Physics-Assisted Explainable Anomaly Detection in Power Systems

Matthew Lau[1], Fahad Alsaeed[1], Kayla Thames[1], Nano Suresettakul[1], Saman Zonouz[1], Wenke Lee[1] and Athanasios P Meliopoulos[1]

[1]Georgia Institute of Technology

Abstract. Detection of cyber-attacks in power systems is crucial for rapid corrective actions like isolation, disinfection and asset restoration. For real-time deployment, detection methods must not only be accurate and computationally efficient, but also interpretable for further action. While physics models can reliably detect cyber-attacks, diagnosing where and how assets were attacked is computationally demanding. To supplement detection models, we propose *Physics-Assisted Statistics for Anomaly Localization (PASAL)*, a domain-informed data-driven method that directly identifies anomalous devices. PASAL leverages domain knowledge of the grid topology and incorporates correlation and variance statistics to model inter-sensor causal relationships. Consequently, PASAL offers inherent interpretability and computational efficiency. Our study demonstrates that PASAL swiftly localizes data integrity attacks with minimal false positives and has the potential to identify the type of attack.

1 Introduction

Physical systems, such as power systems, are now controlled and operated by cyber assets. Anomalies can arise not only from faults [23] but also from cyber-attacks. Such systems are huge, so anomalies must not only be detected, but detections must also be explainable for targeted incident response. Explainable anomaly detection can encompass localizing *where* the anomaly is (anomaly localization) and identifying *how* it came about (anomaly classification). In this paper, we mainly focus on anomaly localization. It is challenging to localize anomalies in large systems due to the complexity of modeling temporal and inter-sensor dependencies in high dimensions [15] and collecting representative training data [11].

We propose 3 metrics to evaluate a method for localizing anomalies: *performance*, *interpretability* and *computational efficiency*. Performance evaluates the ability to detect and localize attacks. Interpretability is a qualitative metric for users to understand and trust their system, the best being inherently interpretable models [12]. More understanding can provide more diagnostics, such as anomaly classification (e.g. the type of attack). Finally, to ensure quick response and mitigation of attacks, methods must be computationally efficient to run in real-time. Solutions are generally dichotomized into domain-based and data-driven methods. We summarize key comparisons in Table. 1.

Table 1: Comparison of anomaly localization methods on interpretability (int.), computational efficiency (eff.) and performance (perf.) from our experiments. Legend: ● for best, ○ for worst.

Method Type	Localization Method	Int.	Eff.	Perf.
Physics-Based	Dynamic State Estimation [16]	●	○	NA
	$D\chi^2$ [6]	●	◐	○
Data-Driven	Causal Discovery (RCD) [8]	●	○	◐
	Spectral [21]	◐	◐	○
	Masked Modelling [15]	○	●	NA
Hybrid	Sparse Group Lasso [10]	◐	○	NA
	kNN [18]	●	◐	◐
	PASAL (our method)	●	●	●

1.1 Domain-/Physics-based Anomaly Localization

Purely domain-based methods trade-off between computational efficiency and performance to detect anomalies. Traditional bad data detection has detection delays of up to 0.15 seconds [28], 5 times longer than protection latency requirements [2]. Direct current (DC) state estimation is much faster, but its liberal approximations for alternating current (AC) systems makes it vulnerable to data integrity attacks [14]. Meanwhile, Dynamic State Estimation (DSE) maintains a balance between modeling AC systems and being within the latency requirement to detect anomalies, running every cycle [16, 26].

After detecting the anomaly, state estimation methods like DSE perform hypothesis testing to localize compromised devices. Without prior knowledge, hypothesis tests require naive subset scanning, which has exponential complexity. To mitigate this, weighted least squares state estimation with decoupled χ^2 test ($D\chi^2$) [6] and Sparse Group Lasso [10] assume that one region is attacked to localize the attack. However, solving optimization problems like these in real-time can be inefficient and lacks convergence guarantees.

1.2 Data-driven Anomaly Localization

Exploration of anomaly localization methods lags behind that of anomaly detection in data-driven methods too. Typically, these methods trade-off performance, interpretability and efficiency. Some are computationally inefficient, such as looping over the training data for k-Nearest Neighbors (kNN) [18], performing causal discovery [8], performing real-time optimization (e.g. [9]) or require high-dimensionality and thus more computing [21]. Others lack the capacity to model dependencies across devices which produces more false positives [12]. For instance, spectral methods [21] require real-time optimization likened to solving polynomial roots, while root cause discovery (RCD) [8] involves numerous conditional independence tests. In contrast, efficient anomaly detection methods heavily rely on their interpretability for localization. Explainable artificial intelligence (XAI) tools like SHAP [20] designed to interpret outputs of complex models (e.g. neural networks) lack strong explainability power in detecting anomalies [12].

Conversely, inherently interpretable statistical models (e.g. linear-time subset scanning [4]) require unrealistically strong parametric assumptions or are inefficient (for non-parametric methods e.g. linear in amount of training data).

1.3 Hybrid Methods: Best of Both Worlds

By contextualizing data-driven methods with domain knowledge, hybrid models model anomalies of interest to increase interpretability and performance [24]. One approach is to model inter-sensor dependencies implicitly such as by using neural networks to test for non-linear Granger causality [15]. Without a causal model, Granger causality only asserts correlation and not causation. Coupled with an uninterpretable neural network, it is not trivial to remove spurious correlations (see [19]). One way is to construct causal models for root-cause analysis (RCA) [11]. However, RCA requires a directed acyclic graph while the power system topology is mostly undirected, so RCA cannot be applied directly. Another way is to embed the physics of a system in data-driven methods. Grid topology can be used to monitor changes in groups of the 3-phase current and voltage sensor measurements across time to detect and localize anomalies in the distributed energy resources (DERs) [27]. Similarity networks [3] is a more flexible framework so sensors that are not grouped can still be monitored with arbitrary sensor-pairwise similarity functions. If similarity functions can be chosen to reflect the physics of the system, this framework can mimick reality more closely.

In this paper, we evaluate the performance and computational efficiency of interpretable methods for unsupervised anomaly localization. We propose a novel hybrid approach, *Physics-Assisted Statistics for Anomaly Localization (PASAL)*, focusing on protecting power networks against data integrity attacks. Our contributions are as follows:

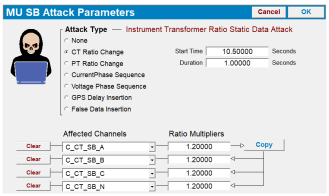
1. We use physical laws via **grid topology** to **directly model inter-sensor causal relationships**, where causality implies correlation.
2. We dichotomize realistic attacks into **phase- and amplitude-based interventions** on sensors, using correlation for the former and introduce *variance ratio* statistics for the latter.
3. Our experiments demonstrate that our hybrid model is not only **efficiently localizes** anomalies to specific sensors, but also remains **computationally efficient**. Moreover, it provides **inherent interpretability**, which facilitates further anomaly classification.

2 Anomaly Localization Problem

2.1 Data-driven Problem Formulation

Let the distribution of normal (non-anomalous) data be D_X, where X refers to the data. The data contain time-series measurements of currents and voltages, measured through the current transformers (CT) and potential transformers (PT) in the system. Let $\mathbf{x}_m$ be the d-dimensional vector that contains the data of the d sensors at sample m. Let an observation with a window of size T before sample m be denoted as $\mathbf{x}_{\leq m} = (\mathbf{x}_{m-T+1}, ..., \mathbf{x}_m) \in \mathbb{R}^{T \times d}$ i.e. a sequence of window size T that has data up to sample m.

Before deployment, we only have limited access to D_X through a collected training dataset that establishes baseline operation (e.g. all possible normal events not captured). During inference, the power system streams data and we want to know if there are anomalous sensors in $\mathbf{x}_{\leq m}$ and if so, which ones.



MU SB Attack Parameters

Attack Type — Instrument Transformer Ratio Static Data Attack

○ None
● CT Ratio Change
○ PT Ratio Change
○ CurrentPhase Sequence
○ Voltage Phase Sequence
○ GPS Delay Insertion
○ False Data Insertion

Start Time 10.50000 Seconds
Duration 1.00000 Seconds

Affected Channels	Ratio Multipliers
C_CT_SB_A	1.20000
C_CT_SB_B	1.20000
C_CT_SB_C	1.20000
C_CT_SB_N	1.20000

Figure 1: Attack User Interface.

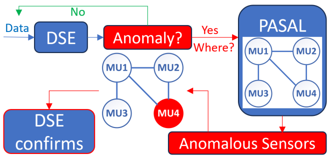


Figure 2: Overview of PASAL.

2.2 Realistic Data Integrity Threat Model

In the literature, most false data injection attacks against power systems are performed by manually altering a subset of meter measurements [13]. This overlooks how data is injected through cyber-attack vectors [1, 17], which is more constrained. We aim to capitalize on the constraints.

Instead of manually editing measurement data, we launch more realistic attacks with a cyber model via its cyber assets. We use high-fidelity three-phase models of the physical system with the cyber system. This cyber-physical co-model includes the instrumentations that link the physical system to the cyber system, and its performance is validated with state estimation. The cyber model has all possible cyber devices located in the power system facility (e.g. merging units (MUs), relays, meters, fault recorders). The existing literature uses Phasor Measurement unit (PMU) data for the measurements [7], which cannot capture the transient behavior in power systems and only represents the steady-state operation. Hence, we use Sampled Values (SV) data in this study, which are high-resolution samples of current and voltage waveforms that capture the transient conditions. Modeling transient conditions resembles the time-series formulation in Section 2.1 for us to use data-driven change-point methods.

We inject false data by GPS spoofing and editing software settings of MU devices (Fig. 1). MU devices collect and convert analog signals from the CT/PT into digital signals. When a cyber device such as a MU is compromised, the attacker cause harm by setting incorrect values for the CT/PT ratio. The displayed data scales by a multiple of the true physical quantity. Phase sequence change attacks change the input channels settings, resulting in a swapped pair of phase measurement channels or phase measurements that are identical to another. GPS spoofing attacks send falsified signals to the GPS clock to add a phase shift delay to the measurement. These false data can cause misoperation, especially in devices that rely on synchronized measurement (e.g. differential relay devices). From the nature of data manipulation, we dichotomize attacks into phase-based (phase sequence and GPS attacks) and amplitude-based (CT/PT ratio attacks).

Phase-based Attacks For a measurement at sample m denoted as $x_{real}(m)$, the attacked measurement $x_{att}(m)$ in a phase-based attack is defined as

$$x_{att}(m) = x_{real}(m - \gamma) \tag{1}$$

where γ is the delay in samples introduced to shift the phase of the measurement signal. A phase sequence attack will introduce a delay corresponding to a ±120 degree phase shift in a three-phase system, while a GPS attack has delays up to 360 degrees.

Amplitude-based Attacks In the amplitude-based attack, the attacked measurement $x_{att}(m)$ is represented as

$$x_{att}(m) = \beta \cdot x_{real}(m) \tag{2}$$

for scaling factor $\beta \in (0, \infty) \backslash \{1\}$ of the attack. In practice, β belongs to a finite set determined by the settings in the MU device.

3 Proposed Method: PASAL

We propose Physics-Assisted Statistics for Anomaly Localization (PASAL), which flags out sensors with violations in their inter-sensor causal relationships. We first outline how PASAL localizes anomalies.

Before deployment, we use physical laws via the grid topology to construct a graph of sensors (vertices) and inter-sensor relationships (edges). We denote this graph as our Markov network. Concurrently, a training dataset of the power system in normal operation is generated offline. The training dataset acts as a baseline and only has normal events, such as load changes, with no attacks. For each sensor, we monitor the changes in the correlations and variance ratio between its neighbors in the Markov network within a sliding window across time. For each sensor, we set the threshold on the sum of these change statistics based on the training data.

In deployment, after DSE detects an anomaly [16, 26], PASAL executes to flag out sensors whose statistics change beyond the baseline threshold (Fig. 2). Then, DSE performs hypothesis testing on PASAL's suspected sensors to confirm the attacked sensors/location.

3.1 Data Preprocessing

Data from power systems vary widely based on the kilo Voltage (kV) level of the part of the system. Also, power systems are designed such that for any loop in the system network, the net phase shift at certain equipment (e.g. delta-wye transformer) is zero. In a normal system, data can be scaled and abrupt phase shifts can be removed, resulting in data with a magnitude near 1.0 per unit (pu) and a small phase shift between the data. We refer to this as *normal Per-Unit (NPU) normalization*. To prepare the data for statistical analysis, we run *NPU normalization*. Appendix A shows an example. Section 3.3 motivates this preprocessing.

3.2 Causal Modeling with a Markov Network

We propose using a causal graph of inter-sensor relationships with the grid topology, which is an undirected graph that shows how each component in the electrical network are interconnected via conductors like wires. Inter-sensor relationships arise from the flow of electricity through these conductors, which are determined by Kirchhoff's Current Law (KCL), Kirchhoff's Voltage Law (KVL) and Ohm's Law. We provide an example in Appendix B. To model undirected inter-sensor relationships reflected by the topology, we relax the requirement for directed edges in [11] and use a Markov network instead. In a Markov network $\mathcal{G} = (V, E)$, where V represents the set of vertices and E represents the set of edges between these vertices, a vertex (v) is conditionally independent of all other vertices given its neighbors. Mathematically, for all $v \in V$, $v \perp V \backslash N_{\mathcal{G}}(v) \backslash \{v\} \mid N_{\mathcal{G}}(v)$, where $N_{\mathcal{G}}(v)$ is the set of neighbors of vertex v. Instead of probabilistic inference, we adapt the idea of conditional independence for causal analysis: given the information provided by the sensor's neighborhood in the Markov network (which is a *Markov boundary* of the sensor), any data from other sensors outside the boundary will not give additional information on the sensor's state.

Markov networks have 3 favorable properties. First, it factors a global (probability) model into local (probability) models. Attacks are seen as interventions on variables [8, 11], so we can perform direct localization with these local models due to its interpretability. Second, local probability models are based on conditionals $\mathbb{P}(X|N_{\mathcal{G}}(X))$ and not joint distributions $\mathbb{P}(X \cup N_{\mathcal{G}}(X)) = \mathbb{P}(X|N_{\mathcal{G}}(X))\mathbb{P}(N_{\mathcal{G}}(X))$, so we only model causal relations.[1] By not modeling marginals, our model is more robust to a limited access to the marginals in the collected training dataset. Third, a Markov network imposes a sparsity constraint for causal modeling, removing spurious correlations [22].

3.3 Anomaly Localization with Graph-Based Statistics

To localize anomalies, we perform hypothesis testing with two parallel alternative hypotheses (H_p and H_a) on each sensor:

$$H_p : \begin{cases} H_0 : \text{Normal.} \\ H_1 : \text{Phase attack.} \end{cases} \qquad H_a \quad : \begin{cases} H_0 : \text{Normal.} \\ H_1 : \text{Amplitude attack.} \end{cases}$$

Measurements are near sinusoidal, so they are characterized by their phase and amplitude. Thus, violations in these attributes could indicate anomalous behavior. Modeling attacks as different interventions on attacked sensors, we use the correlation and the ratio of variances (which we term *variance ratio*) of each sensor with every sensor in its neighborhood as the pairwise similarity metrics for phase and amplitude interventions respectively. These statistics are interpretable and are a direct quantitative measure of the strength of the causal relations (since causality implies correlation). We estimate the correlation (ρ) and variance ratio (VR) between sensors X and Y with a window of size w ending at sample m:

$$\rho_{m,w}(X, Y) = \frac{\text{Cov}_{m,w}(X, Y)}{\sigma_{m,w}(X) \cdot \sigma_{m,w}(Y)}$$

$$\text{Cov}_{m,w}(X, Y) = \frac{1}{w} \sum_{i=m-w+1}^{m} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{VR}_{m,w}(X, Y) := \frac{\sigma^2_{m,w}(Y)}{\sigma^2_{m,w}(X)}, \ \sigma^2_{m,w}(X) = \frac{1}{w} \sum_{i=m-w+1}^{m} (x_i - \bar{x})^2$$

3.3.1 Attack Modeling

To motivate the validity of using correlation and variance ratio, we first consider an ideal case. Let X and Y be a pair of adjacent sensors in a Markov network. They collect data following sine waves with no phase shift from each other and no noise. These assumptions (which are relaxed later) illustrate the validity of using the correlation and variance ratio as effective statistics. We analyze examples of events within a 2-cycle window with p samples per cycle during steady state operation (X, Y), as well as normal events (X', Y') and when X is attacked (X^a, Y) in the second cycle.

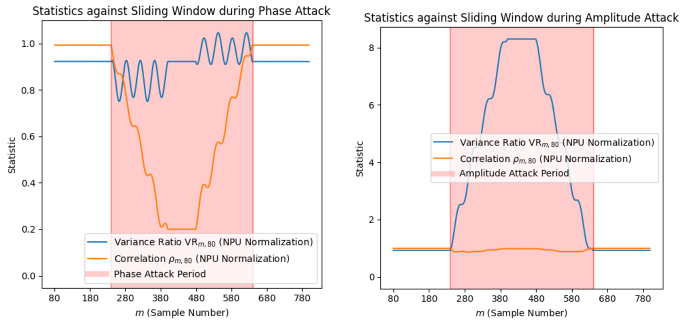
During steady state operation, without loss of generality, let the variance be $\sigma^2_{2p,2p}(X) = \frac{1}{2p}\sum_{i=1}^{2p}(x_i - \bar{x})^2 = \frac{1}{2p}\sum_{i=1}^{2p} x_i^2 := 1 =: \sigma^2_{2p,2p}(Y)$ for mean zero waves. Then,

$$\rho_{2p,2p}(X, Y) = \frac{1}{2p}\sum_{i=1}^{2p}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{p}\sum_{i=1}^{p} x_i y_i = 1,$$

$$\text{VR}_{2p,2p}(X, Y) = \frac{\frac{1}{2p}\sum_{i=1}^{2p}(y_i - \bar{y})^2}{\frac{1}{2p}\sum_{i=1}^{2p}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{p} y_i^2}{\sum_{i=1}^{p} x_i^2} .$$

During normal events (e.g. load change), the phase/amplitude of a sensor and its neighbors change equally. To illustrate, let the phase

[1] There are many ways to model the probability distributions. In our paper, we model $\mathbb{P}(X|N_{\mathcal{G}}(X))$ as the probability distribution of sensor X at sample $m+1$ given data from its neighbors in the stream $\mathbf{x}_{\leq m}$. We model a time-series to account for possible time lags in inter-sensor relationships.



(a) Phase-Based Attack

(b) Amplitude-Based Attack

Figure 3: Visualizations of correlation and variance ratio statistics across time for data with NPU normalization. Statistics drastically change when the attack is launched. We set $w = 80$.

shift by angle ϕ (equivalent to shifting by γ samples) and the amplitude scale by $\beta > 0$ for $\beta \neq 1$. Then, these statistics do not change:

$$\rho_{2p,2p}(X', Y') = \frac{1}{2p}\left[\sum_{i=1}^{p} x_i y_i + \sum_{j=1}^{p} x_{j+\gamma} y_{j+\gamma}\right] = \rho_{2p,2p}(X, Y)$$

$$\mathrm{VR}_{2p,2p}(X', Y') = \frac{\sum_{i=1}^{p} y_i^2 + \sum_{i=1}^{p} (\beta y_i)^2}{\sum_{i=1}^{p} x_i^2 + \sum_{i=1}^{p} (\beta x_i)^2}$$
$$= \frac{(1+\beta^2)\sum_{i=1}^{p} y_i^2}{(1+\beta^2)\sum_{i=1}^{p} x_i^2} = \mathrm{VR}_{2p,2p}(X, Y).$$

During an attack, the attacked sensor is intervened on. Depending on the attack, the phase shifts by ϕ or amplitude scales by β for the attacked sensor but not its neighbors, so the statistics change:

$$\rho_{2p,2p}(X^a, Y) = \frac{1}{2p}\left[\sum_{i=1}^{p} x_i y_i + \sum_{j=1}^{p} x_{j+\gamma} y_j\right] \overset{p\to\infty}{=} \frac{1}{2}\left(1 + \cos(\phi)\right)$$

because the correlation of ϕ out-of-phase sinusodial waves is $\cos(\phi)$ [5], and

$$\mathrm{VR}_{2p,2p}(X^a, Y) = \frac{2\sum_{i=1}^{p} y_i^2}{\sum_{i=1}^{p} x_i^2 + \sum_{i=1}^{p} (\beta x_i)^2}$$
$$= \frac{2}{1+\beta^2}\mathrm{VR}_{2p,2p}(X, Y).$$

In a real system, the data are not ideal as above. First, there are phase shifts ϕ between the measurements due to the physics (e.g. voltage drop due to Ohm's law, load conditions, etc.), so the correlation will be $\cos\phi = 1 - \epsilon$ for small $\epsilon > 0$. Second, some transformers introduce discrete 30-degree phase shifts between primary and secondary voltages (Section 3.1), so we use *NPU normalization* to remove these discrete shifts.

For a realistic demonstration, we use the primary and secondary voltage measurements from a delta-wye connected transformer to visualize the statistics during an attack. Fig. 3 shows the correlation and variance ratio during phase and amplitude attacks of NPU data. We make 2 observations. First, these statistics with NPU data are close to the ideal case during normal operation. Second, our interpretable statistics (correlation *and* variance ratio) produce anomaly signals.

Data variability could also weaken some of the assumptions. To mitigate this, we use 2 approaches. First, rather than monitor the actual values of correlation and variance ratio, we monitor the changes across time. Second, we refer to a training dataset as a baseline for normal operation.

3.3.2 *Measuring Change across Time*

Differencing is useful in filtering trends out to push the time-series towards stationarity. We measure the average change of correlation and variance ratio across the neighborhood and across s pairs of windows for all sensors

$$\Delta_{\rho_{m,w}}(X) := \frac{1}{s|N_{\mathcal{G}}(X)|} \sum_{i=m-s+1}^{m} \sum_{Y \in N_{\mathcal{G}}(X)} |\rho_{i,w}(X,Y) - \rho_{i-1,w}(X,Y)| \tag{3}$$

$$\Delta_{\mathrm{VR}_{m,w}}(X) := \frac{1}{s|N_{\mathcal{G}}(X)|} \sum_{i=m-s+1}^{m} \sum_{Y \in N_{\mathcal{G}}(X)} |\mathrm{VR}_{i,w}(X,Y) - \mathrm{VR}_{i-1,w}(X,Y)| \tag{4}$$

and refer to these second-order statistics as change statistics. Aggregation of edge statistics means that a representative training dataset only has $|V|$ dimensions rather than $|E| = \mathcal{O}(|V|^2)$ dimensions, which we empirically observe to be more robust to limited training data. Meanwhile, the Markov network imposes a sparsity constraint to filter which elements in the correlation matrix and which variance ratios we monitor, removing spurious correlations. Moreover, using a sliding window of size $w + s$ allows us to implicitly model the time-series, because the statistics themselves do not explicitly model it.

3.3.3 *Obtaining Baselines*

Baselines of the change statistics are set under normal operation. Training data models the normal variability of the system (e.g. load changing), enabling us to estimate an upper bound on these change statistics. In deployment, we diagnose a sensor as anomalous if its change statistics go beyond the baseline q-quantile. To reduce false positives, we mandate that an anomaly needs to persist for k consecutive samples before being flagged out. Mathematically, the probability of large change across k samples (i.e. false positive) shrinks exponentially ($(1-q)^k$) given the null hypothesis of independent and identically distributed noise in baseline operation.

After PASAL localizes a set of anomalous sensors, DSE runs hypothesis testing using this set (and possibly subsets via binary search) to confirm the localization. Here, the efficiency of DSE depends on PASAL's ability to localize anomalies to a small and correct set of sensors.

4 Experiments

4.1 *Metrics for Performance and Computational Efficiency*

We evaluate methods on their localization ability, false positives and computational efficiency. For localization, we measure the localization delay (LD), the mean samples before an attack is localized. We compare this to the detection delay (DD), the mean number of samples before an attack is detected. We also measure no localizations (NL), the number of attacked sensors not localized. Since the attacks target inter-sensor relationships, we consider an attack localization to be effective when the attack is correctly localized to the 1-hop neighborhood of the attacked sensor, as defined by the Markov network.[2] False positives are measured in terms of samples and sensors. For samples, we use the mean false positive

[2] If desired, the localized sensors can be prioritized by ranking them by their degree in the induced subgraph of anomalous sensors.

rate across all sensors (FPR), where a false positive refers to an incorrect localization. For sensors, we use the false localization rate (FLR), the proportion of incorrectly localized sensors to sensors that count as effective localizations, averaged across all attacks. A higher FLR denotes a potentially more expansive search by DSE to isolate the attacked sensors. For computational efficiency, we measure the average time taken for a method to run for 1 sample (window).

4.2 Dataset for Training

Our example test system is a 115/13.8 kV substation (Fig. 4b) which is part of a large electric network (Fig. 4a) and one of the distribution circuits connected to that substation (Distribution Feeder Group 1, Fig. 4c). We generate data using the system of Fig. 4 as per Section 2.2. The time-series (sampled values) data of the 3-phase CT and PT measurements from 20 MUs (the black squares in Fig. 4b) spread over the system are collected. 17 MUs include both 3-phase current and voltage measurements, while 3 MUs only have 3-phase current measurements, resulting in a total of 111 collected measurements.

The training dataset consists of 96k samples for each measurement collected over a 20-second simulation of the system (with 80 samples per cycle). To capture the normal variability of the system in the training dataset, we conduct 32 load change events by varying loads between 800 kilo Watts (kW) and 9000 kW for real power and 70 kilo Volt-Amps Reactive (kVar) and 3000 kVar for reactive power. Changing a load at a bus close to a sensor results in changes in the voltage and current measurements of that sensor and its neighboring sensors. Hence, all possible locations for a load change event in the system are performed to capture the inter-sensor relationships.

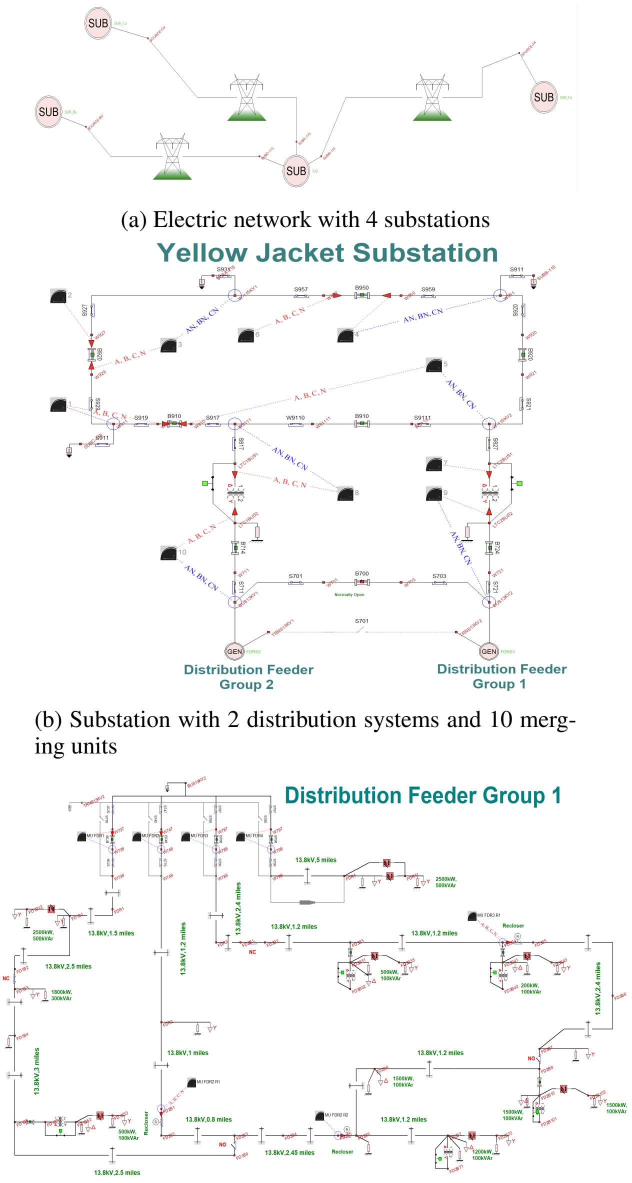
4.3 Experimental Details

During test time, we launch data integrity attacks on the power system (see Section 2.2). The testing dataset is used to evaluate the performance of the proposed method, which involves 9 different cyber attack events and 19 load changing events distinct from the training dataset (simulating covariate shift). The testing dataset comprises 96k samples for each measurement from 20 seconds of simulation (with 80 samples per cycle). To illustrate, Fig. 5 displays a subset of load change events and attacks in first 10.5 seconds of our test case. The full timeline is in Appendix D Fig. 11, along with attack parameters.

We use a window size of $w = 320$ (4 cycles) for our statistic and a rolling sum window of $s = 1$, so our overall window size is $T = w + s = 321$. Since our training data does not have any attacks, we use the $q = 99\%$ percentile of change statistics during training as our threshold. We set that violations of thresholds must persist for at least $k = 5$ timesteps before PASAL flags a sensor out as anomalous.

5 Results and Discussion

As expected, DSE uses domain knowledge and detects all attacks quickly within the cycle (80 samples, 16.67ms). PASAL is also efficient, taking 0.0957ms to run per sample on average on four 2.7GHz Intel$^{®}$ Xeon E5 2686 v4 processors. With no dependency on real-time optimization or the number of attacked sensors, PASAL has a computational complexity of $\mathcal{O}(|E|ws) = \mathcal{O}(|V|Dws)$ where $D := \max_{v \in V} \deg_{\mathcal{G}}(v) \ll |V|$, scaling well (linearly) with the number of sensors. As an aside, dependency on w and s in time complexity can also be traded for memory/space complexity. We report results across different attacks in Table. 2.



(a) Electric network with 4 substations

(b) Substation with 2 distribution systems and 10 merging units

(c) Distribution system with 7 merging units

Figure 4: The power system used in this study.

PASAL localizes all attacks early within an LD of 10 samples, well within our goal of latency less than 2 cycles (33ms or 160 samples), with low FPR. Fig. 6 has sample visualizations which show a sharp increase in change statistics at the start of phase- and amplitude-based attacks on the respective attacked sensors. Measurements are labeled I for currents through phases a, b, c, or neutral (n) denoted by the subscript, and V for voltages between a specific phase and n denoted by the subscript. We observe some false positives during normal events, but these are filtered out by DSE. Note that the spike in variance ratio change statistics in Fig. 6b does not imply an amplitude-based attack, but our statistics are designed to ensure the converse for effective localization. We further validate that our *pairwise* correlation and variance ratio statistics give good localization signals too, changing only for sensors close to the attack. We provide 2 examples in Fig. 7.

The low FPR with low FLR for CT/PT ratio and GPS spoofing attacks indicates that only a few sensors have high FPR while most are close to 0% FPR. This suggests that these few sensors may have

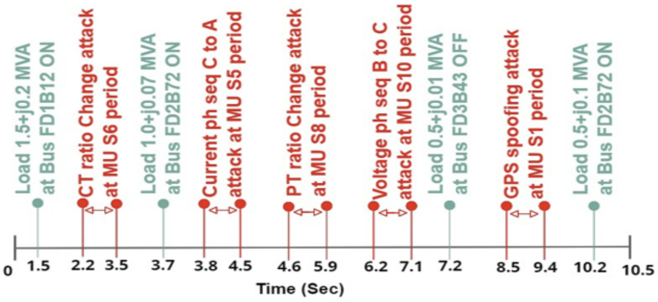
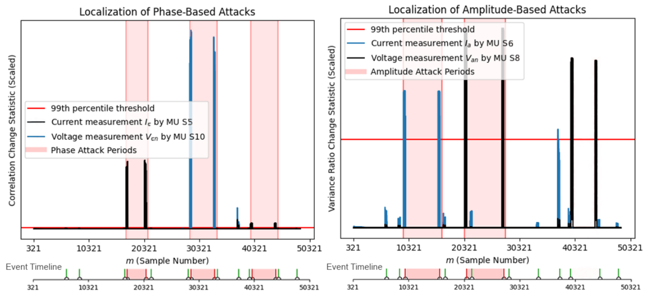

Load 1.5+j0.2 MVA at Bus FD1B12 ON
CT ratio Change attack at MU S6 period
Load 1.0+j0.07 MVA at Bus FD2B72 ON
Current ph seq C to A attack at MU S5 period
PT ratio Change attack at MU S8 period
Voltage ph seq B to C attack at MU S10 period
Load 0.5+j0.01 MVA at Bus FD3B43 OFF
GPS spoofing attack at MU S1 period
Load 0.5+j0.1 MVA at Bus FD2B72 ON
0
1.5
2.2
3.5
3.7
3.8
4.5
4.6
5.9
6.2
7.1
7.2
8.5
9.4
10.2
10.5
Time (Sec)

Figure 5: Rough timeline of the first ∼10s of a test case.



(a) Correlation: phase-shift attacks.

(b) Variance ratio: amplitude-based attacks.

Figure 6: Sample visualizations of correlation and variance ratio change statistics during attacks. Event timeline shows green bars for normal events and red highlighted portions for relevant attack periods.

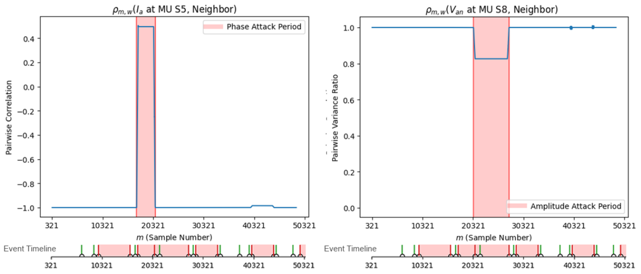
inter-sensor relationships not modeled by our Markov network. Despite only using grid topology as domain knowledge, our Markov network captures most inter-sensor relationships, leading to swift localization for all attacks. We observe that phase sequence change attacks have similar FPRs but higher FLR, suggesting that there are more sensors which violate the threshold, but they do so for a shorter period of time. This suggests that a higher persistency parameter k can reduce the FPR. Our choice of k preserves a good trade-off between a low LD and a low FPR.

5.1 *Comparison to Other Localization Methods*

To benchmark PASAL against the literature, we compare other state-of-the-art anomaly localization methods which are inherently interpretable: Dχ^2 [6] (domain-based), spectral statistics on the correlation matrix [21] and RCD [8] (data-driven), and kNN [18] (hybrid). Dχ^2 and kNN account for the grid topology like our method, while the data-driven methods do not. For RCD, we provide knowledge on the number of attacked sensors, which is unknown during deployment but required for the method. We report results in Table. 3.

We observe that all methods have LDs below 160 samples (2 cycles), which is within the latency requirement. We also observe that PASAL significantly outperforms all methods on FLR and time taken to run, so we discuss other metrics.

The spectral method suffers from 99.6% FPR, suggesting that it does not produce useful signals during attacks. Similarly, Dχ^2 seems to have noisy signals with a high FPR, FLR and NL. Aside from these 2 methods, PASAL achieves the best localization (LD, NL, DD). For false positives, RCD achieves a slightly lower FPR than PASAL. However, its 4-second runtime does not meet the latency requirement and has 3 NL despite knowing the number of attacked sensors. Apart from RCD, PASAL achieves the best FPR. kNN runs more efficiently than RCD, but is too reliant on training data. Its lack of robustness to covariate shift compared to PASAL explains why PASAL dominates it in every metric. Holistically, PASAL is the only method that is fast enough for real-time deployment and produces useful signals for swift localizations and minimal false positives.



(a) Correlation: phase-shift attack.

(b) Variance ratio: amplitude-based attack.

Figure 7: Sample visualizations of pairwise correlation and variance ratio statistics during an example attack. Statistics change only during attacks in the sensor's neighborhood.

Table 2: PASAL Results for different cyber-attacks: DD, LD (#samples); NL (#sensors).

Cyber-attack	DD	LD	NL	FPR	FLR	Time/ms
CT ratio	4.0	8.5	0	2.68%	0.39	0.0957 for all
PT ratio	4.0	4.0	0	3.24%	0.81	
GPS spoofing	4.0	6.3	0	5.10%	0.35	
I-Phase swap	4.0	7.5	0	3.42%	6.50	
V-Phase swap	4.0	10.0	0	2.78%	5.33	

5.2 *Ablation: Importance of Domain Knowledge*

To demonstrate the utility of domain knowledge, we test the performance of a variant, PASALv, and report results in Table. 4. Instead of using the grid topology, PASALv uses the complete graph $K_{|V|}$ as its "Markov network". Both localize all attacks within the latency requirement, but the high FPR and FLR for PASALv suggests that domain knowledge to model inter-sensor relations is crucial in removing spurious correlations. Moreover, PASALv is slower because of the quadratic complexity on sensors $|E| = \binom{|V|}{2}$ because the complete graph is used.

5.3 *Flexibility of PASAL: Topology Changes*

An advantage of PASAL over machine learning (e.g. neural networks) is the flexibility of analyzing a variable number of devices due to its modularity in monitoring conditionals of each sensor's neighborhood. These conditionals are stable, only changing under specific and predictable circumstances such as topology changes.

To account for topology changes, there are two possible approaches. The first is to train PASAL with different topologies through different Markov networks. During deployment, anomalies are localized based on the topology determined by DSE. This procedure is like training a mixture of experts model, except that domain knowledge informs us which expert to use as per the topology. The second approach is to not to adapt and accept these topological changes as noise, using thresholds from the base topology.

To test these 2 approaches, new topologies (Topo A and B) are introduced by opening/closing some of the reclosers and switches in the distribution system in Fig. 4c. We report results in Table. 5. Localization is rapid in both approaches. Since the topological changes are minimal (usually a few edges), most sensors can still be effectively monitored without retraining. Nevertheless, as predicted by the No

Table 3: Results comparing other methods: DD, LD (#samples); NL (#sensors). Best: **in bold**; Worst: underlined.

Method	DD	LD	NL	FPR	FLR	Time/ms
Dχ^2 [6]	4.8	4.8	3	22.73%	16.6	31.940
Spectral [21]	**4.0**	**4.0**	**0**	99.61%	22.0	14.548
RCD [8]	NA	14.0	3	**1.37%**	10.2	4109.1
kNN [18]	NA	14.6	5	7.73%	8.6	15.853
PASAL(ours)	**4.0**	6.5	**0**	3.43%	**2.4**	**0.0957**

Table 4: Results of PASAL and PASALv (no domain knowledge): DD, LD (#samples); NL (#sensors).

Method	DD	LD	NL	FPR	FLR	Time/ms
PASAL(ours)	**4.0**	6.5	**0**	**3.43%**	**2.4**	**0.0957**
PASALv	**4.0**	**4.9**	**0**	57.74%	16.1	4.5490

Free Lunch theorems [25], we notice a slight increase in FPR and FLR without retraining. PASAL trained on the relevant topologies suppresses false positives while swiftly localizing anomalies.

5.4 *Pursuing Interpretability: Anomaly Classification*

After localization, pairwise correlation and variance ratio statistics can help classify the type of attack (e.g. Fig. 7). For amplitude attacks, an attack on a CT/PT device corresponds to a CT/PT ratio attack respectively. For phase-based attacks, changes in correlation signals in both current and voltage measurements simultaneously suggest that the attack is probably a GPS attack, because one GPS signal is usually used to synchronize measurements from multiple CT and PT devices. Otherwise, the attack is probably a phase sequence attack, especially considering the challenge of simultaneously attacking multiple sensors in a real system.

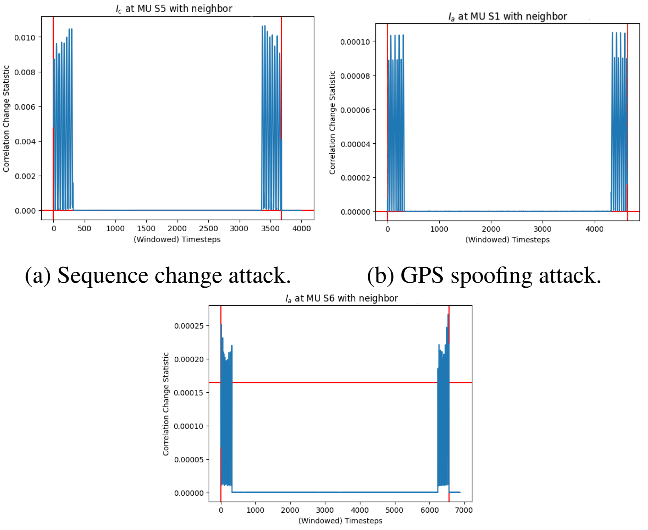
Signals from PASAL also reveal that PASAL can identify all types of attacks. We can use correlation statistics to help classify the attack type, as seen in Fig. 8.[3] For all attacks, we see a pattern where the correlation change statistics ('signals') will peak and drop, but a distinct trend exists for each attack. For the subsequent discussions, we will ignore scale of change and focus on patterns in the signals, because patterns are less sensor-specific.

Sequence change attacks display a general upward trend in the signals as the attack persists (Fig. 8a). Meanwhile, GPS spoofing attacks have a cluster of roughly equally sized signal peaks (Fig. 8b). These suggest that the physical effect of both attacks are different, even though they are both phase-based attacks – sequence change attacks seem to result in increasing change. One possibility is that GPS spoofing attacks use the same sensor's data, which is more stable to noise with respect to its own measurement than with respect to other sensor's data, which is what sequence change attacks use. Meanwhile, CT/PT ratio attacks display a decrease then increase in their signal, with a final peak lower than its initial peak (Fig. 8c). Moreover, these peaks seem more frequent than the phase-based attacks. Note that these percentage violation of amplitude-based attacks are not as much as those for phase-based attacks, which sheds some light on how variance ratio is more reliable for localizing the attacks.

These signals can help diagnose how the sensor is being attacked. If desired, this classification can be automated, such as using machine learning on Fourier-transformed signals. Moreover, these consistent signals can be used to investigate potential false positives or false negatives due to the design choice of setting the threshold.

Table 5: Results for PASAL on changed topology (topo). We compare our proposed retrained PASAL with no retraining: DD, LD (#samples); NL (#sensors).

Topo	Method	DD	LD	NL	FPR	FLR
A	Retrain	5.4	6.4	**0**	**1.29%**	**2.4**
	No Retrain	**4.0**	**6.2**	**0**	3.79%	2.8
B	Retrain	5.9	**5.8**	**0**	**1.59%**	**2.1**
	No Retrain	**4.0**	**5.8**	**0**	4.01%	2.8



(a) Sequence change attack.

(b) GPS spoofing attack.

(c) CT/PT ratio attack.

Figure 8: Sample visualizations of correlation change statistics during attacks with distinct patterns. Red vertical lines denote the start and end of the attack, while the red horizontal line is the threshold.

5.5 *Evading PASAL requires more resources*

PASAL monitors the *effect* of attacks via the change in causal relationships. To evade PASAL, attacks need to maintain small change statistics over time, requiring the attacker to persist for a longer time. However, operators will be alerted by DSE (or any intrusion detection system) early and will have more time for incident response to compromised devices. By trading-off stealth from PASAL localization, the attacker gives the defender more time for diagnosis before the attack does significant damage to the system.

Another way to evade PASAL is to preserve all conditionals but edit the marginals of sensors. This would require an attack on all sensors, which is costly and would be flagged out by DSE. Controller-based attacks could be a cheaper alternative, and future work can extend PASAL to localize controller-based attacks.

6 Conclusion

We present PASAL, an explainable anomaly detector for realistic data integrity cyber-attacks. Integrating domain knowledge through the grid topology, PASAL directly localizes attacked sensors by measuring the change in causal relationships with other sensors in its neighborhood. This interpretability makes anomaly localization computationally efficient, suppresses false positives by removing spurious correlations and has potential for anomaly classification. PASAL can be generalized to other domains by defining the corresponding Markov network and inter-variable relationships.

[3] Since PASAL monitors changes across time, these signals will be present for the duration of the window at the start (and the end).

Acknowledgements

This work has been supported by the ARPAe award DE-AR0001031, the SETO award DE-EE0009020, ACTION (NSF grant number IIS-2229876) and DOE (DE-CR0000018). Their support is greatly appreciated.

Supplementary Materials

The accompanying appendix and code can be found in https://github.com/mattlaued/Physics-Assisted-Statistics-for-Anomaly-Localization.

References

[1] F. Almutairy, L. Scekic, R. Elmoudi, and S. Wshah. Accurate detection of false data injection attacks in renewable power systems using deep learning. *IEEE Access*, 9:135774–135789, 2021. doi: 10.1109/ACCESS.2021.3117230.

[2] P. M. Anderson, C. Henville, R. Rifaat, B. Johnson, and S. Meliopoulos. *Protective Device Characteristics*, pages 47–107. 2022. doi: 10.1002/9781119513100.ch3.

[3] S. Cao and Y. Xie. Dynamic change-point detection using similarity networks. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 117–121, 2016. doi: 10.1109/ACSSC.2016.7869006.

[4] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1166–1175, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623619. URL https://doi.org/10.1145/2623330.2623619.

[5] J. D. Cook. Correlation of two sine waves. https://www.johndcook.com/blog/2016/03/06/correlating-two-sine-waves. Accessed: 2024-04-28.

[6] M. Gol and A. Abur. A modified chi-squares test for improved bad data detection. *IEEE Eindhoven PowerTech*, pages 1–5, 2015.

[7] A. A. Habib, M. K. Hasan, A. Alkhayyat, S. Islam, R. Sharma, and L. M. Alkwai. False data injection attack in smart grid cyber physical system: Issues, challenges, and future direction. *Computers and Electrical Engineering*, 107:108638, 2023. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2023.108638. URL https://www.sciencedirect.com/science/article/pii/S0045790623000630.

[8] A. Ikram, S. Chakraborty, S. Mitra, S. Saini, S. Bagchi, and M. Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31158–31170. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c9fcd02e6445c7dfbad6986abee53d0d-Paper-Conference.pdf.

[9] S. Kulinski and D. I. Inouye. Towards explaining distribution shifts. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17931–17952. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kulinski23a.html.

[10] D. Li, N. Gebraeel, K. Paynabar, and A. P. S. Meliopoulos. An online approach to covert attack detection and identification in power systems. *IEEE Transactions on Power Systems*, 38(1):267–277, 2023. doi: 10.1109/TPWRS.2022.3167024.

[11] M. Li, Z. Li, K. Yin, X. Nie, W. Zhang, K. Sui, and D. Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, aug 2022. doi: 10.1145/3534678.3539041. URL https://doi.org/10.1145%2F3534678.3539041.

[12] Z. Li, Y. Zhu, and M. van Leeuwen. A survey on explainable anomaly detection. 2022. doi: 10.48550/ARXIV.2210.06959. URL https://arxiv.org/abs/2210.06959.

[13] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong. A review of false data injection attacks against modern power systems. *IEEE Transactions on Smart Grid*, 8(4):1630–1638, 2017. doi: 10.1109/TSG.2015.2495133.

[14] Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Trans. Inf. Syst. Secur.*, 14 (1), jun 2011. ISSN 1094-9224. doi: 10.1145/1952982.1952995. URL https://doi.org/10.1145/1952982.1952995.

[15] P. Lymperopoulos, Y. Li, and L. Liu. Exploiting variable correlation with masked modeling for anomaly detection in time series. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*, 2022. URL https://openreview.net/forum?id=TCJuzs585W.

[16] A. P. S. Meliopoulos, G. Cokkinides, R. Fan, and L. Sun. Data attack detection and command authentication via cyber-physical comodeling. *IEEE Design & Test*, 34(4):34–43, 2017. doi: 10.1109/MDAT.2017.2682233.

[17] M. Ozay, I. Esnaola, F. T. Yarman Vural, S. R. Kulkarni, and H. V. Poor. Machine learning methods for attack detection in the smart grid. *IEEE Transactions on Neural Networks and Learning Systems*, 27(8):1773–1786, 2016. doi: 10.1109/TNNLS.2015.2404803.

[18] V. Saligrama and M. Zhao. Local anomaly detection. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 969–983, La Palma, Canary Islands, 4 2012. PMLR. URL https://proceedings.mlr.press/v22/saligrama12.html.

[19] B. Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference*, pages 765–804. ACM, feb 2022. doi: 10.1145/3501714.3501755. URL https://doi.org/10.1145%2F3501714.3501755.

[20] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza. Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data. *Applied Sciences*, 11(16), 2021. ISSN 2076-3417. doi: 10.3390/app11167376. URL https://www.mdpi.com/2076-3417/11/16/7376.

[21] X. Shi, R. Qiu, X. He, Z. Ling, H. Yang, and L. Chu. Early anomaly detection and localisation in distribution network: a data-driven approach. *IET Generation, Transmission & Distribution*, 14(18):3814–3825, July 2020. doi: 10.1049/iet-gtd.2019.1790. URL https://doi.org/10.1049/iet-gtd.2019.1790.

[22] S. Strelnikoff, A. Jammalamadaka, and T.-C. Lu. Causanom: Anomaly detection with flexible causal graphs. *The International FLAIRS Conference Proceedings*, 36(1), May 2023. doi: 10.32473/flairs.36.133298. URL https://journals.flvc.org/FLAIRS/article/view/133298.

[23] T. Takagi, Y. . Yamakoshi, M. Yamaura, R. Kondow, and T. Matsushima. Development of a new type fault locator using the one-terminal voltage and current data. *IEEE Transactions on Power Apparatus and Systems*, PAS-101(8):2892–2898, 1982. doi: 10.1109/TPAS.1982.317615.

[24] A. Vafaei Sadr, B. A. Bassett, and M. Kunz. A flexible framework for anomaly detection via dimensionality reduction. *Neural Computing and Applications*, 3 2021. ISSN 1433-3058. doi: 10.1007/s00521-021-05839-5. URL https://doi.org/10.1007/s00521-021-05839-5.

[25] D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

[26] J. Xie and A. S. Meliopoulos. Sensitive detection of gps spoofing attack in phasor measurement units via quasi-dynamic state estimation. *Computer*, 53(5):63–72, 2020. doi: 10.1109/MC.2020.2976943.

[27] J. Yu, H. Cheng, J. Zhang, Q. Li, S. Wu, W. Zhong, J. Ye, W. Song, and P. Ma. Congo²: Scalable online anomaly detection and localization in power electronics networks. *IEEE Internet of Things Journal*, 9(15):13862–13875, 2022. doi: 10.1109/JIOT.2022.3143123.

[28] R. Zhu, C.-C. Liu, J. Hong, and J. Wang. Intrusion detection against mms-based measurement attacks at digital substations. *IEEE Access*, 9:1240–1249, 2021. doi: 10.1109/ACCESS.2020.3047341.

A Data Processing

Figure 9 illustrates a basic electrical network used as an example to demonstrate the process of *NPU normalization*. As shown in Figure 9, the transformers divide the system into 4 voltages zones, the *NPU normalization* procedure starts with voltages and currents being divided by the corresponding base, then the phase is shifted by the indicated amount to make the data normal.

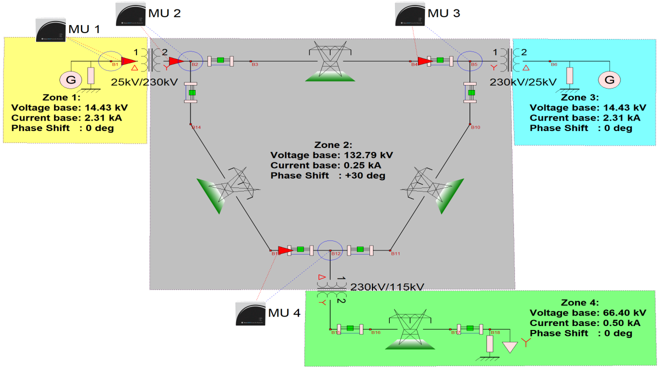


Figure 9: An example of a basic electrical network.

B Creating the Markov Network

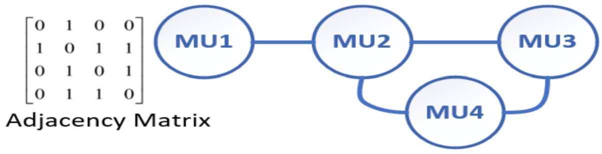


Figure 10: Markov network (graph and adjacency matrix form).

In grid topology modeling, given the topology information a Markov network is constructed using Kirchoff's current, voltage laws, and Ohm's law, which depict current-current, voltage-voltage, and current-voltage relations, respectively. To illustrate, we use a circuit example in Figure 9, where four MU devices gather both current and voltage measurements. The corresponding Markov network is in Figure 10. Edges between MU1 and MU2 indicate transformer primary and secondary relationships, with changes in MU1 affecting MU2. Edges between MU2, MU3, and MU4 show interrelations in measurements, and due to Kirchoff's laws, a change in one affects the others. In contrast, given MU2's measurements, MU1's measurements will not provide more information about MU3 and MU4 (and vice versa), so there are no edges between MU1 and MU4, and MU1 and MU3.

The same can be applied to radial networks: downstream alterations influence upstream readings and, by Ohm's Law, when line current alters, its voltage varies inversely.

C Graph-Based Statistics

The other two cases not mentioned in the main paper is how correlation and variance ratio changes under normal amplitude- and phase-based events respectively. Under our ideal assumptions, we also note that the correlation under normal amplitude-based events are different while the variance ratio under normal phase-based events remains the same.

C.1 Potential False Positives Mitigated

Since DSE first flags out anomalies before PASAL is run, a false positive during localization will occur when an attack is launched just when a normal amplitude-based event occurs. When this happens, the attack and the normal amplitude-based event will be flagged out. Although it is undesirable that the normal amplitude-based event is flagged out, the impact of this false positive is not big, because (1) only a constant number of such normal events are flagged out (and can be checked with physical equations: binary search with DSE on the proposed anomalous devices to confirm the identity of the truly anomalous device(s)) and (2) PASAL still localizes the true attack location. Hence, this attack is not particularly successful at incurring more cost to the defender.

C.2 Proofs

We show that correlation under normal amplitude-based events are different while the variance ratio under normal phase-based events remains the same. For normal amplitude-based events, we see the correlation statistics change:

$$\begin{aligned}\rho_{2p,2p}(X',Y') &= \frac{1}{2p}\left[\sum_{i=1}^{p} x_i y_i + (Ax_i)(Ay_i)\right] \\ &= (\beta^2+1)\rho_{2p,2p}(X,Y).\end{aligned}$$

On the other hand, for phase-shifts, the variance ratio statistics do not change:

$$\begin{aligned}\mathrm{VR}_{2p,2p}(X',Y') &= \frac{\sum_{i=1}^{p} y_i^2 + \sum_{i=1}^{p} y_{i+\gamma}^2}{\sum_{i=1}^{p} x_i^2 + \sum_{i=1}^{p} x_{i++\gamma}^2} \\ &= \frac{2\sum_{i=1}^{p} y_i^2}{2\sum_{i=1}^{p} x_i^2} = \mathrm{VR}_{2p,2p}(X,Y).\end{aligned}$$

D Attack Details

The full attack timeline is shown in Fig. 11. For the CT and PT ratio attacks, β values were set at 1.2 and 1.1. In the GPS attack, delays of 0.9 and 0.5 msec were implemented.

E Details of Other Methods

E.1 Data-driven methods

For the data-driven methods we compared to in Section 5 that required more tuning (namely, the spectral method and RCD), we explain how we used them for our study for reproducibility.

The spectral / linear eigenvalue method [21] This is the closest method to PASAL, but does not have domain knowledge and runs online (i.e. without training data). It monitors how the correlation matrix is changing by observing that

$$\sum_{j=1}^{d}\left(\frac{d\lambda_{\boldsymbol{\Sigma},k}}{d\epsilon_{ij}}\right)^2 = \left(v_{\boldsymbol{\Sigma},k}^{(i)}\right)^2 \tag{5}$$

where $\lambda_{\boldsymbol{\Sigma},k}$ is the k^{th} eigenvalue of the correlation matrix $\boldsymbol{\Sigma}$ and $v_{\boldsymbol{\Sigma},k}^{(i)}$ is the i^{th} entry of the k^{th} eigenvalue of correlation matrix $\boldsymbol{\Sigma}$. The anomaly score for the i^{th} sensor is

$$\eta_i := \frac{\sum_{\lambda_{\boldsymbol{\Sigma},k}\in\{\lambda>b\}} \lambda_{\boldsymbol{\Sigma},k}\left(v_{\boldsymbol{\Sigma},k}^{(i)}\right)^2}{\sum \lambda_{\boldsymbol{\Sigma},k}} \tag{6}$$

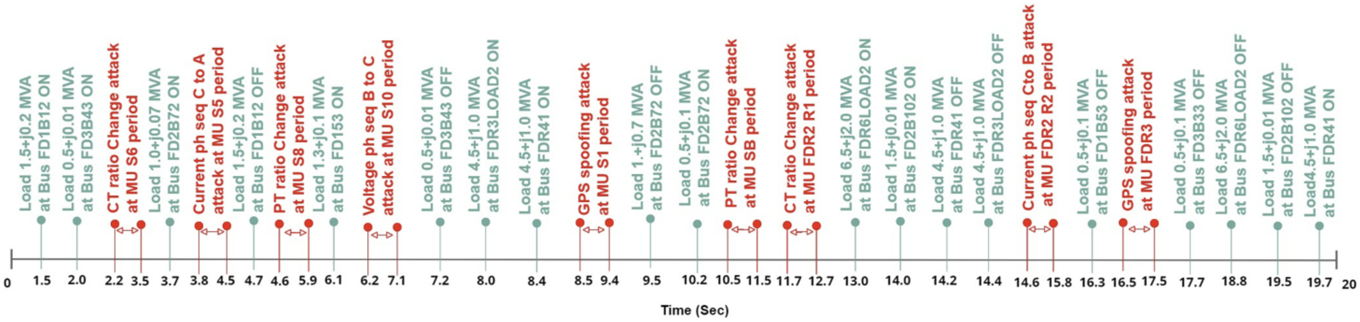

Load 1.5+j0.2 MVA at Bus FD1B12 ON
Load 0.5+j0.01 MVA at Bus FD3B43 ON
CT ratio Change attack at MU S6 period
Load 1.0+j0.07 MVA at Bus FD2B72 ON
Current ph seq C to A attack at MU S5 period
Load 1.5+j0.2 MVA at Bus FD1B12 OFF
PT ratio Change attack at MU S8 period
Load 1.3+j0.1 MVA at Bus FD153 ON
Voltage ph seq B to C attack at MU S10 period
Load 0.5+j0.01 MVA at Bus FD3B43 OFF
Load 4.5+j1.0 MVA at Bus FDR3LOAD2 ON
Load 4.5+j1.0 MVA at Bus FDR41 ON
GPS spoofing attack at MU S1 period
Load 1.+j0.7 MVA at Bus FD2B72 OFF
Load 0.5+j0.1 MVA at Bus FD2B72 ON
PT ratio Change attack at MU SB period
CT ratio Change attack at MU FDR2 R1 period
Load 6.5+j2.0 MVA at Bus FDR6LOAD2 ON
Load 1.5+j0.01 MVA at Bus FD2B102 ON
Load 4.5+j1.0 MVA at Bus FDR41 OFF
Load 4.5+j1.0 MVA at Bus FDR3LOAD2 OFF
Current ph seq Cto B attack at MU FDR2 R2 period
Load 0.5+j0.1 MVA at Bus FD1B53 OFF
GPS spoofing attack at MU FDR3 period
Load 0.5+j0.1 MVA at Bus FD3B33 OFF
Load 6.5+j2.0 MVA at Bus FDR6LOAD2 OFF
Load 1.5+j0.01 MVA at Bus FD2B102 OFF
Load4.5+j1.0 MVA at Bus FDR41 ON
0
1.5
2.0
2.2
3.5
3.7
3.8
4.5
4.7
4.6
5.9
6.1
6.2
7.1
7.2
8.0
8.4
8.5
9.4
9.5
10.2
10.5
11.5
11.7
12.7
13.0
14.0
14.2
14.4
14.6
15.8
16.3
16.5
17.5
17.7
18.8
19.5
19.7
20
Time (Sec)

Figure 11: Full timeline for the test case.

(for some threshold b) is fed into a 2-sided t-test with significance level of 5%. Values of $b = 10^{-5}, 10^{-18}$ were tried based on the eigenvalues of the training data (which filtered most and few of the eigenvalues respectively), and we used 10^{-18} to report results. We found that the hyperparameter b is not particularly sensitive, so long as it is large enough to remove noisy changes (ie. small eigenvalues). Another suggestion by the authors were also that tensor products can be used to inflate the dimensionality of the input data to approach the Marchenko-Pastur Law, which dictates the distribution of eigenvalues for which their t-testing assumption would hold. However, we observe that calculating the eigenvalues is an optimization step that is already inefficient and not preferable for deployment, so we only report results without using tensor products. Tensor products also inhibit the interpretability of the model (somewhat like kernel methods) which is another downside.

Root Cause Discovery (RCD) [8] RCD also runs online and requires hyperparameter k that determines the size of the localized set of anomalous variables. This is generally not straightforward to choose in practice, because k asserts apriori knowledge on the number of sensors that will be attacked. Nevertheless, to remove ambiguity of hyperparameter choice, we *assume* apriori knowledge on the number of attacked sensors. This inflates the performance of RCD, but as we observe in our experiments, the lack of domain knowledge in RCD degrades the localization strength while being too inefficient to be run in deployment. Even with inflated results, RCD still performs poorly compared to PASAL.

E.2 Physics-based Method

To make this paper self-contained for an AI audience, we also detail the physics-based method used in Section 5.

Decoupled Chi-Squared (Dχ^2) [6] The conventional chi-square test operates under the assumption that the computed metric follows a chi-squared distribution. However, the assumption is not fully valid, introducing an approximation that could yield erroneous outcomes, such as failing to detect existing anomalies in the data. Hence, we chose to use the modified decoupled chi-squared test (which we refer to as Dχ^2) with test statistic

$$\Psi_m(\hat{x}) \simeq \sum_{i=1}^{m} \frac{(z_i - h_i(\hat{x}))^2}{\Omega_{ii}} \tag{7}$$

where $(z_i - h_i(\hat{x}))$ is the measurement residual vector and Ω_{ii} is the variance of the i^{th} measurement residual. The performance of Dχ^2 surpasses the conventional chi-squared method [6], it performed poorly when compared to PASAL.