Detection of False Data Injection Attacks in Cyber-Physical Systems using Dynamic Invariants

Kiyoshi Nakayama
NEC Laboratories America
San Jose, CA USA
knakayama@nec-labs.com

Nikhil Muralidhar
Virginia Tech
Arlington, VA USA
nik90@vt.edu

Chenrui Jin
NEC Laboratories America
San Jose, CA USA
cjin@nec-labs.com

Ratnesh Sharma
NEC Laboratories America
San Jose, CA USA
ratnesh@nec-labs.com

***Abstract*—Modern cyber-physical systems are increasingly complex and vulnerable to attacks like false data injection aimed at destabilizing and confusing the systems. We develop and evaluate an attack-detection framework aimed at learning a dynamic invariant network, data-driven temporal causal relationships between components of cyber-physical systems. We evaluate the relative performance in attack detection of the proposed model relative to traditional anomaly detection approaches. In this paper, we introduce Granger Causality based Kalman Filter with Adaptive Robust Thresholding (G-KART) as a framework for anomaly detection based on data-driven functional relationships between components in cyber-physical systems. In particular, we select power systems as a critical infrastructure with complex cyber-physical systems whose protection is an essential facet of national security. The system presented is capable of learning with or without network topology the task of detection of false data injection attacks in power systems. Kalman filters are used to learn and update the dynamic state of each component in the power system and in-turn monitor the component for malicious activity. The ego network for each node in the invariant graph is treated as an ensemble model of Kalman filters, each of which captures a subset of the node's interactions with other parts of the network. We finally also introduce an alerting mechanism to surface alerts about compromised nodes.**

***Index Terms*—Cyber-Physical Systems, Anomaly Detection, False Data Injection Attacks, Bad Data Detection, State-Estimation, Bayesian Filtering, Kalman Filter, Dynamic Invariant Network, Robust Estimation**

I. Introduction

Cyber-physical systems are becoming more and more complex especially with IoT (Internet of Things) devices integrated into various platforms as in digital-twin systems that represent the physical devices using data. The power grid in particular is a representative example of a complex cyber-physical system consisting of multiple power generation, transmission, and distribution components all interacting with each other to maintain the stability of the system in a large geographic area. The security and reliability of the power system has a significant impact on the smooth functioning of society.

To ensure data fidelity, bad data detection (BDD) techniques are employed by the control center to filter data before it is used for state estimation purposes. If however, false measurements are able to circumvent the BDD layer, they could adversely affect the outcome of state estimation, thus misleading the automatic control algorithms, resulting in catastrophic consequences such as brownouts or blackouts in a power grid.

The security of a system as complex as the modern electric grid is non-trivial to ensure as the large number of inter-dependent components render the system vulnerable to cyber-attacks. One such type of attack involves malicious entities manipulating data from devices like smart meters, being transmitted to the control center for state estimation. Essentially such an attack comprises of the attacker trying to inject an attack vector into a subset of devices they have successfully compromised that transmit diagnostic data to the control center. The effect of such an attack is that the control center receives data that is not representative of the true state of the components transmitting the data and hence such an attack is termed a False Data Injection Attack (FDIA).

Traditional bad data detection approaches based on weighted least squares techniques are susceptible to stealthy false data injection attacks as outlined in [1]. The authors in [1] outline two different types of false data injection attacks, namely *random false data injection attacks* and *targeted false data injection attacks*. In this paper, we focus on detecting random false data injection attacks although the proposed model can also be used to detect targeted false data injection attacks. We showcase how the proposed approach is able to detect stealthy random false data injection attacks that the traditional weighted least squares based state estimation procedures are unable to detect according to [1].

We address the problem of false data injection attacks by proposing a dynamic invariant network with *Granger Causality based Kalman Filter with Adaptive Robust Thresholding* (G-KART), leveraging latent functional relationships between the components in a power system that may not necessarily be represented by explicit power flow equations. By using data-driven learning techniques for state estimation, we are able to model a wider array of component-level relationships at any given time because the pair-wise data driven invariant learning is not restricted to learning relationships only between components that are directly connected in the power system topology.

Instead of considering topological relationships between components, we learn statistically significant predictive causal relationships between any pair of components in the power system through an inductive learning procedure. We model the state of each component in the power system as an aggregation of individual pairwise functional relationships learned between said component and other components of the power network using the meter measurements obtained from them.

Our contributions are as follows:

- We propose a data-driven dynamic invariant learning framework with a temporal causal network learning based approach to detect stealthy False Data Injection attacks in power systems.
- We introduce a novel adaptive robust thresholding procedure to accommodate for effective anomaly detection even in the context of noisy training data laden with outliers and gradual data distribution changes.
- We augment the invariant learning approach proposed in [2] to be governed by using a Granger Causality F-test to learn only statistically significant causal relationships and eliminate the need for manual thresholding (we eliminate the minimum acceptable threshold τ as defined in [2]). The F-test however is flexible enough to accept lag values and significance levels if the user wishes to specify them but will return an invariant network without either of these specified.

II. Background and Related Work

A. Background

1) Power System: A power system is defined as a complex network consisting of generation centers, transmission lines and transformers [3]. A power system has a matrix termed the Jacobian $H \in R^{m \times n}$ that is constructed depending on line impedances and the network topology of a power system. The variable m represents the number of meters providing measurements of active power P or reactive power Q flowing from or to a bus (P_f, Q_f), and n represents the number of state variables to be estimated. A control center is employed to monitor and control the various operations of devices in the power systems.

2) State Estimation: To ensure resilient operation of power systems even when some components fail, power engineers employ meters to monitor different parts of the network and collect their readings at the control center. These meter measurements (active power P, and reactive power Q) recorded at each bus are used to estimate the states of power system variables (voltage V and phase angle θ) of the buses in the power system.

If we define $\mathbf{x} = \{x_1, x_2, .., x_n\}$ to be the set of n state variables and $\mathbf{z} = \{z_1, z_2, ..., z_m\}$ as the m bus measurements recorded by the meters, the goal of the state estimation problem is to estimate the values of the vector of state variables $\mathbf{x}$ using the measurements $\mathbf{z}$. If the Jacobian matrix is represented as $\mathbf{H}$, and measurement errors are defined by $\mathbf{e} = \{e_1, e_2, .., e_m\}$, then the state estimation problem can be defined as in equation 1 wherein the goal is essentially to find a vector $\hat{\mathbf{x}}$ that is a good approximation of the vector of state estimates $\mathbf{x}$.

$$\mathbf{z} = \mathbf{H}\hat{\mathbf{x}} + \mathbf{e} \tag{1}$$

If measurement noise is assumed to be normally distributed with zero mean, then equation 2 represents the solution to the state estimation problem [1]. Here, the matrix $\mathbf{W} \in R^{m \times m}$ is a diagonal matrix where each entry w_{ii} is a reciprocal of the variance of meter error of meter i.

$$\hat{\mathbf{x}} = (\mathbf{H^T W H})^{-1}\mathbf{H^T W z} \tag{2}$$

3) Bad Data Detection: Many techniques for bad data detection have been proposed to protect power systems [4], [3]. The L_2 norm of the measurement residual $||\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}||_2$ has essentially been proposed as a bad data detection procedure. Essentially if $||\mathbf{z} - \mathbf{H}\hat{\mathbf{x}}||_2 > \tau$ for some expert defined threshold τ, the BDD procedure indicates the presence of bad measurements.

4) False Data Injection Attacks: For a power system with m measurements and n state estimates, the measurements and estimates are related through the Jacobian matrix $\mathbf{H} \in R^{m \times n}$. For our purposes, we assume the attacker has complete knowledge of the composition of matrix $\mathbf{H}$. This can be obtained by compromising the control center network or through social engineering or other such approaches. We consider the scenario of *Random False Data Injection Attacks*, in which the goal of the attacker is to generate a random attack vector $\mathbf{a} \in R^{k \times 1}$ for any subset of k compromised meters to cause a wrong state estimation of state variables at the control center [1]. For ease of notation we assume $\mathbf{a} \in R^{m \times 1}$ for a subset of k compromised meters where values at all $m - k$ indices of uncompromised meters are 0 in $\mathbf{a}$. If $\mathbf{z_a} = \mathbf{z} + \mathbf{a}$ then it has been shown that any vector $\mathbf{a}$ can be injected into measurements to circumvent the L_2-norm based BDD procedures as long as $\mathbf{a} = \mathbf{Hc}$ where $\mathbf{c}$ is any arbitrary non-zero vector. Essentially, an attack vector $\mathbf{a}$ is a linear transformation of $\mathbf{H}$.

In contrast to traditional BDD approaches where the matrix $\mathbf{H}$ is used, we ignore $\mathbf{H}$ and instead focus on learning purely data-driven functional relationships between the various components in a power system that we refer to as an invariant graph. This has two effects

1) It ignores the system topology and properties to a certain extent and hence the attacker now only has partial information (as he only has access to the Jacobian matrix $\mathbf{H}$ of the system and is unaware of the functional relationships learnt by the invariant network.)
2) Any component in the network could essentially be connected to any other component in the network through a functional relationship. So, to launch a successful attack, the attacker must on average compromise more components for the FDI attack to remain undetected.

The aforementioned effects ensure that the FDI attack is no longer always stealthy and even if a stealthy attack is launched, the invariant network significantly increases the average cost of the attacker to launch successful stealthy FDI attacks.

B. Related Work

There have been many efforts undertaken for detecting false data injection attacks in power systems.

In the context of an FDI attack, the Jacobian matrix has been compromised and the attacker has complete access to it [1]. This is imperative because the Jacobian matrix is used to estimate the values of phase angles and voltages of the buses from the active and reactive power measurement values respectively. Naturally, if an attacker were to inject false values at certain points in the active or reactive power readings before they were recorded at the control center, the subsequent estimates of voltage and phase angle would be affected. This process of false data injection could be used by the attacker to govern how certain processes occur in the power system or to destabilize the system.

Hence, if BDD methods were only based on system topology and weighted least squares procedure as is the case with bad data detection in traditional power systems, it would be relatively simple for the attacker (who now has an inherent understanding of the line and component properties) to launch an attack that would be stealthy and pass the Bad Data Detection (BDD) procedure. It is here that our procedure of data driven invariant learning models comes into play.

In [4], the author describes a system for obtaining measurements of active power (P), reactive power (Q), voltage (V) and phase angle (θ). The National Communications System document on SCADA systems alludes to the risks posed by FDI attacks on SCADA systems and gives a high priority of 2 to the development of bad data detection systems for False Data Injection attacks on SCADA based systems.

Another related problem which is prone to false data injection attacks is one of energy theft detection problem, as outlined by [5]. The authors in the aforementioned paper outline methods both for power theft and possible detection techniques.

There have also been efforts undertaken to develop different attack scenarios to aid the development of more robust detection procedures. The authors in [6] propose a "least-effort" attack on a power system [1] as mentioned previously in addition to developing two separate paradigms of stealthy false data injection attacks also develop a low-cost attack strategy wherein the attacker can de-stabalize the system compromising a minimal number of components.

The authors in [7], [8], [9], [10] propose alternate formulations for detecting False Data Injection Attacks in a DC power setting. A generic survey of recent anomaly detection techniques in statistical learning literature is provided in [11]. An exhaustive survey for Cyber Security in the smart grid, along with a good description of the structure of the smart grid is provided by [12].

III. Problem Formulation

Let us assume we have a set of n time series $S = \{X_1, .., X_n\}$ where each X_i represents a sensor at bus i measuring a particular metric like active-power flowing out from bus i (P_{f_i}) in a power system. X_i^t represents a measurement recorded at a particular bus i at time t. Each component in the power system is affected either directly or indirectly by the other components.

Modeling pair-wise component relationships in sensor networks has proven useful in anomaly detection tasks in cyber-physical systems as demonstrated in [2], [13], [14].

A. Dynamic Invariant Graph Construction & Model Learning

1) Temporal Causal Networks: We model the pair-wise component relationships to construct a Temporal Causal Network as a dynamic invariant network. Learning temporal predictive causality is a popular concept in many fields like biology, social science and climate science. Although many approaches based on randomization, cross-correlation etc. have been adopted [15], [16], [17], we adopt a popular regression-based method for uncovering temporal causality called Granger causality [18] for construction of the causality graph.

The basic idea as enumerated in [19] states that a variable X_j is the cause of another variable X_i if the past values of X_j are helpful in predicting the future values of X_i. If we were to consider the two autoregressions in equations 3 and 4:

$$X_i^t = \sum_{l=1}^{L} a^l X_i^{l-1} \tag{3}$$

$$X_i^t = \sum_{l=1}^{L} a^l X_i^{l-1} + b^l X_j^{l-1} \tag{4}$$

with L being the maximum time lag, X_j is said to Granger cause X_i if the predictions of equation 4 are significantly better than predictions of X_i by equation 3.

For each pair of time series in S, we utilize the F-test to determine statistical significance wherein if the null hypothesis (X_j does not cause X_i) is discounted with a confidence level of higher than α, we assume that X_j has a relationship of predictive causality with X_i denoted by a directed edge from X_j to X_i in the temporal causal network (a.k.a invariant network). Here, α usually called the significance level indicates the probability of type 1 errors i.e. the probability of wrongly indicating that X_j causes X_i. In order to retain only strong relationships of temporal causality, we set $\alpha = 0.01$.

The temporal causal network learning procedure culminates yielding a graph $G = (V, E)$ wherein an edge $e_{ij} \in E$ from node v_i to $v_j | \{v_i, v_j\} \in V$ indicates that time series X_i has a temporal causal effect on time series X_j.

2) Kalman Filter Modeling: Once the temporal causal network $G = (V, E)$ is learned, each predictive causal relationship $e_{ij} \in E$ in G is modeled with a Kalman Filter. Each node $v_j \in V$ represents a bus in the original power system and has a set of k Kalman filters monitoring its state at each time step where k represents the number of incoming edges (temporal causal relationships) that node v_j is involved in. A Kalman filter K_{ij} represents the model monitoring the state of node v_j at each time step, using the historical data from X_i, X_j. In addition to pairwise data, we also incorporate non-linear system states through a deep robust autoencoder mechanism as depicted in [20].

$$\hat{X}_{ij}^{t} = a_i X_{ij}^{t-1} + a_j X_{ji}^{t-1} + \epsilon \quad (5)$$

The predictions of model K_{ij} at each time step t are calculated according to equation 5. Here, the predicted state of the measurement at bus j at time t is represented by $\hat{X}_{ij}^{t}$. ϵ represents the prediction error and is assumed to be normally distributed. X_{ij}^{t-1} and X_{ji}^{t-1} are state estimates for the states of X_j, X_i respectively at time $t-1$. The weights $A = [a_i, a_j]$ are estimated from a subset of the data using expectation maximization. A detailed account of the Kalman filter formulation has been provided by [21].

B. Anomaly Detection with Adaptive Robust Thresholding

1) Constant Threshold: If $\hat{X}_{ij}^{t}$ and $\bar{X}_{j}^{t}$ represent the predicted and actual values respectively of bus_j at time t, a particular invariant relationship $e_{ij} \in G$, e_{ij} is said to be broken if equation 6 is violated.

$$|\bar{X}_{j}^{t} - \hat{X}_{ij}^{t}| < \epsilon_{ij}^{0} \quad (6)$$

This residual based anomaly threshold used by [22] is adopted in an effort to reduce false positive rates of broken invariant relationships in the graph G during testing. According to [22], ϵ_0 can be estimated from the residuals in the training phase to be 10% larger than the tolerance of residuals as given by equation 7.

$$\epsilon_{ij}^{0} = 1.1 * arg_r\{Prob(|\bar{X}_{j}^{t} - \hat{X}_{ij}^{t}|) < 0.995\} \quad (7)$$

2) Adaptive Robust Thresholding: However, in the context of FDI attacks, the assumption that the training data is free of a significant portion of outliers is not a sound assumption to make. Hence, we use data laden with noise both for training and testing due to which the models and thresholds learned need to be *robust*. Unfortunately the thresholding methodology in equation 7 is susceptible to yielding an overestimated anomaly threshold as it is sensitive to noise and outliers.

We augment the learned thresholding procedure with an adaptive component in the testing phase as outlined in equation 8.

$$\epsilon_{ij}^{t+1} = \beta * \epsilon_0 + (1-\beta) * \mu_{|t-w:t|} \quad (8)$$

ϵ_{t+1}^{ij} represents the adaptive anomaly threshold for time step $t+1$ for relationship e_{ij} in the temporal causal graph G. The adaptive threshold is a convex combination of the constant residual based anomaly threshold learned during the training phase and the term $\mu_{|t-w:t|}$ which is a rolling window based *median* of the residuals in the past window of size w. This new procedure is capable of yielding good performance even with noisy data by adapting the anomaly threshold to underlying changes in the data distribution. The procedure is *robust* due to the inclusion of the adaptive median thresholding component as the median is known to be a robust statistic.

Equation 9 represents a modified version of equation 6 in the context of adaptive thresholding.

$$|\bar{X}_{j}^{t} - \hat{X}_{ij}^{t}| < \epsilon_{ij}^{t} \quad (9)$$

Algorithm 1: Bi-variate Temporal Causality Model Training

```
Input  : $S = \{X_1, .., X_n\}$: Input time series,
         $t_s$: Training Period Start,
         $t_e$: Training Period End,
         $G = (V, E)$: Temporal Causality Inv. Network
Output: K: Temporal Causality Model Matrix
K ={};
for $e_{ij} \in E$ do
    /* Fit Kalman Filter using $X_j$ and
       $X_i$ to estimate state of $bus_j$    */
    $K_{ij}$ = KalmanFilter($X_j^{t_s:t_e}$,$X_i^{t_s:t_e}$) ;
    K = K $\cup K_{ij}$;
end
```

C. Problem Formulation Summary

With more accurate predicted values $\hat{X}_{ij}^{t}$ and adaptive threshold ϵ_{ij}^{t}, we try to increase or maintain a good range of the True Positive Rate (TPR) for FDI attack detection when the systems are compromised with random attacks. At the sane time, we want to minimize the False Positive Rate (FPR) as much as possible since it is also equally important not to be bothered by the overfitting issues as otherwise critical attacks could be overlooked. We also aim at achieving a higher F1-score that is a combination of the precision and recall of the model.

IV. FDI Attack Detection Framework

We extend the framework described in [2], [14] by incorporating Granger-causality based invariant learning to model the power of predictive causality of different components in the power network. We also augment the anomaly detection procedure with an adaptive robust thresholding mechanism.

A. Model Training

In the model training phase, we describe how to produce a Bi-Variate Temporal Causality Model as a Temporal Causality Model Matrixl. A Kalman filter K_{ij} represents the model monitoring the state of node v_j at each time step, using the historical data. Based on Algorithm 1, the set K of Kalman Filters is obtained as an output.

Multiple Kalman filter models are trained to monitor the state of each bus in the power network, and the training process is governed by the aforementioned Granger Invariant Network, i.e. for each directed edge (from a source node to a sink node) in the invariant network, we train a Kalman Filter to predict the state of the sink node at the next time step, given historical data from the sink and the source nodes. At the end of the training process, a node with k invariant relationships essentially has an ensemble of k different Kalman filters monitoring its state at each time step.

B. Ensemble Anomaly Detection and Alerting Framework

At any time step, if the error in state prediction of an invariant model is greater than a pre-calculated threshold, the invariant edge between the two nodes in question is said to be broken/invalidated. We can consider that in this case, one of the k models in the ensemble for the sink node has predicted that an anomaly has occurred at the current time step in the sink node. If a majority ($> 50\%$) of the bivariate invariant relationships for a particular node in the system are invalidated, we declare the component to have experienced an anomaly. Hence we employ a majority-voting ensemble model for anomaly detection at each bus.

The procedure of the ensemble anomaly detection and alerting framework is described in Algorithm 2. Consider a bus i in the power system that has k temporal causal relationships (incoming edges) monitoring its state. This signifies that there are k Kalman filter models (K_{*i}) offering predictions for the state of bus i at each time step t. Each of these models has a concomitant value for the adaptive anomaly threshold which along with the actual measurement at bus_i at time t is used to determine whether or not the edge e_{ji} is broken for each of the k invariants. If greater than 50% of the invariants of bus_i are broken, a FDI attack is said to have occurred at bus_i at time t and an alert is sent out.

This is similar to a majority voting paradigm in ensemble modeling. Hence, we might consider the state estimation and anomaly detection framework for each bus in the power system, as a majority voting based ensemble model that sends out alerts if a majority of the temporal causal relationships of a particular bus are broken at a particular time step.

Algorithm 2: Ensemble Anomaly Detection & Alerting

```
Input  : $S = \{X_1, .., X_n\}$: Input time series,
         $t_e$: Testing Period Start,
         $t_{end}$: Testing Period End,
         $\xi = 0.5$: Alert Threshold,
         $G = (V, E)$: Temporal Causality Network,
         K: Temporal Causality Model Matrix
foreach $t \mid t_e < t < t_{end}$ do
    foreach $v_j \in V$ do
        votes = 0;
        Let $E(v_j)$ be edges connected to $v_j$;
        /* Iterate over all incoming
           edges of $v_j$                     */
        foreach $e_{ij} \in E(v_j)$ do
            $\hat{X}^t_{ij}$ = $K_{ij}.predict(\bar{X}^{t-1}_j, \bar{X}^{t-1}_i)$;
            Calculate the adaptive robust threshold $\epsilon^t_{ij}$
              based on eq. (8);
            if $|\bar{X}^t_j - \hat{X}^t_{ij}| > \epsilon^t_{ij}$ then
                votes += 1;
            end
        end
        $score_j = \frac{votes}{|E(v_j)|}$;
        if $score_j \geq \xi$ then
            Invoke Alert: FDI at $v_j$ at time step $t$;
        end
    end
end
```

V. Experimental Results

A. Dataset Description

We conducted experiments on an IEEE 33 bus dataset consisting of one generation station (Bus1), a PV setup attached to Bus 33 and an electric vehicle charging station at Bus 10. The network topology of the IEEE 33 bus network has been depicted in figure 1a. For the purposes of this experiment, we focus on the active power flowing out of each bus in the IEEE 33 bus network denoted P_f. We denote P_f flowing from Bus i to j as $Bus_i - Bus_j$. Figure 1b represents the data-driven functional relationships learned between buses that may be explicitly or implicitly connected. In this case, it shows the Bivariate Granger Invariant Network modeling the active power flowing out of each bus (P_f) in an IEEE 33 bus dataset. Figure 1b indicates each node Bus_i-Bus_j signifies measurements of active power flowing from Bus i to Bus j.

B. Random False Data Injection attack

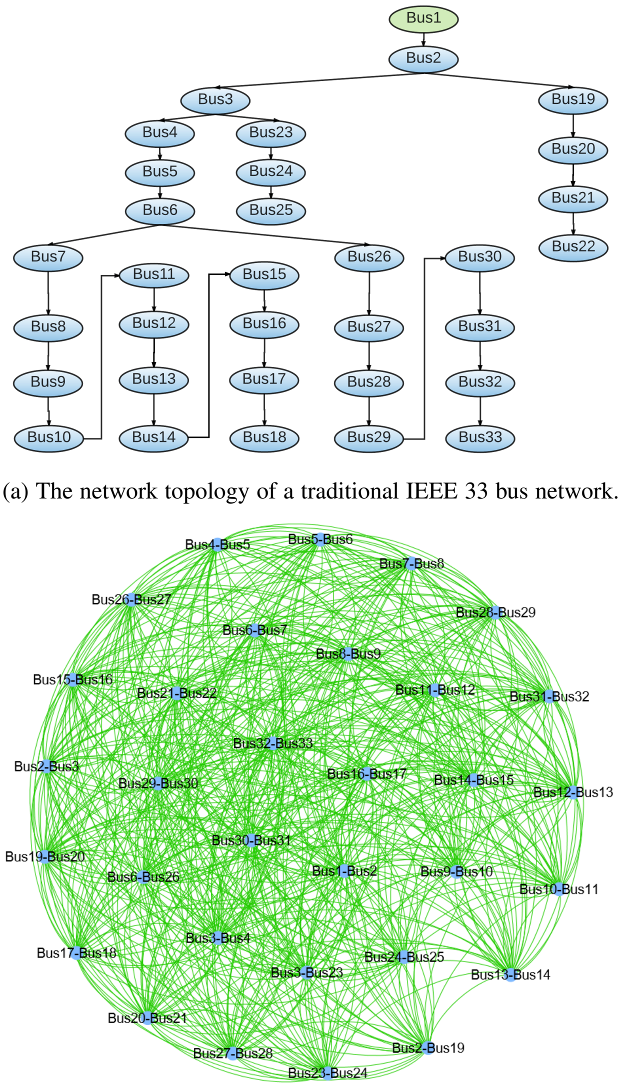
We injected anomalies on the synthetic data of active power flowing from each branch (P_f) at specific points (depicted as red dots in figure 2).

We first randomly select 30% of the 33 buses to be attacked. These buses can be considered to have been compromised by the attacker. We then design an attack vector using the lightweight attack vector construction method for stealthy random false data injection attacks as outlined by [1]. We developed a false data injection mechanism that randomly injects the chosen attack vector at different time steps in each of the compromised buses so the total attack percentage at each compromised bus is δ. We discuss results for $\delta = 10\%$ i.e. when 10% of the data points have been affected by false or noisy data injected by the attacker.

We first baseline by using popular anomaly detection algorithms namely, the One Class SVM and Isolation Forest. The performance comparison between the baselines and G-KART have been depicted in table I. We observe that the G-KART model, which is the dynamic invariant learning model we propose in this paper based on Granger causality based adaptive robust thresholding, has the highest F1-score that is a combination of the precision and recall of the model. The G-KART also has a much lower False Positive Rate than the Isolation Forest and One Class SVM models which is an essential characteristic for any anomaly detection procedure.

VI. Discussion

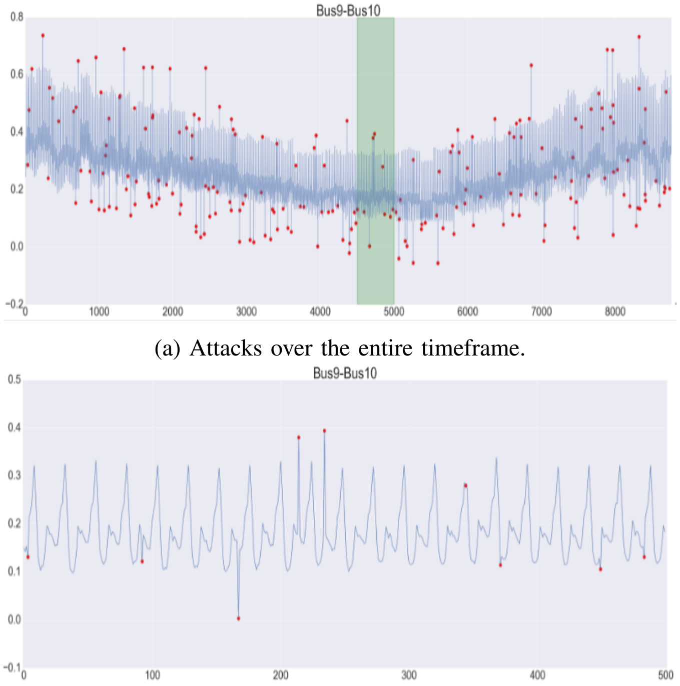
We discuss the performance of detection frameworks of FDI attacks using data distribution analyses. In figure 3a, the green curve showcases the original data distribution of active power (P_f) from Bus9 to Bus 10 and the yellow curve showcases the data distribution with false data injected. Figure 3b indicates clean data distribution in green and noisy data distribution in yellow of only those time steps when false data was injected into the active power flowing from bus 9



(a) The network topology of a traditional IEEE 33 bus network.

(b) The Bivariate Granger Invariant Network modeling the active power flowing among IEEE 33 bus systems.

Fig. 1: IEEE 33 bus network and its invariant graph.



(a) Attacks over the entire timeframe.

(b) Attacks between $t = 4500$ and 5000.

Fig. 2: Random FDIA on active power flowing from Bus 9 to Bus 10.

Method Metric	G-KART	One Class SVM	Isolation Forest
FPR	**0.0658**	0.467	0.285
TPR	0.356	**0.39**	0.322
F1-Score	**0.141**	0.0717	0.0617

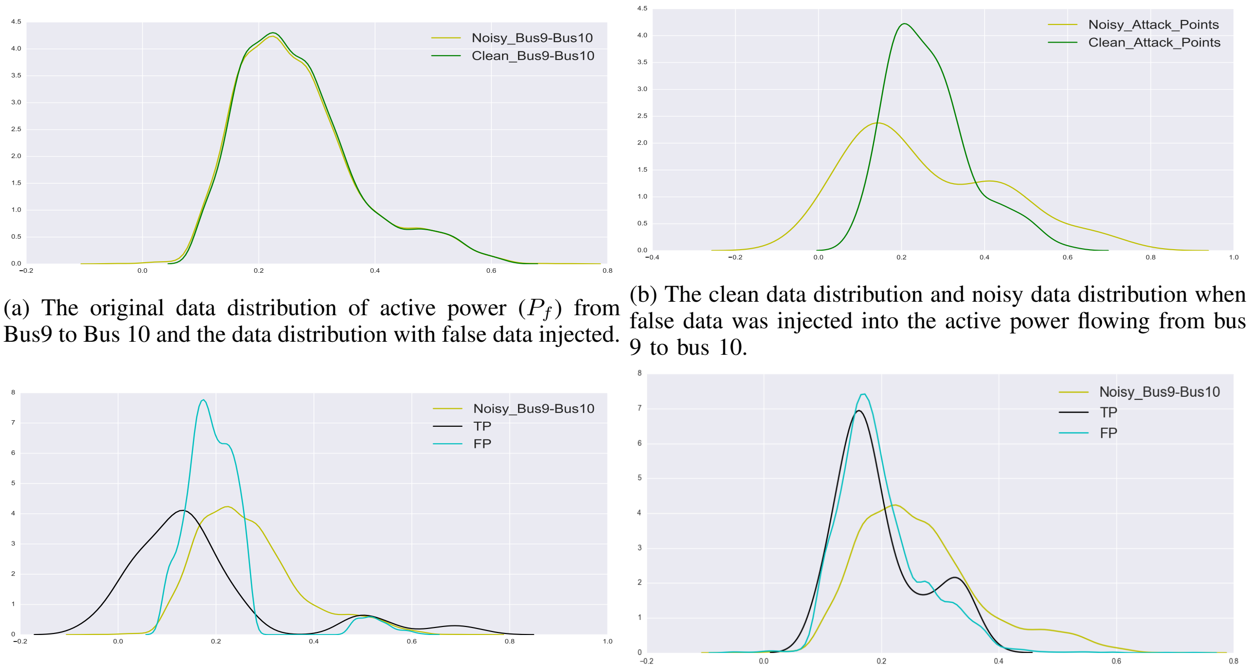
TABLE I: Performance Comparison of a Random False Data Injection Attack with an attack percentage of 10% of the time steps across 30% of the buses in the IEEE 33 Bus system using the P_f metric.

Method Metric	G-KART	G-KCT	G-KAT
FPR	0.0658	**0.0076**	0.048
TPR	**0.356**	0.1309	0.333
F1-Score	**0.141**	0.0754	0.14

TABLE II: Performance Comparison between constant and adaptive thresholding methodologies of a Random False Data Injection Attack with an attack percentage of 10% of the time steps across 30% of the buses in the IEEE 33 Bus system using the P_f metric.

to bus 10. Figure 3c presents the classification performance of the 1-class SVM model by depicting the distributions of the True Positive (in black) and the False Positive (in cyan) classifications compared with the full data distribution for a particular bus (in yellow). Figure 3d showcases a similar true positive, false positive classification distribution along with the true data distribution. We notice that in both the case of the One-Class SVM, and the isolation forest, the true positive and false positive distributional overlap indicates the inability of both these models to effectively detect false data injection attacks.

1) One Class SVM and Isolation Forest: Figures 3a-3d and table I indicate that the One-class SVM and Isolation Forest algorithms are unable to perform effectively in the context of stealthy random FDI attack detection. Traditional SVMs are maximum margin classifiers where a separating hyperplane is learned between each pair of distinct classes such that the distance separating the classes is maximized. In the same vein, One class SVMs learn a single class which they consider the normal class. In our case, this class should ideally contain all instances where there is no attack. All other instances that do not belong to the normal class are deemed anomalies. In our case, this should ideally contain all time steps where false data was injected into a particular bus. If we observe figure 3a, we notice that overall, there is no discernible difference



(a) The original data distribution of active power (P_f) from Bus9 to Bus 10 and the data distribution with false data injected.

(b) The clean data distribution and noisy data distribution when false data was injected into the active power flowing from bus 9 to bus 10.

(c) The classification performance of the 1-class SVM model by depicting the distributions of the True Positive (in black) and the False Positive (in cyan) classifications compared with the full data distribution for a particular bus (in yellow).

(d) The classification performance of the Isolation Forest by depicting the distributions of the True Positive (in black) and the False Positive (in cyan) classifications compared with the full data distribution for a particular bus (in yellow).

Fig. 3: Analyses with data distribution.

in the densities of the original data and the original data with false data injected. However, a significant difference in the plotted densities exists in figure 3b, wherein only the subset of points at which false data is injected is considered, primarily to showcase that the attack vector injection does indeed have a local effect on the data distribution.

However, if we observe the false positive (cyan) and true positive (black) curves, the significant overlap between them in figures 3c and even more so in 3d, indicates that the One class SVM and Isolation Forest are unable to effectively discriminate instances of FDI attack data from instances where the data is free from attack. This is primarily owing to there being no discernible change in the global distribution when false data is injected but only a change in the local property of the time series around the region of attack. The inability of these models to capture this local data infidelity accounts for their high false positive rate and concomitantly low f1-score depicted in table I. This inability to discriminate between true and false data distributions is because of the subtle perturbations in the time series that are caused by stealthy false data injection as indicated in figure 2a and 2b.

Figure 2a depicts the full time series of an example bus i.e. the active power flowing from Bus_9 to Bus_{10} with false data points depicted as red dots. The green box highlights the snapshot of the full time series that has been depicted in the figure 2b, showcasing a subset of the time series. In this figure we are able to observe that while some instances of random FDI cause significant perturbations in the data distribution, there also exist a number of time steps in which the resulting false data point lies within the global data distribution but does not conform to the local time series property. Hence, we argue that, a dynamical system model like the Kalman filter is able to model this temporal variation in the time series as opposed to density based anomaly detection as in the case of One class SVM and Isolation Forest which are unable to capture this local temporal variation. This in addition to the adaptive robust thresholding based on the rolling median leads to the superior performance of the proposed G-KART model.

2) G-KCT, G-KAT, G-KART: We experimented with three different thresholding schemes and hence developed three different models. The comparative performance of each of the three models in the FDI attack detection task has been depicted in table II.

The first model *Granger Causality based Kalman Filter with Constant Thresholding* (G-KCT) uses equation 6 and 7 for constant thresholding and hence is unable to adapt to changes in data during the testing phase as the threshold once set in the training phase does not get updated and the model is also affected by noise in the data.

The second model *Granger Causality based Kalman Filter with Adaptive Thresholding* (G-KAT) uses 8 and 9 for updating the threshold in the testing phase. In this case, the term $\mu_{|t-w:t|}$ is a rolling window *mean* of the residuals in the past window of size w. This model although an improvement over G-KCT, is perturbed by the noise in the training and testing phases.

The third model *Granger Causality based Kalman Filter with Adaptive Robust Thresholding* (G-KART) also uses 8 and 9 for adaptive thresholding. The thresholding methodology in this case is more robust compared to the previous methods because we use the *median* which is a robust statistic. Hence, it is less perturbed by noisy data and is able to adapt to true changes in the underlying data distribution.

VII. Conclusion and Future work

We have proposed a novel robust anomaly detection procedure for detecting false data injection attacks in power systems based on the dynamic invariant learning approach as a temporal causality network learned using the Granger causality framework. It is shown through the simulation results that not only this method maintains the high standard of TPR that is competitive among other techniques that are designed to detect FDIs, but also FPR is significantly reduced compared with One Class SVM and Isolation Forest approaches. In addition, with the adaptive robust thresholding approach, we could significantly improve the TPR while maintaining the low FPR that is the advantage of the FDI detection with the dynamic invariant network analysis.

We have currently showcased preliminary results of the procedure on a single metric of an IEEE 33 Bus power system. We further wish to evaluate the proposed framework against more sophisticated algorithms and other evaluation metrics like the NAB (Numenta Anomaly Benchmark) score and incorporate more sophisticated attacks like targeted FDI attacks and Denial-of-Service (DoS) attacks. Although the FDI detection framework has been applied to energy management of power systems as a critical infrastructure, there would be many digital twin platforms with which the dynamic invariant analysis is integrated such as factory management systems and automobile manufacturing management systems.

References

[1] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[2] M. Momtazpour, J. Zhang, S. Rahman, R. Sharma, and N. Ramakrishnan, "Analyzing invariants in cyber-physical systems using latent factor regression," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2009–2018, ACM, 2015.

[3] A. J. Wood and B. F. Wollenberg, *Power generation, operation, and control*. John Wiley & Sons, 2012.

[4] A. Monticelli, *State estimation in electric power systems: a generalized approach*, vol. 507. Springer Science & Business Media, 1999.

[5] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. S. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Science and Technology*, vol. 19, no. 2, pp. 105–120, 2014.

[6] Q. Yang, J. Yang, W. Yu, D. An, N. Zhang, and W. Zhao, "On false data-injection attacks against power system state estimation: Modeling and countermeasures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 717–729, 2014.

[7] L. Liu, M. Esmalifalak, and Z. Han, "Detection of false data injection in power grid exploiting low rank and sparsity," in *Communications (ICC), 2013 IEEE International Conference on*, pp. 4461–4465, IEEE, 2013.

[8] K.-M. Lee, B. Min, and K.-I. Goh, "Towards real-world complexity: an introduction to multiplex networks," *arXiv preprint arXiv:1502.03909*, 2015.

[9] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on dc state estimation," Preprints of the First Workshop on Secure Control Systems CPSWEEK, 2010.

[10] A. Teixeira, G. Dán, H. Sandberg, and K. H. Johansson, "A cyber security study of a scada energy management system: Stealthy deception attacks on the state estimator," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 11271–11277, 2011.

[11] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[12] W. Wang and Z. Lu, "Cyber security in the smart grid: Survey and challenges," *Computer Networks*, vol. 57, no. 5, pp. 1344–1371, 2013.

[13] G. Jiang, H. Chen, and K. Yoshihira, "Discovering likely invariants of distributed transaction systems for autonomic system management," in *Autonomic Computing, 2006. ICAC'06. IEEE International Conference on*, pp. 199–208, IEEE, 2006.

[14] N. Muralidhar, C. Wang, N. Self, M. Momtazpour, K. Nakayama, R. Sharma, and N. Ramakrishnan, "illiad: Intelligent invariant and anomaly detection in cyber physical systems," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, 2017.

[15] T. La Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects," in *Proceedings of the 19th international conference on World wide web*, pp. 601–610, ACM, 2010.

[16] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[17] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller, "Robustly estimating the flow direction of information in complex physical systems," *Physical review letters*, vol. 100, no. 23, p. 234101, 2008.

[18] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[19] M. T. Bahadori and Y. Liu, "Granger causality analysis in irregular time series," in *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 660–671, SIAM, 2012.

[20] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 665–674, ACM, 2017.

[21] G. Bishop and G. Welch, "An introduction to the kalman filter," *Proc of SIGGRAPH, Course*, vol. 8, no. 27599-23175, p. 41, 2001.

[22] A. B. Sharma, H. Chen, M. Ding, K. Yoshihira, and G. Jiang, "Fault detection and localization in distributed systems using invariant relationships," in *Dependable Systems and Networks (DSN), 2013 43rd Annual IEEE/IFIP International Conference on*, pp. 1–8, IEEE, 2013.