electronics

MDPI

Article

Gaussian Mixture Modeling for Detecting Integrity Attacks in Smart Grids

Stavros Ntalampiras [1,*] and Yannis Soupionis [2]

[1] Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milan 20133, Italy
[2] European Commission, Joint Research Center, Ispra 21027, Italy; yannis.soupionis@jrc.ec.europa.eu
[*] Correspondence: stavros.ntalampiras@polimi.it or dalaouzos@gmail.com; Tel.: +39-02-23993491

Academic Editors: Alfredo Vaccaro and Jin (Wei)Kocsis
Received: 17 September 2016; Accepted: 15 November 2016; Published: 23 November 2016

Abstract: The thematics focusing on inserting intelligence in cyber-physical critical infrastructures (CI) have been receiving a lot of attention in the recent years. This paper presents a methodology able to differentiate between the normal state of a system composed of interdependent infrastructures and states that appear to be normal but the system (or parts of it) has been compromised. The system under attack seems to operate properly since the associated measurements are simply a variation of the normal ones created by the attacker, and intended to mislead the operator while the consequences may be of catastrophic nature. Here, we propose a holistic modeling scheme based on Gaussian mixture models estimating the probability density function of the parameters coming from linear time invariant (LTI) models. LTI models are approximating the relationships between the datastreams coming from the CI. The experimental platform includes a power grid simulator of the IEEE 30 bus model controlled by a cyber network platform. Subsequently, we implemented a wide range of integrity attacks (*replay*, *ramp*, *pulse*, *scaling*, and *random*) with different intensity levels. An extensive experimental campaign was designed and we report satisfying detection results.

Keywords: critical infrastructure protection; linear time invariant modeling; hidden Markov model; fault diagnosis; cyber-physical system; Gaussian mixtures

1. Introduction

Modern critical infrastructures (CI) rely on advanced information and communication technologies (ICT) which have an important role in their monitoring and control. CIs are assets of high importance to the society, such as electricity networks including generation, transmission and distribution, gas network for production, transportation and distribution, financial services (banking, clearing), transportation systems (fuel supply, railway network, airports, harbours, inland shipping), etc. Typically, an ICT layer is employed since it improves the control of the CI leading to better performance while reducing the operational cost. Controlling a CI by means of an ICT framework aims at (a) production profit maximization; (b) satisfying the service demand while ensuring that the parameters of the underlying process are within predetermined limits; and (c) achieving the required operational performance while preventing the occurrence of any type of undesirable behavior.

In principle, each CI should employ its own ICT infrastructure to maximize security. However, in practice, the tendency is to share the ICT layer for minimizing the cost of installation, operation, maintenance, etc. [1–3]. The specific advantage comes with the drawback of increased vulnerability since a successful attack on the ICT layer may provide control to multiple CIs. The main type of cyber threats occur on Supervisory Control And Data Acquisition (SCADA) systems [4,5]. A recent paradigm is the Stuxnet worm [6] which compromised industrial control systems. Even though cyber attacks are not common, they have a hazardous socioecomonic impact [7]. There have been several recent cyber attacks with very unfortunate consequences, such as (a) in August 2012, the Saudi oil giant Aramco was subjected to a large cyber attack [8] that affected about 30,000 workstations; (b) in April 2012, the big payment processing provider Global Payments confirmed a massive breach [9] that compromised about 1.5 million cards; (c) in January 2013, the U.S. Department of Energy underwent an intrusion [10] to 14 of its servers and many workstations located at the Department's headquarters, aimed at exfiltrating personal information about its employees; and (d) a very recent attack [11] that lasted two years was detected in the Woods Hole Oceanographic Institution. The organization sustained a "sophisticated, targeted attack" which has resulted in the theft of US commercial secrets, potentially sensitive government information, and military data.

The present article focusses on detecting integrity attacks occurring in the ICT layer. These can be implemented either by affecting the power grid components/equipment, which are responsible for distribution systems, or by manipulating the exchanging protocol messages in order for the attacker to send malicious data to the field device or the control center operator [12]. There exist different lines of thought in the literature: (a) methods detecting *known time-profiles*, so-called signatures, of intrusions, e.g., [13]; (b) approaches using *countermeasures*, i.e., inserting information in the involved signals for facilitating the detection of cyber attacks [14], and (c) *novelty detection* algorithms, i.e., algorithms which search for patterns not existing within the nominal data which may comprise evidence of cyber attacks.

This work concentrates on the third class which is the most generic one, as it can detect integrity attacks of unknown patterns without increasing the computational complexity of the operation of the system. The first class comes with limited capabilities as the attack should have been previously encountered and recorded, not to mention that an experienced attacker would design a novel kind of attack. In our opinion, it may be used as the first line of defence. The drawback of the second class is the fact that it raises computational complexity while there is always the danger of an attacker understanding the authentication scheme by long-term monitoring of the CI and then possessing the ability to permanently exploit the CI undetected.

There are several papers that approach the problem based on the third line of thought. A probabilistic neural network is proposed in [15] for detecting faults appearing in power utilities which are controlled by a broadband network. However, integrity attacks were not considered. A rough classification algorithm is employed in [16], which provides a set of rules for detecting anomalies on data associated with a power system control center. The algorithm requires a quite large amount of data while it is affected by noise which may alter the set of rules significantly. An n-gram analytical model is designed in [17] for detecting anomalies in SCADA systems. The anomalies include simple fault cases without the presence of integrity attacks. It is worth mentioning a similar line of works focusing on the detection of bad data provided by phasor measurement units [18,19].

This paper addresses the problem of integrity attacks occurring in an ICT-controlled CI. Even though the topic is of significant importance, it still needs attention from the scientific community since most approaches examine it only on the side of the cyber layer [20]. The main objective is to design a detector providing a very low (close to zero) false negative rate, while a low false positive rate would be advantageous. The proposed technique is applied on the data coming from the physical CI layer, which in the experiments is the IEEE 30-bus model. We take advantage of the relationships existing within the voltage measurements of each network node. Linear time invariant models are built to approximate each relationship, and the distribution of their parameters is learned by means of Gaussian mixture models. Subsequently, a clustering scheme based on the Kullback–Leibler divergence is used to detect abnormalities in the incoming data. The main novelties of this approach are the usage of a limited amount of data, and its modularity (meaning that new nominal states can be easily inserted in the framework) while there is no need to know any information regarding the nature of the integrity attacks a priori. Finally, it is the first time that a generative modelling technique, such as Gaussian Mixture Models (GMM), is applied onto the specific problem.

In this work, the objective is to quickly detect the attack jeopardizing the operation of the network without localizing it. In such cases, prompt detection is of critical importance and may be proven particularly beneficial towards limiting or even eliminating the hazardous consequences of such events. After an attack detection, the network operator may decide to switch to an operational state which ensures service while restricting, confining or even overriding the usage of the ICT layer suffering the attack. This way, the network can still satisfy the service demand (even without working while operating in an optimal manner) while the attack will have minimum affects on its components. Concurrently, a localization algorithm can be employed (as the one presented for example in [21]) to isolate the component that is not functioning properly and restore its operation. In addition, prompt detection and switching to ICT-free operation may be useful for identifying the root cause of the problem and even locating the attacker due to the very fast detection. It should be noted that most infrastructures can already operate in such a mode since the ICT layer was added later for automatic control of the resources and optimizing the overall performance of the infrastructure. The drawback is the non-optimal control of the infrastructure, which, however, is advantageous when compared to the catastrophic consequences that an integrity attack may cause.

The rest of this article is organized as follows: Section 2 formalizes the problem while Section 3 explains Linear Time Invariant (LTI) and GMM modeling, and the clustering process. The experimental set-up and results are provided in Section 4 while Section 5 concludes this work.

2. Problem Definition

We consider a network of interdependent infrastructures composed by N nodes which can produce either *homogeneous* measurements, i.e., of the same physical quantity (e.g., only voltages), or *heterogeneous* meaning that they are associated with different physical quantities (e.g., voltage phasor, frequency, and rotor angle measurements). The underlying assumption in this work is that the produced measurements are related to some extent since altogether they form a framework providing a particular service, e.g., an electrical distribution network. Thus, the data coming from the nodes are indirectly related to each other. The proposed method does not assume the existence of an analytical model of the specific relationship [22], it rather aims at learning the dynamics of the system solely by inspecting the data. In addition, the method is independent of network topology or routing/synchronization protocol.

Let $X_{i,T_0} = \{X_i(t), t = 1, \ldots, T_0\} : \mathbb{N} \to \mathbb{R}$ be the scalar-in-time datastream produced by the i-th node. At an unknown time instant $T^* > T_0$, an abnormal situation, i.e., an integrity attack, affects node i. No assumptions are made with respect to the magnitude or the time profile of the incident influencing the data generation process. The ultimate goal is to detect promptly and identify correctly the nature of the occurred incident.

3. Probabilistic Modeling

This section describes the way the dataset is modeled for optimizing the detection of malicious events. The basic idea behind the proposed method is to create points in the probabilistic space, where each one describes a case of the normal operation. The larger the dataset, the more the respective points and thus the more accurate the mathematical description of the normal operating modality. The specific points comprise an abstract representation of the system while working under nominal conditions. When we need to test novel data coming from the network for potential abnormalities, we measure the distance between the respective point and the ones describing the normal network state. If the distance is over a threshold, the specific data is associated with an abnormality; otherwise, it is considered that the system is within the normal operating limits. The threshold is set equal to the maximum distance existing among the training points. In the following subsections, we explain the proposed modeling type and algorithm.

3.1. Gaussian Mixture Modeling of LTI Coefficients

The various normal system states are represented by Gaussian mixture models, an important characteristic of which is their ability to approximate with high accuracy every possible distribution as long as a sufficient amount of data is available [23]. A weighted sum of Gaussian functions is fit to the available data while the associated parameters are learned using the Expectation-Maximization (EM) algorithm.

A GMM composed of M Gaussian components is given by the next equation: $p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \sigma_i)$, where x is a continuous-valued data vector, w_i the weight of the i-th component ($\sum_{i=1}^{M} w_i = 1$), and g denotes the Gaussian component. In our case, x represents the coefficients of linear time-invariant models approximating the relationship between the datastreams of the power grid nodes.

The LTI model is described by the general *discrete-time linear Multiple-Input Single-Output* structure [24]:

$$A(z)X_i(t) = \sum_{j=1}^{m} \frac{B(z)}{F(z)} X_j(t) + \frac{C(z)}{D(z)} d(t), \tag{1}$$

where $d(t)$ is an independent and identically distributed random variable accounting for the noise, m is the number of inputs, z is the time-shift operator, while $A(z), B(z), C(z), D(z)$ and $F(z)$ represent z-transfer functions, whose parameter vectors are $\theta_A, \theta_B, \theta_C, \theta_D$ and θ_F, respectively. Consequently an element f_θ in the approximating model family $\mathcal{M}(\subseteq)$ is fully described with a $\theta \in \mathbb{R}^p$, which comprises the above parameter vectors, i.e., $x = \theta$. Following the logic of [25], we create an ensemble of dynamic models (e.g., Autoregressive with Exogenous Input, Autoregressive Moving Average with Exogenous Input, Output Error, etc.) with various orders and select the one which best fits the datastreams (i.e., lowest reconstruction error) while low-order models are preferred. The model search algorithm minimizes a robustified quadratic prediction error criterion [26]. It should be mentioned that the methodology is independent of the model type selection and can be applied using another type without any modifications.

Moving on, it should be mentioned that each density has the following form:

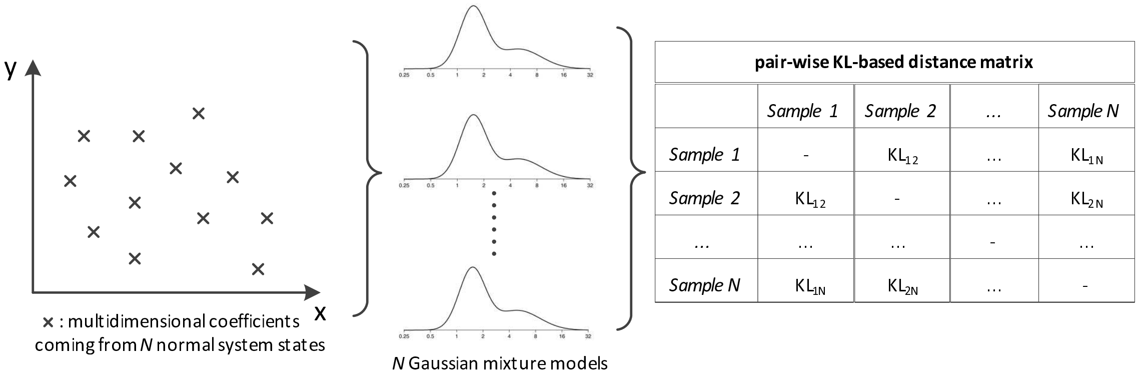
$$g(x|\mu_i, \sigma_i) = \frac{1}{(2\pi)^{D/2}} \exp^{-1/2(x-\mu_i)' \sigma_i^{-1} (x-\mu_i)}, \tag{2}$$

where D denotes the size of x, μ_i the mean vector and σ_i the covariance matrix.

A GMM may be completely defined by the mean vectors μ, the covariance matrices σ and the mixture weights w with respect to every component density. These parameters are typically assembled as $\lambda = \{\mu_i, \sigma_i, w_i\}, i = 1 \ldots M$ while the covariance matrices can belong to a variety of forms, e.g., full, constrained, etc. This choice depends on the requirements of the application and here we employ Gaussian functions of diagonal covariance because they can potentially be equally effective to the full ones [27,28] at a much lower computational cost due to their simplicity. Moreover, by combining Gaussians with diagonal covariance bases, one is able to capture the correlations existing between the elements of the feature vector.

3.2. k-*Nearest Neighbor Clustering Using Kullback–Leibler Divergence*

This subsection details the way the LTI coefficients are represented in the stochastic plane for achieving clustering. Here, the aim is to assess the similarity, denoted as $\mathcal{S}$, between the data under investigation and ones coming from the nominal state which were processed during the training of the system. This work proposes that $\mathcal{S}$ can be estimated via Kullback–Leibler (KL) distances computed in the probabilistic space (see Figure 1). Thus, the normal state is decomposed into a group of GMMs rather than using only one universal model.


y
x
× : multidimensional coefficients coming from *N* normal system states
0.25 0.5 1 2 4 8 16 32
N Gaussian mixture models

pair-wise KL-based distance matrix

	Sample 1	*Sample 2*	…	*Sample N*
Sample 1	-	KL_{12}	…	KL_{1N}
Sample 2	KL_{12}	-	…	KL_{2N}
…	…	…	-	…
Sample N	KL_{1N}	KL_{2N}	…	-

Figure 1. The process leading to the matrix which depicts the distribution of the LTI parameters of the dataset in the Kullback–Leibler sense.

More precisely, each data sequence is transformed to LTI parameters, the distribution of which is then estimated by means of a GMM, i.e.,

$$(X_{i,T_0}, X_{j,T_0}) \rightarrow x_i \rightarrow p(x_i|\mu_i, \sigma_i, w_i), \tag{3}$$

where X_i denotes the voltage datastream of the i-th bus up to time T_0, x the respective sequence of LTI parameters and p its distribution in a Gaussian form.

The next step of the proposed methodology is the computation of the pairwise distances between every GMM pair in the set. For the computation of each distance, a Monte Carlo approximation of the KL divergence is employed, the outcome of which is inversely analogous to the proximity among the involved distributions. For two distributions denoted as $p(x_i|\mu_i, \sigma_i, w_i)$ and $p(x_j|\mu_j, \sigma_j, w_j)$, the KL distance is defined as follows:

$$KL(M||N) = \int p(x_i|\mu_i, \sigma_i, w_i) \log \frac{p(x_j|\mu_j, \sigma_j, w_j)}{p(x_i|\mu_i, \sigma_i, w_i)} dF_n. \tag{4}$$

Due to the absence of a closed-form solution, the above formula is approximated by the empirical mean [29]:

$$KL(M||N) \approx \frac{1}{t} \sum_{k=1}^{t} \log \frac{p(p(x_j|\mu_j, \sigma_j, w_j))}{p(x_i|\mu_i, \sigma_i, w_i)}. \tag{5}$$

This metric represents the distance quite accurately given that t is sufficiently large. In the experiments included in this work, the number of Monte Carlo draws is $t = 2000$. An important detail here is the fact that KL divergence is used as a distance metric, and thus symmetricity is required. However, in general, the quantity defined above is not symmetric, i.e., the distance $KL(M||N)$ may be different than $KL(N||M)$. In order to satisfy this requirement, the following symmetrized form was inferred: $KL(M||N) \rightarrow KL(N||M) \rightarrow KL(M||N) + KL(N||M)$.

Thus, the distance matrix $\mathcal{D}$ is derived with respect to every nominal system state as depicted in Figure 1. This scheme can be parallelized to a $k - NN$ logic where the distance metric is the KL divergence. $\mathcal{D}$ is summed column-wise and the maximum value comprises the threshold $\mathcal{T}$ which discriminates between the normal vs. the rest of the data samples. Algorithm 1 demonstrates the process leading to integrity attack detection while a representative paradigm is depicted in Figure 2. The following section describes the experimental set-up and analyses the results.

Algorithm 1: The algorithm for detecting integrity attacks by receiving data from the physical layer of the interdependent CIs and modeling them by means of GMMs operating on the parameter space of LTI models.

```
1. Build the GMMs representing the nominal class, $G_N = \{\mu_N, \sigma_N, w_N\}$ from the vectors of parameters $\theta_1 ... \theta_d$ each of which associated with a linear dynamic model applied to the training data $O_{i,T_0,i=1,...,d}$ windowized using length $M$ overlapping by $M - 1$;
2. Compute the pairwise distances with respect to all GMMs using the KL divergence $KL(M||N) = \int p(x_i|\mu_i, \sigma_i, w_i) \log \frac{p(x_j|\mu_j,\sigma_j,w_j)}{p(x_i|\mu_i,\sigma_i,w_i)} dF_n$ and derive distance matrix $\mathcal{D}$;
3. Sum $\mathcal{D}$ column-wise and determine the threshold for nominal data as $T_h = \max(sum(\mathcal{D}))$ ;
4. Windowize the incoming novel data as above, which results in windows $W = W_1 ... W_x$;
repeat
    5. j = 1;
    6. Compute the parameter vectors of the $j - th$ dynamic model $\theta_j$ with respect to $W_j$;
    7. Estimate the respective GMM $G_{novel} = \{\mu_{novel}, \sigma_{novel}, w_{novel}\}$;
    8. Compute the KL distances with respect to $G_{novel}$, i.e.,
       $D_{novel} = KL(novel||N) = \int p(x_{novel}|\mu_{novel}, \sigma_{novel}, w_{novel}) \log \frac{p(x_N|\mu_N,\sigma_N,w_N)}{p(x_{novel}|\mu_{novel},\sigma_{novel},w_{novel})} dF_n$;
    9. if $D_{novel} > T_h$ then
        Raise alarm: integrity attack detection;
    else
        Nominal data detected;
    end
    10. $j = j + 1$;
until Detector turned OFF;
```

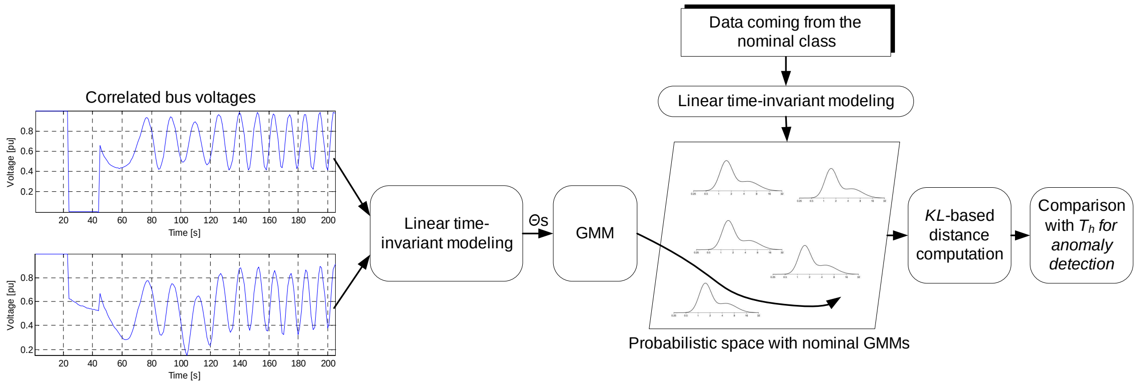

Data coming from the nominal class
Linear time-invariant modeling
Correlated bus voltages
Voltage [pu]
0.8
0.6
0.4
0.2
20
40
60
80
100
120
140
160
180
200
Time [s]
Linear time-invariant modeling
Θs
GMM
KL-based distance computation
Comparison with T_h *for anomaly detection*
Probabilistic space with nominal GMMs

Figure 2. A representative paradigm where the proposed system applies the detection scheme to correlated bus voltages.

4. Experimental Set-Up and Results

This section contains information regarding the experimental platform which was employed in this work along with the presentation and analysis of the experimental protocol and results.

4.1. The Platform

The experimental framework is shown in Figure 3 and includes simulation of border gateway routing protocol, SCADA servers, and the corporate network. The specific representation enables studying of the failures appearing on each component of the cyber layer while simplifying the energy network due to the existence of accurate models. MATPOWER [30] and Matdyn [31] were employed at this phase while the simulation step was carefully chosen so as to realize the real-world process as accurately as possible. To this end, the simulation step was set equal to 20 ms after verifying that the output of real-time simulation reproduces as accurately as possible the real-world process. The case of IEEE 30 Model Bus was used which is considered to be representative of a portion of the American Electric Power System (in the Midwestern US) as of December, 1961 [32]. The equivalent system has 15 buses, two generators, and three synchronous condensers, while it is worth mentioning that the 30-bus test case does not have line limits. The dynamic data of the model were used by the Matdyn software while the optimal power flow problematic was addressed by Matpower. The integrity attacks were performed on the real values of the bus voltages at randomly chosen time intervals as well as buses. It should be mentioned that, during the nominal state, the load fluctuates between margins predefined by the IEEE model while the sampling period is 20 ms.

A relatively recently developed platform is called the Assessment platform for Multiple Interdependent Critical Infrastructures (AMICI) [33]. Here, the physical components are simulated while the cyber layer is emulated for obtaining a more realistic representation of the part where the cyber attacks take place. Emulab [34,35] includes the border gateway routing protocol, SCADA servers, and the corporate network. The specific representation (Figure 3) enables studying of the failures appearing on each component of the cyber layer while simplifying the energy network due to the existence of accurate models.

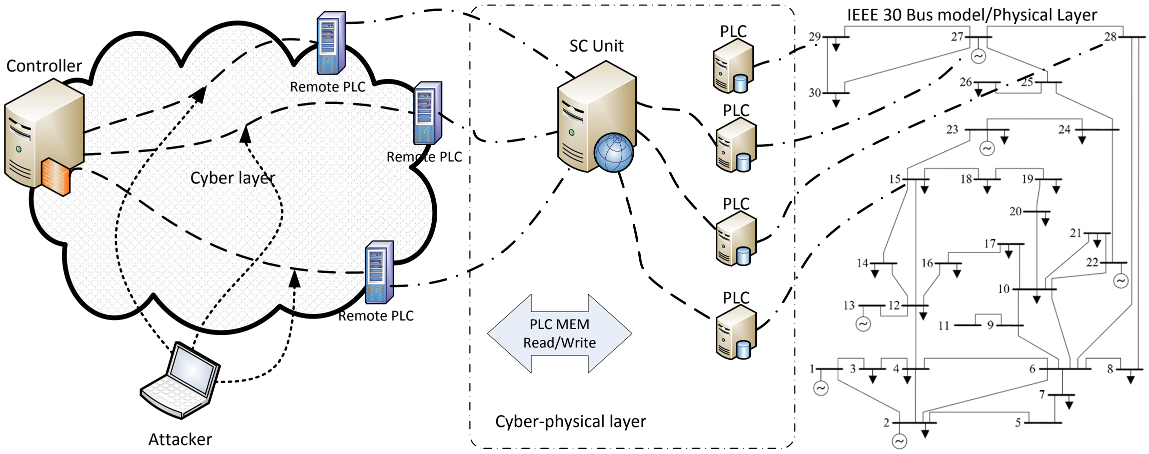


Figure 3. The experimental architecture showing the cyber, cyber-physical and physical layer.

One of the most important issues of the specific set-up is achieving synchronization between the simulated parts. In order to ensure that the simulated model runs at the same rate as the actual physical system, we employed PCs with multitasking OSs for execution. There are five Dell PCs with AMD 2GHz Athlon processor and 2GB RAM, running FreeBSD 10.1 . Simulink graphical programming was used here along with [30] and Matdyn [31] while the simulation step was carefully chosen so as to realize the real-world process as accurately as possible.

AMICI framework facilitates the simulation of Programmable Logic Controllers (PLCs) including their interaction with software models by transforming model values to measured voltage [34] was employed in this work. More specifically, the messages from Remote Procedure Calls (RPC) are sent to Modbus and vice versa in the form of measured voltage. It should be emphasized that the implementation of client-side calls may offer realization of the interactions with other simulation units. The merit of the specific approach is that the data exchange is realized by an emulated network, thus closing the gap between real world conditions and the experimental platform.

4.2. Cyber-Attacks

The problem description considered in this paper assumes that the attacker has full access to the infrastructure, i.e., he may hijack node data according to his best interest, and thus perform a full-extent man-in-the-middle attack. In addition, the attacker may have the network under surveillance over an extended time interval so as to understand its dynamic behaviour and monitor any of the provided functionalities including component states, connectivity statuses, other type of measurements, etc.

During an attack, the control inputs are altered by the party implementing it, thus the information gathered by the controller is incorrect and the system becomes an open-loop, i.e., operator side control is lost. In such cases, the problem can only be tackled by prompt detection of the attack.

Based on the above described logic and the kinds of attacks considered in the literature (see Section 1) in this paper, we encompass a generic set of integrity attacks representing a wide range of scenarios. More specifically:

1. *Pulse*: In this case, the datastream is altered according to an additive pulse : $X_i^*(t) = X_i(t) + rect(t)$, where $X_i^*(t)$ is the compromised data, $X_i(t)$ the data coming from the nominal network state as recorded by the attacker, and

$$rect(t) = \begin{cases} 0, & \text{if } |t| > \frac{1}{2} \\ a_p/2, & \text{if } |t| = \frac{1}{2} \\ a_p, & \text{if } |t| < \frac{1}{2} \end{cases}$$

   where a_p is the attack parameter.
2. *Scaling*: Here the recorded measurements are scaled on the basis of the parameter a_s: $X_i^*(t) = a_s \times X_i(t)$.
3. *Ramp*: During this attack type the recorded measurements are gradually modified by adding a ramp function with parameter a_r: $X_i^*(t) = X_i(t) + ramp(t)$, where $ramp(t) = a_r \times t$.
4. *Random*: This type of attack suggests summing the recorded datastream with a uniform random distribution from the interval (a, b): $X_i^*(t) + rand(a, b)$.
5. *Replay*: The final type of attack which is usually found in the literature with the descriptive name *replay* merely involves the identical repetition of a priori recorded data.

An important note is that the attacker may compromise the network using the same or a different attack type in time, e.g., he may initialize with a *Replay* attack and continue with a *Random* one.

4.3. Parametrization and Results

Due to the nature of the proposed solution, the parametrization needs to take place on two sides: (a) each LTI model was of an autoregressive type (this choice can be conveniently altered thanks to the modularity of the proposed framework) while the order was the one with the highest reconstruction capabilities on a validation set V_S taken out of the {3, 4, 5, 6, 7} set; and (b) for the Gaussian modelling part, the number of components was chosen from the set $\{1, \ldots, 30\}$, where the step is 1. Torch framework [36] was employed for learning the model and during validation. The final number of Gaussian functions was selected after early experimentation while the upper bounds for k-means and EM iterations were 50 and 100, respectively, with a threshold of 0.001 between subsequent iterations.

Special care is needed during the parametrization of the integrity attacks since the attacker knows that the data belonging to the entire network are regularly passing quality checks (see, for example, [37]). As a consequence, the attack should be smart enough not to initialize the respective alarm type. The goal of the attacker is to have an impact on the overall network without compromising its stability in order to remain undetected. This facilitates information theft, usage of the network components according to his/her interests or any other action favouring the attacker. Based on this requirement, initial experimentations were conducted on the framework of Figure 3 for coming up with a feasible set of integrity attack parameters. The attacks were implemented using the following set of parameters: $a_s \in \{0.02, 0.03, 0.04, 0.05\}$, $a_r \in \{0.02, 0.03\}$, $a_p \in \{0.02, 0.04\}$ and random $a = 0.3, b = 0$. Replay attacks merely involve repetition of previously recorded data, thus there is no parameter to set.

The experimental data included a continuous operation of the system for 40,000 samples. The integrity attacks were affecting a randomly chosen bus at various time instances, i.e., at 10,000, 17,000, 23,000, 30,000, and 36,000. The attacks could be of the same or of a different type. The first 5000 samples coming from the nominal system state were employed for learning the probabilistic space shown in Figure 2. The validation set V_S was comprised of the 2000 samples which follow the training ones. The performance of the proposed system was assessed on the rest of the data. The final results are averaged across all types of integrity attacks. Here, it should be noted that each attack scenario, i.e., a case with a specific parameter was executed 50 times. A representative set of integrity attacks is shown in Figure 4.

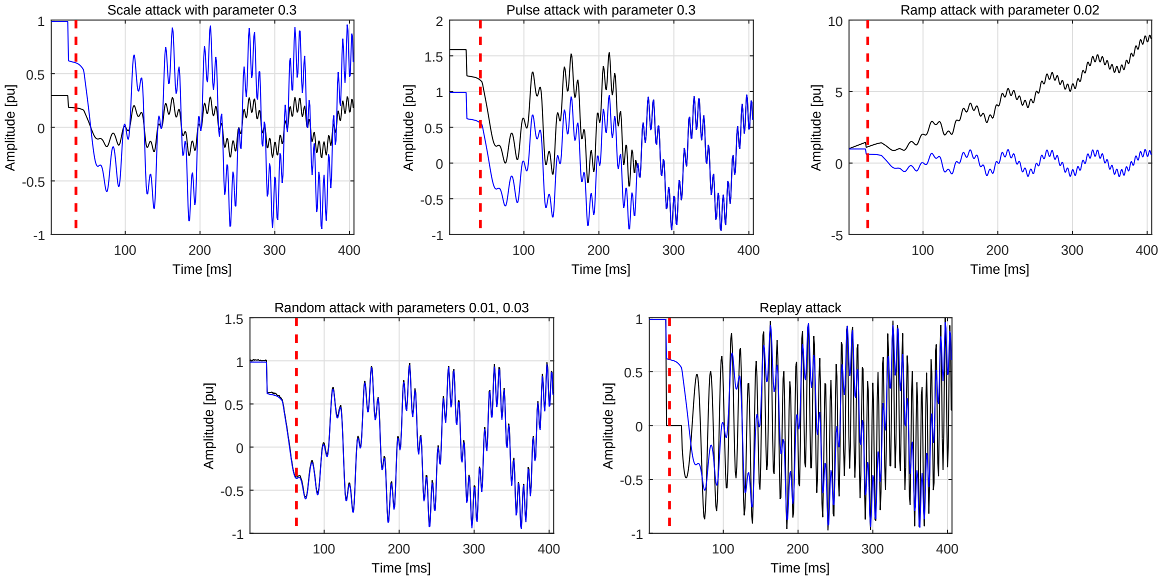


Figure 4. A set of the integrity attacks considered in this work. **Blue** coloured lines denote the nominal datastream while the ones affected by an attack are shown in **black**. The attacks start at $t = 1$ and the dashed vertical line denotes the time instant in which it is detected by the proposed algorithm. The detection delays are 34 ms, 42 ms, 26 ms, 63 ms, and 28 ms with respect to scale, pulse, ramp, random and replay attacks.

The proposed generative modelling approach was contrasted to a discriminative one, i.e., the support vector machine (SVM) technique, which is appropriate for binary classification problems. The fundamental assumption behind the creation of an SVM is that there exists a high-dimensional hyperplane able to separate the two classes. A one-class SVM was trained (using the implementation provided by [38]) on the nominal data using the radial basis function while the LTI parameters were scaled to $[-1, +1]$ in a linear way.

The SVM kernel function is a Gaussian radial basis

$$(k(x_i, x_j) = exp(\gamma||x_i - x_j||^2), \gamma > 0)$$

one, while the soft margin parameter and γ were determined though a grid search guided by cross-validation on the training set.

The LTI processing layer is based on linear ARX models with orders (2,2,1),and the resulting parameters are used for GMM and SVM learning. The orders were chosen as the ones providing the lowest reconstruction error on the validation set V_S.

Since we deal with a detection task [39], we employed the following three typical figures of merit to measure the capabilities of the proposed framework:

1. False positive (FP): it comprises the number of times that the proposed algorithm detects an integrity attack even though it is not present.
2. True positive (TP): it comprises the number of times that the proposed algorithm detects an integrity attack which is present.
3. False negative (FN): it comprises the number of times that the proposed algorithm does not detect an attack which is present.
4. True negative (TN): it comprises the number of times that the proposed algorithm does not detect an attack which is not present.
5. Detection Delay (DD): it measures the time delay for a correct detection by the proposed algorithm.

Comparative results are tabulated in Table 1 with respect to the proposed GMM clustering approach and the one employing SVM. They were trained and tested on identical datasets in order to achieve a fair comparison. As we can see, the proposed approach reaches a higher performance level than the SVM with respect to every type of integrity attack. It achieves lower FP and FN rates along with smaller DDs. Some representative detection examples of the proposed method are demonstrated in Figure 4. The complementary results of TP and TN rates confirm the superiority of the proposed approach over the contrasted one.

Table 1. The detection results of the proposed method and the contrasted one using SVM. The figures of merit are shown with respect to each type of integrity attack (see Section 4.2). The highest performances with respect to each attack are emboldened.

Integrity Attack	Model Type	FP (%)	TP (%)	FN (%)	TN (%)	Detection Delay (# of Samples)
Pulse	*Proposed*	**0.2**	**99.8**	**0.8**	**99.2**	**4**
	SVM	7.1	92.9	3.8	96.2	13
Scaling	*Proposed*	**0.4**	**99.6**	**1.2**	**98.8**	**2**
	SVM	3.5	96.5	6.2	93.8	19
Ramp	*Proposed*	**0.7**	**99.3**	**0.9**	**99.1**	**7**
	SVM	2.9	97.1	4.1	95.9	12
Random	*Proposed*	**1.3**	**98.7**	**1.8**	**98.2**	**6**
	SVM	3.7	96.3	5.9	94.1	8
Replay	*Proposed*	**1.2**	**98.8**	**0.4**	**99.6**	**3**
	SVM	8.8	91.2	4.5	95.5	16
Average	*Proposed*	**0.76**	**99.24**	**1.02**	**98.98**	**4.4**
	SVM	5.4	94.6	4.9	95.1	13.6

The attack which is detected faster by the proposed method is the *pulse* one, while the one with the highest latency is the *ramp* attack. We can observe that all the metrics are kept within very low values, which demonstrates the strength of the proposed approach, especially while considering the non linear characteristics of the specific application.

The merits of generative modelling are evident here since the distribution followed by the nominal state of the system is better captured and subsequently identified. The dependencies existing in the dataset are exploited by means of the parameter space where the differences between the normal operating conditions vs. operation under attack are highlighted. We can safely assume that the main goal of this work, i.e., very fast detection of integrity attacks, is achieved, at least to some extent.

5. Conclusions

This work analysed a novel method for detecting integrity attacks in cyber-physical infrastructures. The proposed framework exploits the correlations existing among the datastreams of the physical part of the network by means of LTI and GMM models. Each normal state comprises a point in the probabilistic space and the algorithm computes the affinity of novel data with the ones encountered during training by computing the respective KL divergences. One of the most important characteristics of the proposed algorithm is its flexibility, which is expressed in a twofold way: a) it can be applied practically unaltered to diverse types of ICT-controlled CIs, and b) it can be updated to deal with unseen types of nominal states by training new GMMs. The present framework is quite modular since new normal system states can be incorporated just by collecting the associated data, training a GMM and inserting it in the probabilistic space.

However, there are several issues which remain to be addressed in our future works. The main two research directions we wish to follow are:

- investigating how a data reconstruction technique might facilitate the controller of the network and substitute the node(s) under attack, and
- cooperating with an operator for applying the overall attack diagnosis framework on real-world data and understanding its limitations.

Acknowledgments: This work is supported by the Politecnico di Milano International Fellowship Program.

Author Contributions: S.N. conceived and designed the experiments; S.N. performed the experiments, analyzed the data, and contributed reagents/materials/analysis tools; S.N., and Y.S. wrote the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, Y.; Qian, Y.; Sharif, H.; Tipper, D. A Survey on Smart Grid Communication Infrastructures: Motivations, Requirements and Challenges. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 5–20.
2. Bao, H.; Lu, R. A New Differentially Private Data Aggregation With Fault Tolerance for Smart Grid Communications. *IEEE Internet Things J.* **2015**, *2*, 248–258.
3. Dau, S.H.; Song, W.; Yuen, C. On Simple Multiple Access Networks. *IEEE J. Sel. Areas Commun.* **2015**, *33*, 236–249.
4. Ten, C.W.; Liu, C.C.; Manimaran, G. Vulnerability Assessment of Cybersecurity for SCADA Systems. *IEEE Trans. Power Syst.* **2008**, *23*, 1836–1846.
5. Ten, C.W.; Manimaran, G.; Liu, C.C. Cybersecurity for Critical Infrastructures: Attack and Defense Modeling. *IEEE Trans. Syst. Man Cybern. Part A* **2010**, *40*, 853–865.
6. Langner, R. Stuxnet: Dissecting a Cyberwarfare Weapon. *IEEE Secur. Priv.* **2011**, *9*, 49–51.
7. Sridhar, S.; Hahn, A.; Govindarasu, M. Cyber-Physical System Security for the Electric Power Grid. *Proc. IEEE* **2012**, *100*, 210–224.
8. Hackers Allegedly Breached Saudi Aramco Again. Available online: http://www.net-security.org/secworld.php?id=13493 (accessed on 21 November 2016).
9. 1.5 million Cards Compromised in Global Payments Breach. Available online: http://www.net-security.org/secworld.php?id=12680 (accessed on 21 November 2016).
10. Hackers Breach U.S. Energy Department Networks. Available online: http://www.net-security.org/secworld.php?id=14353 (accessed on 21 November 2016).
11. One of America's Premier Research Institutions Was Hacked—and the Signs Point to China. Available online: http://qz.com/526287/one-of-americas-premier-research-institutions-was-hacked-and-the-signs-point-to-china/ (accessed on 21 November 2016).
12. Soupionis, Y.; Ntalampiras, S.; Giannopoulos, G. Faults and Cyber Attacks Detection in Critical Infrastructures. In *Critical Information Infrastructures Security, Proceedings of the 9th International Conference (CRITIS 2014), Limassol, Cyprus, 13–15 October 2014*; Panayiotou, C.G., Ellinas, G., Kyriakides, E., Polycarpou, M.M., Eds.; Springer: Cham, Switzerland, 2016; Revised Selected Papers; pp. 283–289.

13. Zhengbing, H.; Zhitang, L.; Junqi, W. A Novel Network Intrusion Detection System (NIDS) Based on Signatures Search of Data Mining. In Proceedings of the 1st International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), Adelaide, Australia, 23–24 January 2008; pp. 10–16.
14. Mo, Y.; Chabukswar, R.; Sinopoli, B. Detecting Integrity Attacks on SCADA Systems. *IEEE Trans. Control Syst. Technol.* **2014**, *22*, 1396–1407.
15. Su, S.; Duan, X.; Zeng, X.; Chan, W.; Li, K.K. Context Information based Cyber Security Defense of Protection System. In Proceedings of the 2007 IEEE Power Engineering Society General Meeting, Tampa, FL, USA, 24–28 June 2007.
16. Coutinho, M.P.; Lambert-Torres, G.; da Silva, L.E.B.; Martins, H.G.; Lazarek, H.; Neto, J.C. Anomaly detection in power system control center critical infrastructures using rough classification algorithm. In Proceedings of the 3rd IEEE International Conference on Digital Ecosystems and Technologies (DEST '09), Istanbul, Turkey, 1–3 June 2009; pp. 733–738.
17. Bigham, J.; Gamez, D.; Lu, N. Safeguarding SCADA Systems with Anomaly Detection. In *MMM-ACNS, Lecture Notes in Computer Science*; Gorodetsky, V., Popyack, L.J., Skormin, V.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2776, pp. 171–182.
18. Li, W.T.; Wen, C.K.; Chen, J.C.; Wong, K.K.; Teng, J.H.; Yuen, C. Location Identification of Power Line Outages Using PMU Measurements With Bad Data. *IEEE Trans. Power Syst.* **2016**, *31*, 3624–3635.
19. Sun, Y.; Li, W.T.; Song, W.; Yuen, C. False data injection attacks with local topology information against linear state estimation. In Proceedings of the 2015 IEEE Innovative Smart Grid Technologies—Asia (ISGT ASIA), Bangkok, Thailand, 3–6 November 2015; pp. 1–5.
20. Singh, S.; Silakari, S. An Ensemble Approach for Cyber Attack Detection System: A Generic Framework. In Proceedings of the 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Honolulu, HI, USA, 1–3 July 2013; pp. 79–84.
21. Ntalampiras, S. Detection of Integrity Attacks in Cyber-Physical Critical Infrastructures Using Ensemble Modeling. *IEEE Trans. Ind. Inform.* **2015**, *11*, 104–111.
22. Ntalampiras, S. Fault Identification in Distributed Sensor Networks Based on Universal Probabilistic Modeling. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *26*, 1939–1949.
23. McLachlan, G.; Basford, K. (Eds.) *Mixture Models: Inference and Applications to Clustering*; Marcel Dekker: New York, NY, USA, 1988.
24. Ljung, L. Convergence analysis of parametric identification methods. *IEEE Trans. Autom. Control* **1978**, *23*, 770–783.
25. Bonissone, P.P.; Xue, F.; Subbu, R. Fast meta-models for local fusion of multiple predictive models. *Appl. Soft Comput.* **2011**, *11*, 1529–1539.
26. Alippi, C.; Ntalampiras, S.; Roveri, M. An HMM-based change detection method for intelligent embedded sensors. In Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, Australia, 10–15 June 2012; pp. 1–7.
27. Reynolds, D.A.; Rose, R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 72–83.
28. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions. *IEEE Trans. Multimed.* **2011**, *13*, 713–719.
29. Aucouturier, J.-J.; Defreville, B.; Pachet, F. The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J. Acoust. Soc. Am.* **2007**, *122*, 881–891.
30. Zimmerman, R.D.; Murillo-Sánchez, C.E.; Thomas, R.J. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Trans. Power Syst.* **2011**, *26*, 12–19.
31. Cole, S.; Belmans, R. MatDyn, a new Matlab-based toolbox for power system dynamic simulation. *IEEE Trans. Power Syst.* **2011**, *26*, 1129–1136.
32. Power Systems Test Case Archive. Available online: http://www.ee.washington.edu/research/pstca/ (accessed on 21 November 2016).
33. Genge, B.; Siaterlis, C.; Hohenadel, M. AMICI: An Assessment Platform for Multi-domain Security Experimentation on Critical Infrastructures. In *LNCS Critical Information Infrastructures Security*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7722, pp. 228–239.
34. Siaterlis, C.; Genge, B.; Hohenadel, M. EPIC: A Testbed for Scientifically Rigorous Cyber-Physical Security Experimentation. *IEEE Trans. Emerg. Top. Comput.* **2013**, *1*, 319–330.

35. White, B. An Integrated Experimental Environment for Distributed Systems and Networks. In Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI 02), Boston, MA, USA, 9–11 December 2002; pp. 255–270.
36. Torch. Available online: http://www.torch.ch (accessed on 21 November 2016).
37. *MRO Under-Frequency Load Shedding (UFLS) Program Midwest Reliability Organization*; Technical Report; Midwest Reliability Organization: St. Paul, MN, USA, 2005.
38. Chang, C.C.; Lin, C.J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27, doi:10.1145/1961189.1961199.
39. Alippi, C.; Ntalampiras, S.; Roveri, M. A Cognitive Fault Diagnosis System for Distributed Sensor Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *24*, 1213–1226.