OPEN ACCESS

2nd International Electronic Conference on Entropy and Its Applications
15-30 November 2015

www.sciforum.net/conference/ecea-2

Conference Proceedings Paper – Entropy

Detection of Integrity Attacks in Cyper Physical Systems Based On Reservoir Networks

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

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32 I-20133, Milano, Italy

[2] European Commission, Joint Research Center, IPSC, Via E. Fermi, 2749, 21027 Ispra (VA), Italy

* Author to whom correspondence should be addressed; E-Mails: stavros.ntalampiras@polimi.it, dalaouzos@gmail.com, https://sites.google.com/site/stavrosntalampiras/home.

Published: 13 November 2015

Abstract: This paper presents an anomaly-based methodology for reliable detection of integrity attacks in cyber-physical critical infrastructures. Such malicious events compromise the smooth operation of the infrastructure while the attacker is able to exploit the respective resources according to his/her purposes. Even though the operator may not understand the attack, since the overall system appears to remain in a steady state, the consequences may be of catastrophic nature with a huge negative impact. Here, we apply a deep learning technique and more specifically reservoir networks. They follow the supervised learning principle for recurrent neural networks, while the fundamental logic is to steer a random, large, fixed recurrent neural network with the input signal to the desired direction (class, probability, etc.). Their great advantage is the fact that the only part in need of training is the output layer which is a linear combination of all of the response signals. The experimental platform includes a simulator of both a power grid and a cyber-network of the IEEE-9 bus model. Subsequently we implemented a wide range of integrity attacks (replay, ramp, pulse, scaling, and random) with different intensity levels. A thorough evaluation procedure is carried out while the results demonstrate the detection capabilities of the proposed method in terms of false positive rate, false negative rate and detection delay.

Keywords: Cyber-physical systems; critical infrastructures; reservoir modeling; fault diagnosis; cyber-attacks

---

1. Introduction

Recently we have observed that Information and Communication Technologies (ICT) play a significant role in monitoring and controlling large scale infrastructures. Critical Infrastructures (CI) are the assets on which the smooth functioning of a society, economy, etc. depends. Some examples are electricity network including generation, transmission and distribution, gas network for production, transport and distribution, financial services (banking, clearing), transportation systems (fuel supply, railway network, airports, harbours, inland shipping), etc. The usage of an ICT layer offers improved exploitation of the CI and increases the performance of the system while reducing the overall cost. The main objectives of such a control structure are: (1) to maintain safe operational goals by limiting the probability of undesirable behavior, (2) to meet the production demands by keeping certain process values within prescribed limits, and (3) to maximize production profit.

A typical example is the Smart Grid (SG) which essentially is a standard electrical network operated via an ICT layer. This design offers the ability to obtain a complete picture of the overall network in terms of the manner in which the assets of the SG are used, the existing demand along with their evolution in time. Thus the operator is able to make informed decisions regarding the future usage of the SG, e.g. steering the power production from generators working close to their limits to lightly used generators. However the cyber-layer opened the door to threats which were previously nonexistent in practice, i.e. cyber-attacks.

In general, any offensive tactic/action against information systems, infrastructures, computer networks conducted by individuals or whole organizations can be considered as a cyber-attack. The purpose of this malicious act can vary and usually includes stealing, falsification, or even destruction of information and/or system components [1,2]. Integrity attacks are the ones which alter the data transmitted within the infrastructure serving the attacker's purposes while remaining undetected. It is achieved in three steps: a) compromising one or more communication buses of the cyber physical system, b) long-term monitoring of the involved datastreams, and c) designing a technique which alters the pattern of the transmitted data while at the same time, it is perceived as normal by the controller.

The specific problematic shares some common characteristics with that of intrusion detection in data packets exchanged among computer systems. In this context, the ultimate aim is to automatically assess data integrity while having as available information only the data content. There are three lines of thought for addressing this problem: *a*) *signature-based*: here the algorithm searches for known patterns of malicious activity in the datastream using a predefined dictionary of attacks (e.g. [3]), *b*) *anomaly-based*: this type of approaches estimates characteristic features of the normal behavior and subsequently detects deviations which may appear during an intrusion (e.g. [4,5]), and *c*) *countermeasure-based*: the methodologies which belong to this category adapt the involved signals so that the task of intrusion detection is simplified (e.g. [6,7]).

This paper presents a method for detecting integrity attacks occurring on SGs based on deep learning and more specifically reservoir networks (RN). The RN learns the temporal relationships of the involved datastreams with respect to each node during nominal conditions. Subsequently it is evaluated while the SG is operating using novel data. When we observe a deviation larger than one observed during the nominal conditions the algorithm detects an integrity attack. It is important to note that the proposed

system does not make any assumption regarding the distribution of the dataset while the types of attacks do not need to be known a-priori. Finally the method is evaluated under a wide range of integrity attacks.

2. The Reservoir Network

2.1. Association with Deep Learning

Reservoir Networks (RNs) represent a novel kind of echo-state networks providing good results in several demanding applications, such speech recognition [8], saving energy in wireless communication [9], etc. RNs are able to capture non-linear relationships existing within data coming from various nodes. An RN, the topology of which is depicted in Fig. 1, includes neurons with non-linear activation functions which are connected to the inputs (input connections) and to each other (recurrent connections). These two types of connections have randomly generated weights, which are kept fixed during both the training and operational phase. Finally, a linear function is associated with each output node.

Recurrent neural networks comprise a deep learning architecture as their main purpose is to capture the characteristics of high-level abstractions existing in the acquired data while designing multiple processing layers of complicated formations, i.e. non-linear functions. The advantage of RN is that the calculations involved in its readout layer are linear, thus of limited computational complexity and relatively small duration of the training process. Reservoir computing argues that since back propagation is computationally complex but typically does not influence the internal layers severely, it may be totally excluded from the training process. On the contrary, the readout layer is a generalized linear classification/regression problem associated with low complexity. In addition any potential network instability is avoided by enforcing a simple constraint on the random parameters of the internal layers.

2.2. Training

The RN parameters are the weights of the output connections and are trained to achieve a specific result, e.g. that a given output node produces high values for observations of a particular class. The output weights are learned by means of linear regression and are called read-outs since they "read" the reservoir state. More details about the RN training and the echo state property can be found in [10].

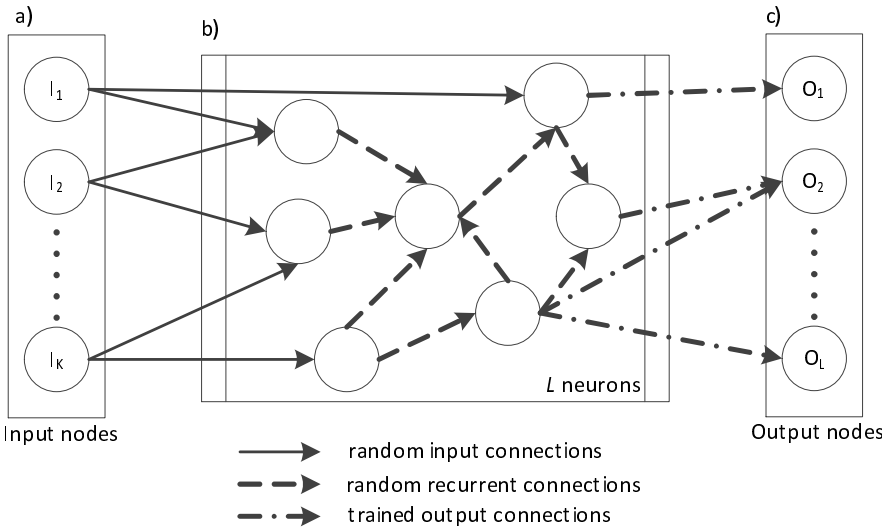
As a general formulation of the RNs, we assume that the network has K inputs, L neurons (usually called reservoir size), M outputs (O_L in Figure 1) while the matrices $W_{in}(K \times L)$, $W_{res}(L \times L)$ and $W_{out}(L \times M)$ include the connection weights. The RN system equations are the following:

$$x(k) = f_{res}(W_{in}u(k-1) + W_{res}x(k-1)) \quad (1)$$

$$y(k) = f_{out}(W_{out})x(k), \quad (2)$$

where $u(k)$, $x(k)$ and $y(k)$ denote the values of the inputs, reservoir outputs and the read-out nodes at time k respectively. f_{res} and f_{out} are the activation functions of the reservoir and the output nodes, respectively. In this work we consider $f_{res}(x) = tanh(x)$ and $f_{out}(x) = x$ and we fix $M = 1$ since we are considering single-output models.

Figure 1. A standard reservoir network consisting of three layers: a) the input, b) the reservoir and c) the readout. The second layer includes neurons with non-linear activation functions. The weights of the input and the recurrent connections are randomly fixed. The weights to the output nodes are the only ones being trained.



Linear regression is used to determine the weights W_{out},

$$W_{out} = \underset{W}{argmin}(\frac{1}{N_{tr}}\|XW - D\|^2 + \epsilon\|W\|^2) \quad (3)$$

$$W_{out} = (X^T X + \epsilon I)^{-1}(X^T D), \quad (4)$$

where XW and D are the computed vectors, I a unity matrix, N_{tr} the number of the training samples while ϵ is a regularization term.

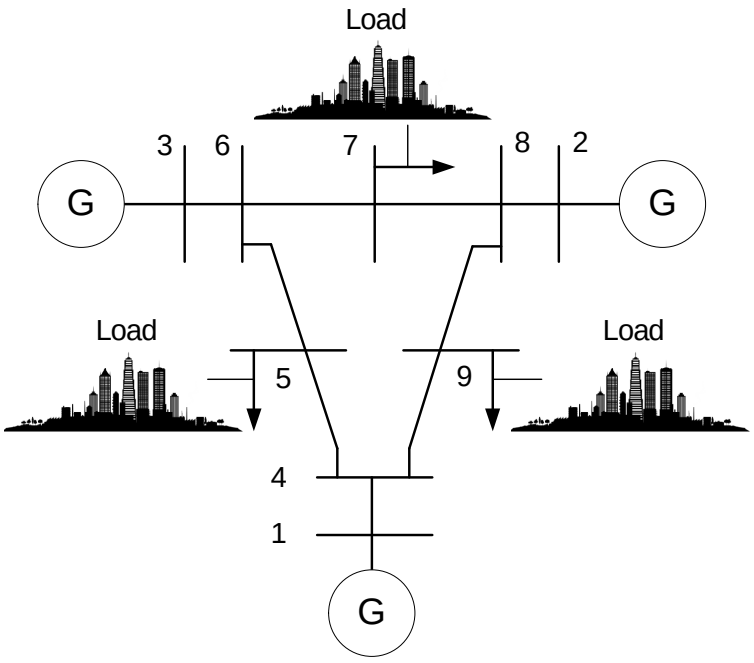
The recurrent weights are randomly generated by a zero-mean Gaussian distribution with variance v, which essentially controls the spectral radius SR of the reservoir. The largest absolute eigenvalue of W_{res} is proportional to v and is particularly important for the dynamical behavior of the reservoir [11]. W_{in} is randomly drawn from a uniform distribution $[-InputScalingFactor, +InputScalingFactor]$, which emphasises/deemphasises the inputs in the activation of the reservoir neurons. It is interesting to note that the significance of the specific parameter is decreased as the reservoir size increases.

To model the temporal redundancy, (1) can be reformulated by substituting $y(k)$ with $X_i(k)$ and $u(k-1)$ with $[X_i(k-1), \ldots, X_i(k-n_i)]$, where $X_i(k)$ denotes the voltage value of bus i at time instant k.

3. Experimental Set-Up and Results

This section describes the experimental set-up and results obtained by applying the proposed RN-based system onto the smart-grid application domain. In this work we employed the IEEE-9 bus model which includes nine buses, three generators and three loads. It is a widely used model which is usually employed in the literature [12–14]. The respective diagram is shown in Fig. 2. The simulation of the system operation was realized by MATPOWER [15] and Matdyn [16], which include the algorithms written in MATLAB for performing dynamic analysis of electric power systems.

Figure 2. The IEEE-9 bus model.



The information $X_i, i = 1 \dots 9$ considered in this work is the voltage of each bus of the network. 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.

The parameters of the RN were selected by means of exhaustive search. They were taken from the following sets: $SR \in \{0.8, 0.9, 0.95, 0.99\}$, $L \in \{100, 500, 1000, 5000, 10000\}$, and $InputScalingFactor \in \{0.1, 0.5, 0.7, 0.95, 0.99\}$. The implementation of the RN was based on the echo state network toolbox which is available at http://reservoir-computing.org/software.

The integrity attacks parameters should be chosen so that the attack has an impact on the system but at the same time the overall network remains stable. We used the following parameters to realize the integrity attacks: $a_s \in \{0.02, 0.03, 0.04, 0.05\}$, $a_r \in \{0.02, 0.03\}$, $a_p \in \{0.02, 0.04\}$ and random $a = 0.3, b = 0$.

Each scenario had a duration of 12000 samples while the attack was injected at sample 9000. The first 4000 samples were used for training the model, the following 4000 for validation and the rest were used for testing. The results (averaged over all types of integrity attacks) are:

1. False positive (FP): **6.1%** (it counts the times the algorithm detects an attack in the datastream when there is not one).
2. False negative (FN): **5.9%** (it counts the times the algorithm does not detect an attack while there is one).
3. Detection Delay (DD): **28 samples** (it measures the time delay that is needed by the algorithm to detect an attack).

The produced results are quite encouraging and characterized by low values of FP, FN and DD. The RN is able to capture the dynamics existing in the time evolution of the datastreams and detect any abnormality induced by an integrity attack. We conclude that reservoir modeling can be effectively applied on the problem of integrity attack detection in smart grids. In addition, the proposed method can be potentially applied to various cyber-physical systems.

4. Conclusions

This paper presented a technique to accurately detect integrity attacks, i.e. a special form of cyber-attacks where the attacker alters the datastreams of the infrastructure. The method was applied on simulated data coming from the IEEE-9 bus model while its datastreams are affected by various kinds of attacks. The proposed method is based on a reservoir network which is able to model the temporal evolution of the datastreams and promptly detect a potential attack. Our future work includes the application of the RN-based framework on other types of cyber-physical systems and cooperate with an operator in order to evaluate the method on data coming from a real infrastructure in order to understand and address its limitations.

Acknowledgements

This work was supported by the Politecnico di Milano International Fellowship Program.

References

1. Hashimoto, H.; and Hayakawa, T. Distributed cyber attack detection for power network systems. *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*; 2011; pp. 5820-5824.
2. Monti, A.; and Ponci ,F. Electric power systems. In *Intelligent Monitoring, Control, and Security of Critical Infrastructure Systems; Kyriakides, E., Polycarpou, M., 2015; pp. 31âĂŞ65.*

3. *Hu Zhengbing and Li Zhitang and Junqi, W. A Novel Network Intrusion Detection System (NIDS) Based on Signatures Search of Data Mining. In* Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on, 2008; pp. 10-16.
4. Coutinho et al., M.P. Anomaly detection in power system control center critical infrastructures using rough classification algorithm. In *Digital Ecosystems and Technologies, 2009. DEST '09. 3rd IEEE International Conference on, 2009; pp. 733-738.*
5. *Neuzil, J. and Kreibich, O. and Smid, R. A Distributed Fault Detection System Based on IWSN for Machine Condition Monitoring.* Industrial Informatics, IEEE Trans. on **2014**, *10*, 1118-1123.
6. Mo, Y. and Chabukswar, R. and Sinopoli, B. Detecting Integrity Attacks on SCADA Systems. *Control Systems Technology, IEEE Transactions on,* **2014***,* 22*, pp. 1396-1407.*
7. *Giani, A. and Bitar, E. and Garcia, M. and McQueen, M. and Khargonekar, P. and Poolla, K. Smart Grid Data Integrity Attacks.* Smart Grid, IEEE Transactions on **2013**, *4*, pp. 1244-1253.
8. Verstraeten, D. and Schrauwen, B. and Stroobandt, D. Reservoir-based techniques for speech recognition. *Neural Networks, 2006. IJCNN '06. International Joint Conference on, 2006; pp. 1050-1053.*
9. *Jaeger, H. and Haas, H.. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication.* Science **2004**, *304*, pp. 78-80.
10. Mantas L. and Jaeger, H. Reservoir computing approaches to recurrent neural network training. *Computer Science Review* **2009***,* 3*, pp. 127-149.*
11. *D. Verstraeten and B. Schrauwen and M. DÃćâĆňâĎćHaene and D. Stroobandt. An experimental unification of reservoir computing methods.* Neural Networks **2007**, *20*, pp. 391-403.
12. Manandhar, K., Xiaojun C., Fei H., and Yao L. Detection of Faults and Attacks Including False Data Injection Attack in Smart Grid Using Kalman Filter. *Control of Network Systems, IEEE Transactions on* **2014***,* 1*, pp. 370-379.*
13. *Saluja, R. and Ghosh, S. and Ali, M.H. Transient stability enhancement of multi-machine power system by novel braking resistor models.* Southeastcon, 2013 Proceedings of IEEE, 2013; pp. 1-6.
14. Zhang, M., Guo, Q., Sun, H., Zhang, B. A sensitivity based simplified model for security constrained optimal power flow. *Innovative Smart Grid Technologies - Asia (ISGT Asia), 2012 IEEE, 2012; 1-4.*
15. *Zimmerman, Ray D. and Murillo S., Carlos E, Robert J. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education.* Power Systems, IEEE Transactions on **2011**, *26*, pp. 12-19.
16. Cole, S., and Belmans, R. MatDyn, a new Matlab-based toolbox for power system dynamic simulation. *Power Systems, IEEE Transactions on* **2011***,* 3*, pp. 1129-1136.*