of 0.874 at one percent, the joint forecaster 0.947 and 0.795, and TALLY with blocks of six tags, which tolerates four compromised tags, 0.915 and 0.739. The self band alone reaches 0.906, because process telemetry sampled every ten seconds is strongly autocorrelated and most released attacks move a tag away from its own trajectory; the cross band alone reaches 0.885; the mixed member, which reads block and own past in one network, is two points better than the two-band design at 0.926, a small price that the feeder justifies. Linear members lose three thousandths of AUROC, so the members are not where the capacity lies. The own rule at $q = 2$ fires on 3.8 percent of benign steps against a design level of 0.1 percent per channel; this is calibration drift, the calibration tails coming from the training weeks and the test traces from other weeks, and the cross band is more sensitive to it than the self band, whose own rule fires on 1.4 percent. Per attack group TALLY reaches 0.935 on the boiler, 0.843 on the water treatment loop, 0.952 on attacks that touch several subsystems and 0.784 on the turbine, where the autoencoder is best at 0.940.

Table 2 Clean detection, three seeds.

				Own rule		
Detector	$k^\star$	AUROC	Recall at 1%	Recall	FPR	Events
HAI						
TALLY	4	0.915 ± 0.005	0.739 ± 0.012	0.787	0.038	0.95
TALLY, cross band only	4	0.885 ± 0.005	0.694 ± 0.007	0.744	0.048	0.95
TALLY, self band only	any	0.906 ± 0.004	0.646 ± 0.008	0.673	0.014	0.93
TALLY, mixed members	4	0.926 ± 0.002	0.733 ± 0.005	0.761	0.014	0.99
TALLY, linear members	4	0.912 ± 0.002	0.705 ± 0.009	0.804	0.044	0.98
Full-block predictor	0	0.961 ± 0.002	0.874 ± 0.004	0.922	0.081	1.00
Autoencoder	0	0.932 ± 0.003	0.705 ± 0.001	0.731	0.017	0.92
Joint forecaster	0	0.947 ± 0.002	0.795 ± 0.002	0.836	0.017	0.98
Mahalanobis	0	0.904 ± 0.000	0.719 ± 0.000	0.727	0.012	0.94
Isolation forest	0	0.742 ± 0.017	0.114 ± 0.041	0.110	0.010	0.36
Feeder						
TALLY	5	0.941 ± 0.000	0.689 ± 0.004	0.718	0.013	0.91
TALLY, cross band only	5	0.946 ± 0.000	0.756 ± 0.005	0.782	0.016	0.91
TALLY, self band only	any	0.677 ± 0.001	0.134 ± 0.003	0.148	0.013	0.87
TALLY, mixed members	5	0.831 ± 0.001	0.274 ± 0.007	0.324	0.014	0.89
TALLY, linear members	5	0.927 ± 0.001	0.616 ± 0.010	0.679	0.017	0.89
Full-block predictor	0	0.937 ± 0.001	0.567 ± 0.019	0.679	0.019	0.92
Autoencoder	0	0.920 ± 0.004	0.346 ± 0.004	0.699	0.031	0.87
Joint forecaster	0	0.915 ± 0.003	0.405 ± 0.019	0.530	0.018	0.94
Mahalanobis	0	0.927 ± 0.000	0.557 ± 0.000	0.677	0.031	0.77
Isolation forest	0	0.701 ± 0.003	0.024 ± 0.000	0.056	0.030	0.09
Morris						
TALLY	1	0.509 ± 0.003	0.034 ± 0.002	0.033	0.009	1.00
TALLY, linear members	1	0.519 ± 0.003	0.031 ± 0.000	0.030	0.009	1.00
Full-block predictor	0	0.535 ± 0.004	0.025 ± 0.001	0.026	0.010	1.00
Autoencoder	0	0.506 ± 0.004	0.032 ± 0.001	0.027	0.008	1.00
Mahalanobis	0	0.499 ± 0.000	0.030 ± 0.000	0.030	0.011	1.00
Isolation forest	0	0.507 ± 0.002	0.020 ± 0.001	0.019	0.010	1.00

Feeder. Here the tolerant design is the best detector outright: TALLY with blocks of three meters, tolerating five compromised meters, reaches 0.941 AUROC and 0.689 recall at one percent, against 0.937 and 0.567 for the full-block predictor, 0.927 and 0.557 for the Mahalanobis detector, 0.920 and 0.346 for the autoencoder and 0.915 and 0.405 for the joint forecaster. The cross band alone is marginally better still at 0.946 and 0.756, and the self band alone collapses to 0.677 and 0.134: at fifteen-minute resolution a customer's demand is weakly predictable from its own last hour, whereas a surge is a large departure from what the other meters predict. The mixed member collapses to 0.831 and 0.274, because the own-past input explains most of the training variance and the block input is under-used; the two-band design avoids this by construction, which is why the bands are kept separate. The full-block predictor loses to the small blocks because one member of fixed width reading 93 channels overfits the calibration split, whereas eleven members reading nine channels each vote through a median. The own rule at $q = 2$ fires on 1.3 percent of benign steps against a bound of 4.8 percent from Theorem 1, calibration and test being exchangeable by construction.

Morris. Every detector is at chance, with AUROC between 0.499 and 0.535: the attack rows include samples recorded before the attack has any physical effect, the replayed fault data are coherent across relays because they are recordings of real faults, and the natural faults in training widen normality until a

remote trip looks like one more disturbance. We keep the deployment for injection and induction, for which its relay structure is right, and report the failure rather than tune it away.

7.2 The tolerance ladder

Figure 5 runs TALLY at every block size from one node to all nodes, with the self band present throughout, at $k = 0$, $q = 2$; Figure 6 then fixes the deployed block size and raises the tolerance k at which the thresholds are set.

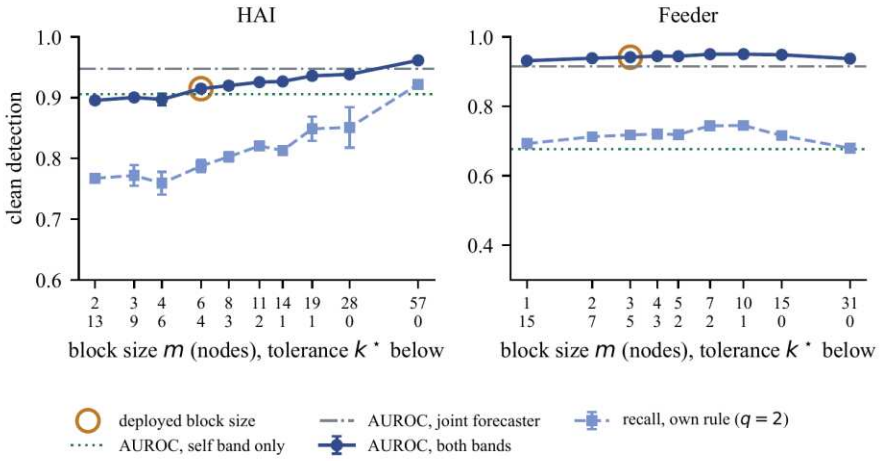


Fig. 5 Clean detection along the tolerance ladder.

HAI pays for tolerance; the feeder does not. On HAI the AUROC falls from 0.961 with one member reading all 56 other tags ($k^\star = 0$) through 0.927 at four members ($k^\star = 1$) and 0.915 at nine ($k^\star = 4$) to 0.900 at nineteen ($k^\star = 9$), where it meets the self band alone at 0.906; blocks of two to four tags are slightly worse than no cross band at all, because three tags of a boiler predict a fourth less well than its own last ninety seconds do, and the noisy residual adds false excesses without adding detections. Recall at the own rule follows the same course, from 0.922 to 0.767. On the feeder every block size between one and ten meters beats the full-block predictor: 0.931 at blocks of one meter ($k^\star = 15$), 0.941 at three ($k^\star = 5$) and 0.950 at seven and ten ($k^\star = 2$ and 1), against 0.937 for the single member reading all 31 other meters. A HAI operator who tolerates four compromised tags gives up five points of AUROC and thirteen of recall; a feeder operator who tolerates five compromised meters gives up nothing.

Raising k at a fixed design. Once the block size is fixed, tolerance k costs detection twice: the guarded rank $r_0 + k$ moves the thresholds outward and the quorum $q = k + 1$ needs more flagged nodes. On HAI with blocks of six tags, recall at the certified configuration goes from 0.892 at $k = 0$ ($q = 1$) to 0.750 at $k = 1$ and 0.606 at $k = 2$, the false-alarm rate of the own rule from 0.239 to 0.020 and 0.007, and the event rate stays at 0.95 and 0.87 because a labelled attack that lasts minutes needs only one flagged step to count. On the feeder recall goes from 0.932 to 0.685 and 0.438 and the event rate from 0.98 to 0.89 and 0.67: surges are visible at two to four meters, so a quorum of three often exceeds what one surge exposes, which is Proposition 2(b) at work.

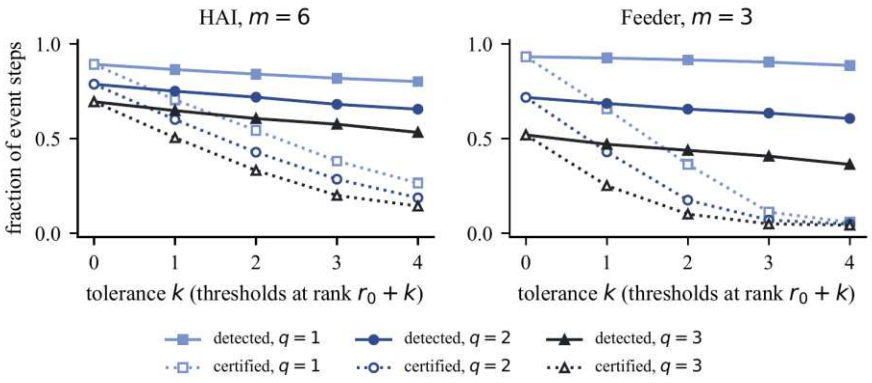


Fig. 6 Detected against certified event steps as tolerance rises.

Certified fraction. The certified fraction is always below the detected fraction, and the gap widens with k because the certificate needs $q + k$ nodes whose inconsistency survives the removal of k members from the honest side of the median. On HAI 0.601 of the event steps are certified at $k = 1$ against 0.750 detected, and 0.331 against 0.606 at $k = 2$, with 0.88 and 0.67 of the labelled attacks containing a certified step; on the feeder 0.429 against 0.685 and 0.101 against 0.438, with 0.67 and 0.48 of the surges. In Figure 6 the detected fraction declines slowly with k while the certified fraction declines quickly, because every tolerated node is one more the attacker may take from the flagged set; Section 7.3 tests whether it can take more.

7.3 Node compromise

Tables 3 and 4 and Figure 7 report what k compromised nodes achieve against each detector: detection of real events under the strongest suppression attack with random or insider node choice, the certified fraction, and the false-alarm rate under the strongest induction attack; a dash in the certified column marks a detector with no tolerance, for which no certificate exists. Morris is used for induction only.

Induction is the decisive experiment. On HAI one compromised tag, chosen by the insider, raises the false-alarm rate of the joint forecaster and the Mahalanobis detector from 0.012 and 0.010 to 1.000 and that of the autoencoder from 0.013 to 0.752; with four tags every joint detector reaches 1.000. The full-block predictor, TALLY's decision layer with a single member per channel, goes from 0.090 to 0.948 with one tag, because with $M = 1$ a tag that appears in every other tag's block can push every honest evidence value wherever it likes: Remark 1 measured. TALLY in its certified configuration goes from 0.020 to 0.043 at $k = 1$ under the insider's choice and to 0.160 under the random choice, which is the stronger choice for induction because the insider spends its nodes on tags already in excess; to 0.013 and 0.071 at $k = 2$; and to 0.003 and 0.024 at $k = 4$. The nominal bound of Theorem 1 at $k = 1$, $q = 2$ is 0.057, and the random-choice rate exceeds it. The excess is not the attacker's doing: honest tags on HAI are in excess at 0.003 per tag without any compromise, three times the calibration level, because the calibration weeks differ from the test weeks; substituting the measured rate into the same bound gives 0.17, which the attacked rate does not exceed, and under attack the honest tags are flagged at 0.002 to 0.004 against 0.001 to 0.003 without it, inside the sandwich of Theorem 1. The fixed quorum $q = 2$ shows what the theorem forbids: it holds at $k = 1$ (0.043), fails at $k = 2$ (0.740 under the insider, 0.907 at random) and fails completely at $k = 4$ (0.998). The isolation forest, which the search attack barely moves

(0.007 to 0.015 at $k = 1$), is the exception among the baselines and also the detector that does not detect (recall 0.125).

Table 3 Compromise on HAI: detection of real events and false alarms, three seeds.

		Detection of real events				False alarms		
Detector	k	clean	random	insider	certified	clean	random	insider
TALLY	1	0.774	0.765	0.645	0.613	0.020	0.160	0.043
	2	0.621	0.599	0.423	0.312	0.007	0.071	0.013
	4	0.312	0.268	0.043	0.028	0.003	0.024	0.003
TALLY, $q = 2$ fixed	1	0.774	0.765	0.645	0.613	0.020	0.160	0.043
	2	0.741	0.719	0.501	0.423	0.014	0.907	0.740
	4	0.677	0.657	0.224	0.154	0.009	1.000	0.998
TALLY, self band only	1	0.698	0.692	0.556	0.556	0.014	0.041	0.014
	2	0.556	0.539	0.329	0.329	0.008	0.040	0.008
	4	0.329	0.288	0.045	0.045	0.004	0.039	0.004
Full-block predictor	1	0.921	0.906	0.868	–	0.090	0.990	0.948
	2	0.892	0.870	0.748	–	0.019	1.000	0.997
	4	0.821	0.762	0.617	–	0.007	1.000	1.000
Autoencoder	1	0.739	0.734	0.652	–	0.013	0.343	0.752
	2	0.739	0.730	0.637	–	0.013	0.709	0.965
	4	0.739	0.726	0.512	–	0.013	0.983	1.000
Joint forecaster	1	0.859	0.835	0.780	–	0.012	0.857	1.000
	2	0.859	0.820	0.724	–	0.012	0.990	1.000
	4	0.859	0.777	0.646	–	0.012	1.000	1.000
Mahalanobis	1	0.752	0.752	0.677	–	0.010	0.810	1.000
	2	0.752	0.748	0.665	–	0.010	0.965	1.000
	4	0.752	0.738	0.632	–	0.010	1.000	1.000
Isolation forest	1	0.125	0.112	0.109	–	0.007	0.014	0.015
	2	0.125	0.098	0.095	–	0.007	0.032	0.028
	4	0.125	0.074	0.077	–	0.007	0.084	0.084

Suppression. Against a real attack in progress a compromised tag is worth less to the attacker than against a benign window, on every detector. With one insider-chosen tag the strongest attack lowers TALLY's detection from 0.774 to 0.645, the joint forecaster's from 0.859 to 0.780, the autoencoder's from 0.739 to 0.652 and the Mahalanobis detector's from 0.752 to 0.677. The certified fraction at $k = 1$ is 0.613, below the 0.645 that survived, as it must be, and no certified window lost its alarm under any attack in any seed, at any k. As k grows the picture turns against TALLY: at $k = 2$ its certified configuration detects 0.423 of the attacked event steps and at $k = 4$ only 0.043, against 0.637 and 0.512 for the autoencoder, because $q = k + 1$ demands five flagged tags at $k = 4$ and the released attacks rarely put five tags in excess once the four most affected are silenced. This must be read with the induction rows: a detector that the same attacker can make alarm on every benign window detects everything and nothing, so the autoencoder's 0.512 is not the detection rate of a usable detector. The self band alone behaves like TALLY with slightly lower recall (0.698 clean, 0.556 attacked at $k = 1$), and its certified fraction equals its attacked recall exactly, because no honest self residual can be moved. Against the joint detectors the gradient attack is the strongest in every cell, whereas against TALLY the gradient-free search is as strong or stronger in every suppression cell and the nominal attack is the strongest at $k = 4$: the attacker of a median-based detector needs no gradients, and the gradient it computes is not masked [46].

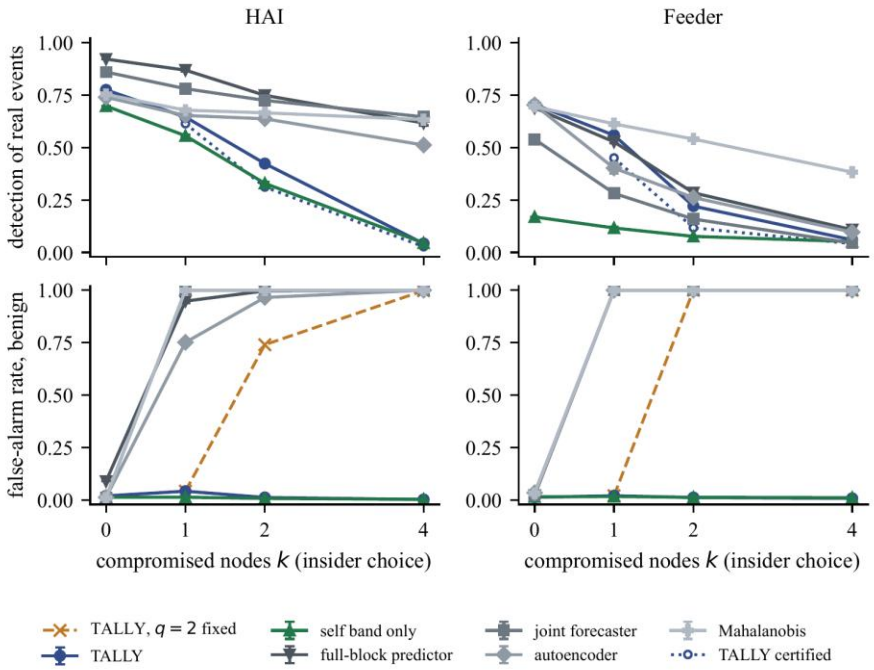


Fig. 7 Detection and false alarms under insider compromise.

Feeder. Where calibration and test are exchangeable the bound of Theorem 1 holds as stated (Table 4). Under induction by one compromised meter TALLY's false-alarm rate goes from 0.014 to 0.021 under the insider and 0.068 under the random choice, against a bound of 0.096; at $k = 2$ to 0.012 and 0.060, and at $k = 4$ to 0.009 and 0.042, with the same bound. Honest meters are flagged at 0.003 to 0.005 with and without the attacker. Every joint detector, the full-block predictor included, reaches a false-alarm rate of 1.000 from one meter, and the full-block predictor flags 0.635 of the honest meters while doing so, against 0.005 without the attacker. The fixed quorum $q = 2$ again holds at $k = 1$ and reaches 1.000 at $k = 2$. Under suppression TALLY loses more than on HAI, from 0.705 to 0.558 at $k = 1$ under the insider, from 0.463 to 0.221 at $k = 2$ and from 0.137 to 0.060 at $k = 4$, with certified fractions of 0.450, 0.118 and 0.042 and no certified alarm removed: the surges touch two to four meters, so the insider who silences the two most affected takes most of the evidence and the quorum of $k + 1$ takes the rest. The joint detectors are hit harder by the same attacker, the autoencoder falling from 0.702 to 0.402, 0.262 and 0.097 and the joint forecaster from 0.539 to 0.282, 0.159 and 0.045, and only the Mahalanobis detector keeps more (0.612, 0.540, 0.382), at a false-alarm rate of 1.000. Here the gradient attack is the strongest against TALLY too and the nominal attack the weakest, the reverse of HAI; reporting the strongest per window is what makes deployments comparable.

Table 4 Compromise on the feeder: detection of real events and false alarms, three seeds.

		Detection of real events				False alarms		
Detector	k	clean	random	insider	certified	clean	random	insider
TALLY	1	0.705	0.686	0.558	0.450	0.014	0.068	0.021
	2	0.463	0.414	0.221	0.118	0.009	0.060	0.012
	4	0.137	0.092	0.060	0.042	0.005	0.042	0.009
TALLY, $q = 2$ fixed	1	0.705	0.686	0.558	0.450	0.014	0.068	0.021
	2	0.682	0.654	0.342	0.192	0.012	1.000	1.000
	4	0.651	0.577	0.131	0.052	0.011	1.000	1.000
TALLY, self band only	1	0.169	0.166	0.117	0.115	0.016	0.053	0.018
	2	0.115	0.107	0.077	0.072	0.012	0.052	0.014
	4	0.072	0.068	0.052	0.051	0.008	0.050	0.012
Full-block predictor	1	0.688	0.588	0.527	–	0.023	1.000	1.000
	2	0.542	0.340	0.284	–	0.015	1.000	1.000
	4	0.359	0.125	0.108	–	0.010	1.000	1.000
Autoencoder	1	0.703	0.665	0.402	–	0.035	1.000	1.000
	2	0.703	0.628	0.262	–	0.035	1.000	1.000
	4	0.703	0.538	0.097	–	0.035	1.000	1.000
Joint forecaster	1	0.539	0.405	0.282	–	0.023	1.000	1.000
	2	0.539	0.320	0.159	–	0.023	1.000	1.000
	4	0.539	0.208	0.045	–	0.023	1.000	1.000
Mahalanobis	1	0.695	0.682	0.613	–	0.033	1.000	1.000
	2	0.695	0.667	0.540	–	0.033	1.000	1.000
	4	0.695	0.617	0.383	–	0.033	1.000	1.000
Isolation forest	1	0.071	0.065	0.065	–	0.028	0.030	0.030
	2	0.071	0.061	0.064	–	0.028	0.034	0.033
	4	0.071	0.053	0.052	–	0.028	0.088	0.090

Morris. A relay node is twenty-nine phasor quantities at once, the design has four members per channel and tolerates one relay. One compromised relay drives the autoencoder, the Mahalanobis detector and the full-block predictor to a false-alarm rate of 1.000 from 0.006, 0.008 and 0.012, the full-block predictor flagging every honest relay while doing so; the isolation forest reaches 0.282 with one relay and 0.978 with two. TALLY at $k = 1$, $q = 2$ goes from 0.005 to 0.012 under the insider and 0.026 under the random choice, against a bound of 0.124, and honest relays are flagged at 0.004 to 0.009 against 0.002 to 0.004 without the attacker, inside the sandwich of Lemma 1. The largest single node in this study moves the honest evidence by less than a hundredth.

One day on the feeder. Figure 8 follows a test day at $k = 2$, $q = 3$, on which two buses raise their demand for 3.75 hours from four in the morning. The cross evidence of the surged meter climbs to 2.7 times its threshold and stays above it for the whole surge, while its self evidence peaks at the onset (1.9 times its threshold) and returns below it within an hour, the surged level having become the meter’s own recent past; an honest meter’s cross evidence never crosses its threshold. The flag count reaches seven meters, the alarm fires on 16 steps, one of them a benign step at the end of the surge, and 12 surge steps carry the certificate: on those steps no choice of two meters, reporting anything, would have silenced the alarm.

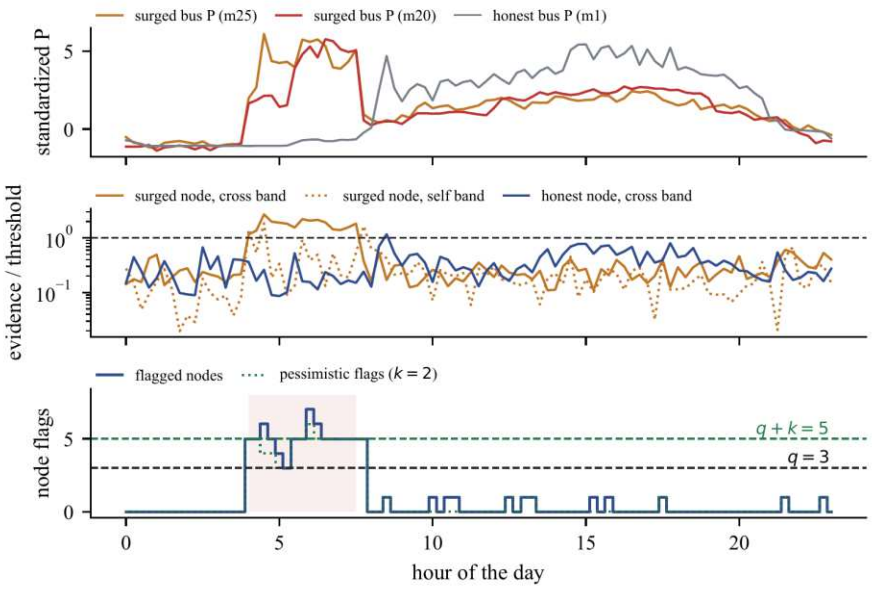


Fig. 8 One surge day on the feeder, with flags and certificates.

7.4 False data injection

Injection is the attack in which the compromised nodes are the event. Table 5 and Figure 9 report, for k lying nodes, whether the liars are ranked highest and flagged, whether honest nodes are flagged, and whether the system alarm fires, the last two being what Theorem 1 and Proposition 2(a) constrain; the joint detectors have no node flags, so those columns carry a dash for them.

HAI. With one lying tag, TALLY ranks it first in 0.121 of the windows for a bias of one standard deviation, 0.349 for two and 0.606 for four, and flags it in 0.070, 0.249 and 0.518 of them; a frozen value or a doubled deviation is caught in fewer than one window in ten. The cross band does the work (0.384 and 0.645 alone at two and four standard deviations); the self band alone stays near 0.05, because a lie held over the whole window is consistent with its own past by construction; the mixed member reaches only 0.117 and 0.200, a second reason to keep the bands separate. Honest tags are flagged at the same rate with and without the lie, 0.003 per node at $k = 1$ and 0.001 at $k = 2$ and 4, and the system alarm stays near its clean rate, 0.020 against 0.022 to 0.050 at $k = 1$ for biases up to two standard deviations, rising to 0.083 only for a four-standard-deviation lie that puts one flag on the liar and needs one honest flag more. The full-block predictor shows the opposite: one lying tag raises the honest flag rate from 0.018 to 0.028 at half a standard deviation and to 0.087 at four, and the system alarm from 0.084 to 0.534, so one lying tag alarms the plant at will. The joint forecaster ranks the liar first in 0.415 of the windows at one standard deviation, better than TALLY at small lies because it reads the tag's own past together with every other tag, but its alarm goes from 0.016 to 0.329 at two standard deviations and 0.495 at four: the attribution is usable, the alarm is not.

Table 5 Injection on HAI: attribution and flags, three seeds.

		Precision at k					Flag rates, bias 2σ		
Detector	k	1σ	2σ	4σ	scale	freeze	liar	honest	alarm
TALLY	1	0.12	0.35	0.61	0.08	0.07	0.25	0.003	0.050
	2	0.16	0.39	0.62	0.12	0.09	0.17	0.002	0.011
	4	0.18	0.35	0.51	0.19	0.08	0.03	0.001	0.003
TALLY, cross band only	1	0.14	0.38	0.65	0.05	0.07	0.29	0.003	0.056
	2	0.19	0.43	0.66	0.10	0.10	0.21	0.001	0.007
	4	0.25	0.47	0.64	0.15	0.14	0.04	0.000	0.000
TALLY, self band only	1	0.05	0.05	0.08	0.11	0.00	0.01	0.001	0.017
	2	0.07	0.08	0.12	0.14	0.01	0.01	0.001	0.011
	4	0.12	0.13	0.17	0.20	0.02	0.01	0.002	0.005
TALLY, mixed members	1	0.05	0.12	0.20	0.13	0.03	0.03	0.001	0.015
	2	0.09	0.16	0.25	0.16	0.05	0.02	0.001	0.007
	4	0.14	0.22	0.30	0.20	0.07	0.01	0.000	0.000
Full-block predictor	1	0.31	0.43	0.48	0.10	0.15	0.72	0.063	0.744
	2	0.40	0.45	0.47	0.22	0.23	0.72	0.105	0.864
	4	0.43	0.45	0.45	0.26	0.26	0.72	0.185	0.942
Autoencoder	1	0.09	0.40	0.58	0.07	0.07	–	–	0.026
	2	0.18	0.47	0.58	0.11	0.12	–	–	0.048
	4	0.31	0.54	0.60	0.16	0.16	–	–	0.196
Joint forecaster	1	0.42	0.50	0.59	0.12	0.13	–	–	0.329
	2	0.42	0.48	0.54	0.15	0.16	–	–	0.592
	4	0.45	0.49	0.52	0.21	0.19	–	–	0.874

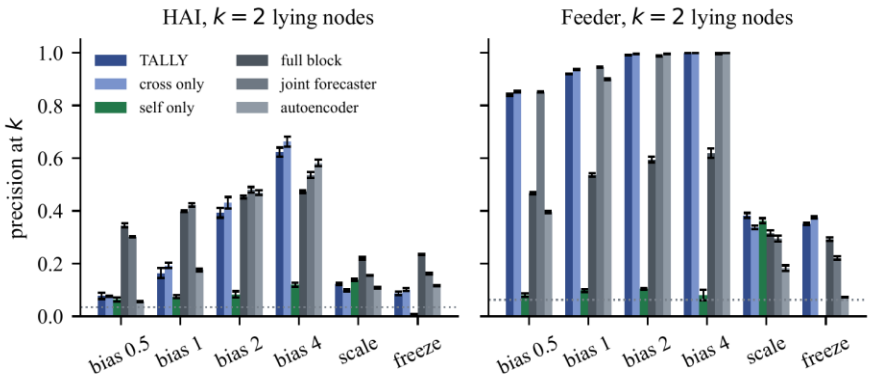


Fig. 9 Precision of the flagged nodes under injection.

Feeder. Where the cross band is strong, attribution is strong (Table 6). A meter with a persistent bias of half a standard deviation is ranked first in 0.838 of the windows, one standard deviation in 0.923, and two or more in 0.990 or better, and the lying meter is flagged in 0.762, 0.854 and 0.961 of them; with two or four liars the precision at k is the same to within a point. Honest meters are flagged at 0.003 with or without the lie, and the system alarm rises from 0.010 to at most 0.055 at $k = 1$, under the bound of 0.096, and to at most 0.036 at $k = 4$. Frozen and rescaled lies are attributed in about a third of the windows, because a demand held for an hour or with doubled swings is inside what a real customer does. The self band alone attributes nothing (0.044 to 0.059), the mixed member is a few points behind the two-band design and the cross band alone a few points ahead. The full-block predictor again shows what zero tolerance means: one lying meter with a half-standard-deviation bias raises the honest flag rate from 0.004

to 0.055 (0.537 at four standard deviations) and the system alarm from 0.017 to 0.584 (0.968 at four); the autoencoder and the joint forecaster alarm on 0.654 and 0.822 of the windows for a two-standard-deviation lie on one meter.

Table 6 Injection on the feeder: attribution and flags, three seeds.

		Precision at k					Flag rates, bias 2σ		
Detector	k	1σ	2σ	4σ	scale	freeze	liar	honest	alarm
TALLY	1	0.92	0.99	1.00	0.32	0.33	0.96	0.003	0.052
	2	0.92	0.99	1.00	0.38	0.35	0.96	0.003	0.044
	4	0.92	0.99	1.00	0.44	0.35	0.95	0.003	0.030
TALLY, cross band only	1	0.94	1.00	1.00	0.29	0.35	0.98	0.003	0.074
	2	0.94	1.00	1.00	0.34	0.38	0.98	0.003	0.061
	4	0.94	1.00	1.00	0.38	0.40	0.97	0.002	0.033
TALLY, self band only	1	0.05	0.06	0.05	0.28	0.00	0.01	0.004	0.013
	2	0.10	0.10	0.08	0.36	0.00	0.00	0.004	0.010
	4	0.18	0.19	0.15	0.46	0.00	0.01	0.004	0.005
TALLY, mixed members	1	0.88	0.97	1.00	0.39	0.33	0.94	0.003	0.056
	2	0.88	0.97	1.00	0.45	0.34	0.93	0.003	0.042
	4	0.89	0.97	1.00	0.50	0.36	0.91	0.002	0.025
Full-block predictor	1	0.85	0.92	0.93	0.26	0.26	0.97	0.392	0.924
	2	0.54	0.59	0.62	0.32	0.29	0.98	0.619	0.994
	4	0.28	0.30	0.32	0.40	0.34	0.98	0.797	1.000
Autoencoder	1	0.86	0.99	1.00	0.13	0.05	–	–	0.654
	2	0.90	1.00	1.00	0.18	0.07	–	–	0.993
	4	0.94	1.00	1.00	0.28	0.13	–	–	1.000
Joint forecaster	1	0.95	0.99	1.00	0.22	0.18	–	–	0.822
	2	0.95	0.99	1.00	0.30	0.22	–	–	0.986
	4	0.95	0.98	0.99	0.40	0.28	–	–	1.000

Morris. A relay that biases its twenty-nine phasor quantities by one standard deviation is ranked first in 0.862 of the windows and by two in 0.976; it is flagged in 0.085 and 0.417 of them and in 0.999 at four standard deviations, while honest relays stay at 0.004 to 0.006 and the system alarm at 0.004 to 0.018. The full-block predictor flags honest relays in 0.154 and 0.395 of the windows at two and four standard deviations and alarms in 0.484 and 0.845 of them; the autoencoder alarms in 0.867 at four. A frozen or rescaled relay is attributed by no detector (precision at the chance level of one in five), because one sample of a frozen phasor is a plausible phasor. That a lying relay is separated so cleanly on a deployment whose released attack labels are at chance says what the two tasks share: the relays are physically coupled, so a lie that breaks the coupling is visible, whereas the released attacks do not break it.

8 Discussion

Tolerance is a dial with a price that depends on the physics. On HAI, where a tag's own trajectory is its best predictor, every step of tolerance bought by shrinking the blocks costs clean detection, from 0.961 with no tolerance to 0.915 at four tolerated tags, and the certified fraction falls faster than the detected fraction because the released attacks touch few tags. On the feeder, where a meter is best predicted by the other meters, the tolerant design is the best detector we have and tolerance up to five meters costs nothing measurable. A deployment can read its own position on this dial from calibration data alone, by comparing the calibration residuals of the self band with those of a cross band at the candidate block size.

What the certificate buys. A certified alarm is a statement that no choice of k nodes, reporting anything, would have silenced it. A numerical audit of Theorem 2, in which the members that read a compromised node are set to arbitrary values, over every subset of one and two compromised meters on the feeder, every single compromised tag on HAI and six hundred random pairs of tags, and the strongest adaptive attacks of Section 7.3, found no certified alarm that was removed. Everything above the certified curve in Figure 6 is detection that depends on the attacker not choosing the right nodes.

Exchangeability. On the feeder, where calibration and test are exchangeable by construction, the measured rates sit under the bound of Theorem 1. On HAI the calibration tails come from the training weeks and the test traces from other weeks, and the cross band's own rule fires on several percent of benign steps at a design level of a tenth of a percent per channel. This is calibration drift, the effect that moves every calibrated detector in Table 2 off its nominal rate, and not something a compromised node causes: under attack the honest-node flag rate stayed within a thousandth of its clean value in every cell, and the alarm rate under compromise is what the drifted honest rate predicts. A deployment that wants the bound as stated must recalibrate on recent attack-free data, which touches only the thresholds.

Design decisions. The quorum counts nodes, not channels, because a node with several channels would otherwise contribute several flags; a first version counted channels and a single compromised meter defeated it. The two bands are kept separate because a member that reads both inputs leans on the one that explains most of the training variance, and on the feeder that is the wrong one.

Limitations. The Markov bound of Theorem 1 is loose by an order of magnitude and would tighten under an independence assumption across nodes that we did not make. The adaptive attacks are three, run once, on subsamples of the test streams; a stronger attacker could use more steps or an attack designed against the median, though the gradient-free search already matches or beats the gradient attack in most cells. The feeder's physics are simulated, its demand is real, and the surge model is a stylized version of a documented attack. Morris is a negative result for every detector, including ours. Three seeds bound the variance of training, not of the released test data. The framework says nothing about an attacker who compromises the aggregator itself or corrupts the training data; those threats belong to authentication, access control and robust learning, which it assumes in place.

9 Conclusion

We built a cybersecurity framework for IoT-based smart grid systems around one question: how many sensing nodes can an attacker control before a learned anomaly detector stops being trustworthy? For the detectors in use today the answer is one; for TALLY it is a design parameter. Every channel is scored by an ensemble of predictors that each read one block of other nodes, plus a predictor of the channel's own past, and the evidence of an honest channel is provably confined between two honest order statistics under any compromise of k nodes; guarded thresholds keep honest nodes from being framed, a quorum keeps k compromised nodes from raising an alarm on their own, and a per-window certificate says when an alarm would have survived every such compromise. On a real generation testbed, a relay testbed and a feeder driven by real demand, the price of tolerance ranges from several points of detection to none, the certificates held against every attack we mounted, injecting nodes were the ones flagged, and the cases where the framework is weaker than an unprotected detector, or where no detector works, are reported with their numbers.

Acknowledgements

During the preparation of this work the authors used large language model tools for language editing and as a coding assistant. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Declarations

Funding.

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. Not applicable. This study used only publicly available datasets and simulations and did not involve human participants or animals.

Consent to Participate declaration: not applicable.

Consent to Publish declaration: not applicable.

Clinical trial number: not applicable.

Data availability. All datasets analysed in this study are publicly available: HAI from the National Security Research Institute of Korea (https://github.com/icsdataset/hai), the power system attack datasets from Mississippi State University and Oak Ridge National Laboratory [22], the ElectricityLoadDiagrams20112014 data from the UCI Machine Learning Repository [25], and the IEEE 33-bus feeder with pandapower.

Author contributions.

References

[1] Ghasempour, A.: Internet of Things in Smart Grid: Architecture, Applications, Services, Key Technologies, and Challenges. Inventions 4(1), 22 (2019) https://doi.org/10.3390/inventions4010022

[2] Saleem, Y.; Crespi, N.; Rehmani, M.H.; Copeland, R.: Internet of Things-Aided Smart Grid: Technologies, Architectures, Applications, Prototypes, and Future Research Directions. IEEE Access 7, 62962-63003 (2019) https://doi.org/10.1109/ACCESS.2019.2913984

[3] Musleh, A.S.; Chen, G.; Dong, Z.Y.: A Survey on the Detection Algorithms for False Data Injection Attacks in Smart Grids. IEEE Transactions on Smart Grid 11(3), 2218-2234 (2020) https://doi.org/10.1109/TSG.2019.2949998

[4] Kravchik, M.; Shabtai, A.: Efficient Cyber Attack Detection in Industrial Control Systems Using Lightweight Neural Networks and PCA. IEEE Transactions on Dependable and Secure Computing 19(4), 2179-2197 (2022) https://doi.org/10.1109/TDSC.2021.3050101

[5] Deng, A.; Hooi, B.: Graph Neural Network-Based Anomaly Detection in Multivariate Time Series. Proceedings of the 35th AAAI Conference on Artificial Intelligence 35(5), 4027-4035 (2021) https://doi.org/10.1609/aaai.v35i5.16523

[6] Mo, Y.; Kim, T.H.; Brancik, K.; Dickinson, D.; Lee, H.; Perrig, A.; Sinopoli, B.: Cyber-Physical Security of a Smart Grid Infrastructure. Proceedings of the IEEE 100(1), 195-209 (2012) https://doi.org/10.1109/JPROC.2011.2161428

[7] Gunduz, M.Z.; Das, R.: Cyber-Security on Smart Grid: Threats and Potential Solutions. Computer Networks 169, 107094 (2020) https://doi.org/10.1016/j.comnet.2019.107094

[8] Liang, G.; Zhao, J.; Luo, F.; Weller, S.R.; Dong, Z.Y.: A Review of False Data Injection Attacks Against Modern Power Systems. IEEE Transactions on Smart Grid 8(4), 1630-1638 (2017) https://doi.org/10.1109/TSG.2015.2495133

[9] Badar, H.M.S.; Qadri, S.; Shamshad, S.; Ayub, M.F.; Mahmood, K.; Kumar, N.: An Identity Based Authentication Protocol for Smart Grid Environment Using Physical Uncloneable Function. IEEE Transactions on Smart Grid 12(5), 4426-4434 (2021) https://doi.org/10.1109/TSG.2021.3072244

[10] Badar, H.M.S.; Obaidat, M.S.; Mahmood, K.; Saher, N.; Ayub, M.F.; Khan, D.M.: An Access Control Protocol for IoT-Based Critical Infrastructure in Smart Grid Environment. International Journal of Communication Systems 35(8), e5115 (2022) https://doi.org/10.1002/dac.5115

[11] Badar, H.M.S.; Mahmood, K.; Akram, W.; Ghaffar, Z.; Umar, M.; Das, A.K.: Secure Authentication Protocol for Home Area Network in Smart Grid-Based Smart Cities. Computers and Electrical Engineering 108, 108721 (2023) https://doi.org/10.1016/j.compeleceng.2023.108721

[12] Zahoor, A.; Mahmood, K.; Saleem, M.A.; Badar, H.M.S.; Le, T.; Das, A.K.: Lightweight Authenticated Key Agreement Protocol for Smart Power Grid Systems Using PUF. IEEE Open Journal of the Communications Society 5, 3568-3580 (2024) https://doi.org/10.1109/OJCOMS.2024.3409451

[13] Badar, H.M.S.; Ahmed, S.; Kajla, N.I.; Fan, G.; Zhang, C.: Q-BLAISE: Quantum-Resilient Blockchain and AI-Enhanced Security Protocol for Smart Grid IoT. IEEE Transactions on Consumer Electronics 71(2), 4959-4971 (2025) https://doi.org/10.1109/TCE.2025.3570969

[14] Liu, Y.; Ning, P.; Reiter, M.K.: False Data Injection Attacks Against State Estimation in Electric Power Grids. ACM Transactions on Information and System Security 14(1), 1-33 (2011) https://doi.org/10.1145/1952982.1952995

[15] Liang, G.; Weller, S.R.; Zhao, J.; Luo, F.; Dong, Z.Y.: The 2015 Ukraine Blackout: Implications for False Data Injection Attacks. IEEE Transactions on Power Systems 32(4), 3317-3318 (2017) https://doi.org/10.1109/TPWRS.2016.2631891

[16] Levine, A.; Feizi, S.: Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation. Proceedings of the 34th AAAI Conference on Artificial Intelligence 34(04), 4585-4593 (2020) https://doi.org/10.1609/aaai.v34i04.5888

[17] Levine, A.; Feizi, S.: (De)Randomized Smoothing for Certifiable Defense Against Patch Attacks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (2020)

[18] Fawzi, H.; Tabuada, P.; Diggavi, S.: Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks. IEEE Transactions on Automatic Control 59(6), 1454-1467 (2014) https://doi.org/10.1109/TAC.2014.2303233

[19] Pasqualetti, F.; Dörfler, F.; Bullo, F.: Attack Detection and Identification in Cyber-Physical Systems. IEEE Transactions on Automatic Control 58(11), 2715-2729 (2013) https://doi.org/10.1109/TAC.2013.2266831

[20] Shin, H.; Lee, W.; Yun, J.; Kim, H.: HAI 1.0: HIL-Based Augmented ICS Security Dataset. Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET) (2020) https://www.usenix.org/conference/cset20/presentation/shin

[21] Shin, H.; Lee, W.; Yun, J.; Min, B.: Two ICS Security Datasets and Anomaly Detection Contest on the HIL-Based Augmented ICS Testbed. Proceedings of the 14th Cyber Security Experimentation and Test Workshop (CSET), 36-40 (2021) https://doi.org/10.1145/3474718.3474719

[22] Pan, S.; Morris, T.; Adhikari, U.: Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems. IEEE Transactions on Smart Grid 6(6), 3104-3113 (2015) https://doi.org/10.1109/TSG.2015.2409775

[23] Baran, M.E.; Wu, F.F.: Network Reconfiguration in Distribution Systems for Loss Reduction and Load Balancing. IEEE Transactions on Power Delivery 4(2), 1401-1407 (1989) https://doi.org/10.1109/61.25627

[24] Thurner, L.; Scheidler, A.; Schäfer, F.; Menke, J.; Dollichon, J.; Meier, F.; Meinecke, S.; Braun, M.: pandapower: An Open-Source Python Tool for Convenient Modeling, Analysis, and Optimization of Electric Power Systems. IEEE Transactions on Power Systems 33(6), 6510-6521 (2018) https://doi.org/10.1109/TPWRS.2018.2829021

[25] Trindade, A.: ElectricityLoadDiagrams20112014. (2015) https://doi.org/10.24432/C58C86

[26] Soltan, S.; Mittal, P.; Poor, H.V.: BlackIoT: IoT Botnet of High Wattage Devices Can Disrupt the Power Grid. Proceedings of the 27th USENIX Security Symposium, 15-32 (2018) https://www.usenix.org/conference/usenixsecurity18/presentation/soltan

[27] Huang, B.; Cárdenas, A.A.; Baldick, R.: Not Everything is Dark and Gloomy: Power Grid Protections Against IoT Demand Attacks. Proceedings of the 28th USENIX Security Symposium, 1115-1132 (2019) https://www.usenix.org/conference/usenixsecurity19/presentation/huang

[28] Urbina, D.I.; Giraldo, J.A.; Cardenas, A.A.; Tippenhauer, N.O.; Valente, J.; Faisal, M.; Ruths, J.; Candell, R.; Sandberg, H.: Limiting the Impact of Stealthy Attacks on Industrial Control Systems. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 1092-1105 (2016) https://doi.org/10.1145/2976749.2978388

[29] Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; Soderstrom, T.: Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 387-395 (2018) https://doi.org/10.1145/3219819.3219845

[30] Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; Zuluaga, M.A.: USAD: UnSupervised Anomaly Detection on Multivariate Time Series. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 3395-3404 (2020) https://doi.org/10.1145/3394486.3403392

[31] Wu, R.; Keogh, E.J.: Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. IEEE Transactions on Knowledge and Data Engineering 35(3), 2421-2429 (2023) https://doi.org/10.1109/TKDE.2021.3112126

[32] Erba, A.; Taormina, R.; Galelli, S.; Pogliani, M.; Carminati, M.; Zanero, S.; Tippenhauer, N.O.: Constrained Concealment Attacks Against Reconstruction-Based Anomaly Detectors in Industrial Control Systems. Proceedings of the 36th Annual Computer Security Applications Conference (ACSAC), 480-495 (2020) https://doi.org/10.1145/3427228.3427660

[33] Zizzo, G.; Hankin, C.; Maffeis, S.; Jones, K.: Adversarial Attacks on Time-Series Intrusion Detection for Industrial Control Systems. Proceedings of the 19th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 899-910 (2020) https://doi.org/10.1109/TrustCom50675.2020.00121

[34] Fan, G.; Huang, Q.; Ma, J.; Badar, H.M.S.; Zhang, C.; Yuan, K.: Anomal-EF: Self-Supervised Learning and Multi-Head Attention Based Anomalous Node Detection in IoT Networks. Cluster Computing 29(9), 544 (2026) https://doi.org/10.1007/s10586-026-06350-z

[35] Fan, G.; Huang, Q.; Ma, J.; Badar, H.M.S.; Zhang, C.; Yuan, K.: Anomal-EFD: A Self-Supervised Model for Anomaly Detection in Dynamic IoT Networks. Peer-to-Peer Networking and Applications 19(1), 33 (2025) https://doi.org/10.1007/s12083-025-02184-5

[36] Ilyas, M.; Kajla, N.I.; Madden, M.G.; Fan, G.; Zhang, C.; Badar, H.M.S.: PoisonShield-FL-NIDS: A Robust Defense Against Poisoning Attacks in Federated Learning Intrusion Detection. IEEE Internet of Things Journal 13(4), 6105-6115 (2026) https://doi.org/10.1109/JIOT.2025.3634357

[37] Shahid, M.I.; Badar, H.M.S.; Asghar, M.N.; Alaulamie, A.: Robust Federated Learning Under Data Heterogeneity and Adversarial Clients. IEEE Internet of Things Journal 13(13), 29015-29027 (2026) https://doi.org/10.1109/JIOT.2026.3685595

[38] Badar, H.M.S.; Kajla, N.I.; Mahmood, K.; Khan, M.W.; Fan, G.; Zhang, C.: ATTA-FL-Lite: Lightweight Byzantine-Robust Federated Learning for Resource-Constrained Medical IoT Devices. IEEE Internet of Things Journal 13(6), 11847-11858 (2026) https://doi.org/10.1109/JIOT.2025.3644701

[39] Huber, P.J.; Ronchetti, E.M.: Robust Statistics. Wiley (2009) https://doi.org/10.1002/9780470434697

[40] Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J.: Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. Advances in Neural Information Processing Systems 30 (NIPS 2017) (2017)

[41] Vovk, V.; Gammerman, A.; Shafer, G.: Algorithmic Learning in a Random World. Springer (2005) https://doi.org/10.1007/b106715

[42] Bates, S.; Candès, E.; Lei, L.; Romano, Y.; Sesia, M.: Testing for Outliers with Conformal p-Values. The Annals of Statistics 51(1), 149-178 (2023) https://doi.org/10.1214/22-AOS2244

[43] Ledoit, O.; Wolf, M.: A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. Journal of Multivariate Analysis 88(2), 365-411 (2004) https://doi.org/10.1016/S0047-259X(03)00096-4

[44] Liu, F.T.; Ting, K.M.; Zhou, Z.: Isolation Forest. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM), 413-422 (2008) https://doi.org/10.1109/ICDM.2008.17

[45] Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks. Proceedings of the 6th International Conference on Learning Representations (ICLR) (2018) https://arxiv.org/abs/1706.06083

[46] Athalye, A.; Carlini, N.; Wagner, D.: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. Proceedings of the 35th International Conference on Machine Learning (ICML) 80, 274-283 (2018) https://proceedings.mlr.press/v80/athalye18a.html