Expert Systems With Applications 58 (2016) 164–173

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automatic identification of integrity attacks in cyber-physical systems

Stavros Ntalampiras*

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy

ARTICLE INFO

Article history:
Received 23 December 2015
Revised 4 April 2016
Accepted 5 April 2016
Available online 9 April 2016

Keywords:
Critical infrastructure protection
Fault diagnosis
Cyber security
Cyber-physical systems
Probabilistic modelling
Deep learning

ABSTRACT

Modern society relies on the availability and smooth operation of complex engineering systems, such as electric power systems, water distributions networks, etc. which due to the recent advancements in information and communication technologies (ICT) are usually controlled by means of a cyber-layer. This design may potentially improve the usage of the components of the cyber-physical system (CPS), however further protection is needed due to the emerging threat of cyber-attacks. These may degrade the quality of the communicated information which is of fundamental importance in the decision making process.

This paper proposes a novel methodology for automatic identification of the type of the integrity attack affecting a CPS. We designed a feature set for capturing the characteristics of each attack in the spectral and wavelet domains while its distribution is learned by pattern recognition algorithms of different modelling properties customized for the specific application scenario. In addition a novelty detection component is incorporated for dealing with previously unseen types of attacks. The proposed approach is applied onto data coming from the IEEE-9 bus model achieving promising identification performance.

1. Introduction

Smart Grids (SG) represent a new generation of electrical grids encompassing information and communication Technologies (ICTs) (Zhaoyang, 2014). The main characteristic of SGs is the ability to exploit electricity and information flows (from power stations to consumers and back) aiming at providing improved delivery of energy and quality of service. In more detail, the joint use of electrical grids and ICTs is motivated by the need for exploiting the assets of the network in a more effective way. The ultimate goal is to take advantage of their full potential by governing the electrical network in an informed manner, i.e. activating, deactivating, or managing differently specific assets which may lead to more cost-effective strategies.

Being part of Critical Infrastructures Systems, such a new generation of electrical grids is heavily interconnected with other critical infrastructures such as water transport/distribution networks, communication networks, gas/oil transport and distribution networks and traffic transport network. Hence, the ability of SGs to guarantee the quality-of-service is crucial to all dependent systems.

In addition SGs wish to integrate high-speed, bidirectional communication technologies for acquiring new energy management capabilities, such as advanced metering infrastructure (AMI) (Zhou, Guo, & Qin, 2010), automatic demand response, (LeMay, Nelli, Gross, & Gunter, 2008; Wang & Lu, 2013), increased data storage (Sancho-Asensio et al., 2014) etc. Even though such a tight coupling with information technologies brings valuable additional functionalities, it is accompanied with an increased risk of intrusions via exploiting potential vulnerabilities of the cyber layer which may have a catastrophic impact on the operation of the SG, such as extended blackouts, loss of one or more infrastructure assets, etc. (Anwar & Mahmood, 2014; Metke & Ekl, 2010). There have been several recent cyber-attacks with very unfortunate consequences: a) a computer worm named Stuxnet was discovered in 2010 infiltrating Windows OSs for affecting Siemens industrial software and destabilising power systems (Karnouskos, 2011), b) in 2003 a large-scale blackout affected portions of the Midwest and Northeast United States and Ontario, Canada due to software failure of the cyber layer controlling the electric network[1], c) a major electricity blackout was suffered in Italy and some parts of Switzerland, in 2003 due to a human error and ineffective communication between the power grid operators (Berizzi, 2004), and d) communication issues and a human error led to another blackout in the

* Tel.: +393888344104.
E-mail address: stavros.ntalampiras@polimi.it, dalaouzos@gmail.com
URL: https://sites.google.com/site/stavrosntalampiras/home

[1] U.S.-Canada Power System Outage Task Force, "Final Report on the August 14, 2003" Blackout in the United States and Canada: Causes and Recommendations, April 2004, available at http://energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf

http://dx.doi.org/10.1016/j.eswa.2016.04.006
0957-4174/

south west Europe in 2006[2]. In general, as the 2011 annual report of the Repository for Industrial Security Incidents states, around 35% of industrial control systems (ICS) security incidents were realized by exploiting vulnerabilities existing in the cyber layer[3].

Thus there is an immediate need to develop systems able to diagnose such attacks for promptly informing the authorized personnel, and provide a decision support to the infrastructure operator. Integrity attack diagnosis systems (IADS) encompass *detection, isolation, identification*, and *provision* of suitable accommodation actions. Even though it is a relatively new research area, it has attracted the interest of the scientific community with most of the researchers have focussed on detection and isolation (Bi, Shen, Plans, and Zhao, 2013; Gulisano et al., 2015; He & Zhang, 2010). Unlike fault identification in sensor networks where the patterns of the involved signals change substantially over time (Ntalampiras, 2015b), integrity attacks are harder to recognize due to the usually minor effect(s) they have on the datastreams of the CPS. Another challenge for IADSs is generality, i.e., the fact that they should be flexible enough to be applicable to different SGs in terms of measurements, topology, etc., with only slight modifications in their functionality, and the need to deal with time-varying load demands that may change in an unpredictable manner (hence inducing a further level of uncertainty in the IADS activity). In addition the problem of integrity attack diagnosis is becoming more interesting due to the availability of information (typically the index of a SG includes the voltage phasor, frequency, and rotor angle measurements) which enables improved control of the infrastructure.

This article proposes a machine learning based framework for identifying integrity attacks occurring in SGs (but the methodology is generic enough to be applied to CPSs in general) as an extension to the previously proposed detection framework (Ntalampiras, 2015a). The overall aim is to facilitate the human experts into taking appropriate actions regarding the accommodation and future usage of the CPS, while having available information regarding the type of the detected integrity attack. The IADS a) gathers the data produced by the power grid, b) extracts features coming from the frequency and wavelet domains, c) performs novelty detection, and d) employs diverse classification techniques to identify the integrity attack. An interesting advantage of the proposed IADS is the absence of an analytical model explaining the process generating the data. This is a major advantage of the present IADS since assuming the existence of an accurate model is a) costly in terms of time and human resources given that a domain expert should be summoned, and b) unrealistic to provide reliable measurements under every operational scenario, e.g. adverse environmental conditions. Finally, we used a simulator of the IEEE-9 bus system for our experiments, where we implemented a gamut of integrity attacks (*ramp, pulse, random, and scaling*) and we thoroughly assessed the performance of the proposed IADS. It should be noted that we consider variations in the load as part of the nominal state.

2. Related literature

Lately, the interest of the scientific community in the area of cyber-attack diagnosis has been significantly increased. This comes as a natural outcome of the fact that everyday, the number of infrastructures controlled by ICT networks is consistently growing due to the advantages it offers (see Section 1). However, most of the works focus solely on the detection of such attacks, (e.g. Almalawi, Yu, Tari, Fahad, and Khalil (2014); Elhag, Fernndez, Bawakid, Alshomrani, and Herrera (2015); Koc, Mazzuchi, and Sarkani (2012); Mo, Chabukswar, and Sinopoli (2014); Ntalampiras (2015a); Shin, Lee, Kim, and Kim (2013). In this work we assume that the detection has already been performed by an appropriate methodology, while the task of identification still remains unaddressed. The specific task bears some similarities with identifying faults in sensor networks (Ntalampiras, 2015b), thus the logic behind the design could be parallel. Nonetheless, one should keep in mind that critical infrastructures operated by an ICT layer comprise systems (or systems of systems) which are large-scale, thus it is particularly difficult to derive the analytical expressions explaining the relationships between different nodes of the network. In addition the problem may potentially be of higher degree of difficulty as the attacks are designed so as to have a minimum effect on the communicated information. Thus one should turn to machine learning techniques to face the problem as they do not require domain specific knowledge which is an important advantage as sometimes obtaining this knowledge is time/resource costly since an expert must be consulted. This work customizes a big gamut of diverse machine learning techniques for identifying integrity attacks affecting cyber-physical systems. We aim to thoroughly assess their performance, understand the merits and weaknesses of each one and potentially design an effective fusion scheme. Unlike papers which assume analytical models with respect to the system under consideration or the data injection process (Deng, Xiao, & Lu, 2015), the proposed framework does not make any similar hypotheses which comprise a significant advantage since it is unrealistic to assume that the model of the system exists (especially for large scale systems), is available and/or is accurate under real-world adverse conditions.

Next, we provide only a brief but representative glimpse of the related work. Paper (Caro, Conejo, Minguez, Zima, & Andersson, 2011) employs statistical correlations among measurements as input to an identification algorithm based on the largest normalized residual test for identifying bad data. Two custom systems with 4 and 39 buses were considered for the experimental evaluation. An algorithm which considers the information from both the active and the reactive power measurements is explained in Sou, Sandberg, and Johansson (2012). Potential anomalies in the measurements are detected by thresholding the residual on the analytical model of the system. The experiment includes a stealth attack targeting the active power flow measurement between bus 2 and bus 5 on IEEE 14-bus benchmark system.

A method for power system static state estimation for bad data detection and identification is presented in Lin and Pan (2007). The method is based on a gap statistic for clustering is used to compute the optimal number of clusters which is an indication of bad data presence. The effectiveness of the method is investigated on a custom 30-bus power system. An approach which first decomposes the system into extended subsystem and then uses chi-square tests is designed in Wang et al. (2014). Three attack cases were simulated in the IEEE 14-, 39-, 118- and 300-bus systems. They are very specific while they affect limited parts of the network and individual measurements in a linear predefined way.

An artificial neural network is proposed in Wu, Onwuachumba, and Musavi (2013) to estimate the power system state and identify manipulated data. Experiments are conducted on GE 6-bus and IEEE 14-bus power systems. Principal components analysis is used in ,Melndez, Herraiz, Prieto, and Bravo (2011) to track the raw measurements, detect and identify the principal forms of gross and historic errors. Two scenarios were designed for the experiments, i.e. the IEEE 42-bus system and a High Volt/Medium Volt substation while the affected measurements were specific power and voltage values.

[2] UCTE, "System Disturbance on 4 November 2006", available at https://www.entsoe.eu/fileadmin/user_upload/_library/publications/ce/otherreports/Final-Report-20070130.pdf

[3] http://www.risidata.com/