Open access • Journal Article • DOI:10.1109/TCST.2013.2280899

Detecting Integrity Attacks on SCADA Systems — Source link

Yilin Mo, Rohan Chabukswar, Bruno Sinopoli

Institutions: California Institute of Technology, Carnegie Mellon University

Published on: 01 Jul 2014 - IEEE Transactions on Control Systems and Technology (IEEE)

Topics: SCADA and Replay attack

Proceedings of the 18th World Congress
The International Federation of Automatic Control
Milano (Italy) August 28 - September 2, 2011

Detecting Integrity Attacks on SCADA Systems

Rohan Chabukswar, Yilin Mo, Bruno Sinopoli

Department of Electrical and Computer Engineering, Carnegie Mellon University.

Abstract: Ensuring security of systems based on supervisory control and data acquisition (SCADA) is a major challenge. In this paper we analyze the effect of integrity attacks on control systems and provide countermeasure capable of exposing such attacks. To validate the results, we apply our findings to an industrial control problem concerning a chemical plant and then a simplified model of a power system.

Keywords: Cyber-Physical Systems, SCADA, Secure, Control

1. INTRODUCTION

Cyber-physical systems (CPS) are systems with a tight co-ordination between its computational and physical elements. Such systems often employ a distributed network of embedded sensors and actuators that interact with the physical environment, monitored and controlled by a Supervisory Control and Data Acquisition (SCADA) system. In daily life, cyber-physical systems can be seen in multifarious applications like smart grids, process control systems, air traffic control, medical monitoring, and so on.

A recent concern in distributed control system security is that an attacker could gain remote access to a large set of sensing and actuation devices and modify their software or their environment to launch a coordinated attack against the system infrastructure. An example of alleged digital warfare is the Stuxnet worm, which seems to have been specially designed to reprogram certain industrial centrifuges and make them fail in a way that was virtually undetectable (Markoff (2010)). Speculations and allegations have flown back and forth, accusing various national intelligence agencies and even the manufacturer, but irrespective of the attacker, target or intention, this worm has indubitably brought to light serious security susceptibilities in industrial control systems. In view of the present threat of global terrorism, a power grid failure, a local breakdown of telecommunications system, a disruption of air traffic control (ATC) at a major hub, could all be executed as an antecedent to a full-fledged invasion. Such threats have been predicted (Carlin (1997)) and even made into movies. CPS infrastructures like power grids, telecommunication networks, ATCs — vital to the normal operation of a society — are safety critical, and a successful attack on one of them, or worse, a co-ordinated attack on two or more of them, can significantly hamper the economy, endanger human lives or even make the community vulnerable to foreign aggression. This makes the design of secure cyber-physical systems of paramount importance.

A conventional method of enforcing security is applying cryptographic principles to the communication network. While this approach might be sufficient for day-to-day usage, in cases of national security a more robust security mechanism is called for. Cryptographic keys can be broken or stolen, or the attacker could directly attack the physical elements of the system, without even hijacking the communication. For example, the attacker could use heaters and coolers to control what a thermometer senses. Such an attack is feasible when sensors and actuators are spatially distributed in remote locations. Consequently, system knowledge and traditional-cyber security are both essential to ensure the secure operation of safety critical cyber-physical systems.

1.1 Previous Work

The importance of addressing the security of cyber-physical systems has been stressed by the research community (Byres and Lowe (2004) and Cárdenas et al. (2008b)). Cárdenas et al. (2008a) discuss the impact of denial-of-service (DoS) and integrity attacks on cyber-physical systems, and DoS attacks are further discussed by Amin et al. (2009). Mo and Sinopoli (2009) developed a methodology to detect replay attacks on a control system. A lot of literature deals with failure detectors in dynamic systems (Willsky (1976)). Sandberg et al. (2010) propose a method of using security indices in large scale power networks. Giani et al. (2009) address in their work the problem of secure and resilient power transmission and distribution.

1.2 Outline of the Paper

The goal of this paper is to develop model-based techniques capable of detecting integrity attacks on the sensors of a control system. First the results by Mo and Sinopoli (2009) are extended. Mo and Sinopoli (2009) analyze the effect of an attack on a control system in steady state. The system is assumed to be equipped with an estimator, a controller and a failure detector. It is assumed that an attacker will record and subsequently replay sensor

⋆ This research was supported in part by CyLab at Carnegie Mellon by grant DAAD19-02-1-0389 from the Army Research Office, by grant NGIT2009100109 from the Northrop Grumman Information Technology, Inc. Cybersecurity Consortium, and by grant 0955111 from the National Science Foundation. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of ARO, CMU, CyLab, NSF, NGC, or the U.S. Government or any of its agencies.

data while conducting the attack on the system. This deception, proposed a year before Stuxnet came to light, was exactly what the worm used to hide its activities — it secretly recorded what normal operations at the nuclear plant looked like, then played those readings back to plant operators, so that it would appear that everything was operating normally while the centrifuges were actually spinning wildly out of control (Broad et al. (2011)).The class of systems incapable of detecting such attacks is identified and a control algorithm is designed, which addresses this vulnerability by adding a zero-mean Gaussian authentication signal to the Linear Quadratic Gaussian (LQG) optimal control. It is further shown that the authentication signal enables the failure detector to detect the replay attack, albeit degrading the control performance of the system. In this paper a way to design the covariance of the authentication signal is provided, to minimize the performance loss while guaranteeing a certain probability of detection. In order to validate the design, the algorithm is implemented on a MIMO system (a chemical plant) that matches the assumptions by Mo and Sinopoli (2009).

The second part of the paper focuses on the specific problem of security for micro-grids. A unique characteristic of a power grid is the difficulty to predict the statistical distribution of the load variation in the system, which differentiates it from the original system model proposed by Mo and Sinopoli (2009). By concentrating on a very simplistic model of a power grid consisting of one generator and several loads, a method is proposed by which an attacker can induce the system to unnecessarily shed load in areas of interest while remaining undetected. Since the detection scheme by Mo and Sinopoli (2009) is not guaranteed to be successful, an alternative detection scheme is proposed and validated in simulations. Like Mo and Sinopoli (2009), an attacker is assumed to perform a man-in-the-middle attack on the sensors and actuators, intercepting the transmission and relaying false data instead. This false data can either be invented by the attacker, or replayed from an earlier recording of the transmitted data.

2. OPTIMAL DESIGN OF AN AUTHENTICATION CONTROL SIGNAL

In this section a methodology is provided to optimally design the authentication signal to detect replay attacks in linear systems. First the methodology proposed in Mo and Sinopoli (2009) for detecting replay attacks in general linear systems is briefly reviewed, after which the new system and attack models are introduced. The design the optimal authentication signal is then shown, to maximize the detection rate while keeping the increase in cost bounded. Finally the findings are validated by applying them to the control of an industrial chemical plant.

2.1 Preliminaries

Consider a linear time-invariant (LTI) system with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$ and $C \in \mathbb{R}^{m\times n}$ as the system matrices. The states and outputs of the system thus evolve according to the following equations:

$$x_{k+1} = Ax_k + Bu_k + w_k \in \mathbb{R}^n, \tag{1}$$

where,

$$x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma),\ w_k \sim \mathcal{N}(0, Q) \tag{2}$$

and

$$y_k = Cx_k + v_k \in \mathbb{R}^m, \tag{3}$$

$$v_k \sim \mathcal{N}(0, R). \tag{4}$$

The minimum variance unbiased estimator of such a system can be approximated by a fixed-gain Kalman filter [1], which takes the following form:

$$\widehat{x}_{k+1|k} = A\widehat{x}_{k|k} + Bu_k, \tag{5}$$

and

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + K\left(y_k - C\widehat{x}_{k|k-1}\right), \tag{6}$$

$$\widehat{x}_{0|-1} = \bar{x}_0, \tag{7}$$

where

$$K \triangleq P_\infty C^T\left(CP_\infty C^T + R\right)^{-1}, \tag{8}$$

and $P_\infty = \lim_{k\to\infty} P_k$ and P_k satisfies the following recursive equation:

$$P_{k+1} = AP_kA^T + Q - AP_kC^T(CP_kC^T + R)^{-1}CP_kA^T.$$

Given this state estimate, the LQG optimal control u_k^* that minimizes the objective function

$$J = \min \lim_{T\to\infty} E\left[\frac{1}{T}\sum_{k=0}^{T-1}\left(x_k^T W x_k + u_k^T U u_k\right)\right], \tag{9}$$

where

$$W, U \succeq 0, \tag{10}$$

is given by

$$u_k^* = Lx_{k|k}, \tag{11}$$

where

$$L \triangleq -\left(B^TSB + U\right)^{-1}B^TSA, \tag{12}$$

and S is the solution of the following Riccati equation:

$$S = A^TSA + W - A^TSB(B^TSB + U)^{-1}B^TSA.$$

To implement a detector for the control system, $\mathcal{P}$ is defined as the covariance of the prediction error ($\mathcal{P} \triangleq CPC^T + R$), a window size of $\mathcal{T}$, and

$$g_k \triangleq \sum_{k-\mathcal{T}+1}^{k}\left(y_i - C\widehat{x}_{i|i-1}\right)^T\mathcal{P}^{-1}\left(y_i - C\widehat{x}_{i|i-1}\right). \tag{13}$$

g_k is χ^2 distributed with parameter $m\mathcal{T}$. The expected value is $Eg_k = \bar{g_k} = m\mathcal{T}$. Thus a detector for such a system at time k is of the form

$$g_k \underset{H_1}{\overset{H_0}{\lessgtr}} \text{threshold}. \tag{14}$$

Here, H_0 denotes the null hypothesis, and H_1 denotes hypothesis that the system is under attack, and a χ^2 test can be used to distinguish between them (Greenwood and Nikulin (1996)) and the detector is dubbed a χ^2-detector.

If we define $\mathcal{A} \triangleq (A + BL)(I - KC)$, then it is proven by Mo and Sinopoli (2009) that if $\mathcal{A}$ is stable, the distribution of g_k under replay attack will converge exponentially to the same distribution as g_k without the attack. As a result the asymptotic detection rate of the χ^2 detector is the same as its false alarm rate, i.e., the detector is unable to distinguish a system under the replay attack from a system that is running normally.

To detect a replay attack, a small random authentication signal $\Delta u_k \sim \mathcal{N}(0, \mathcal{Q})$ is superimposed on the optimal

[1] Time-varying Kalman filters and fixed gain filters will achieve the same asymptotic performance. Since the optimal filter rapidly converges to the fixed gain filter, the latter is usually used in practice.

control input u_k^*, which serves as a time stamp. It is proved that asymptotically the expectation of g_k under the attack will increase to

$$\lim_{k\to\infty} Eg_k = m\mathcal{T} + 2trace(C^T\mathcal{P}^{-1}C\mathcal{U})\mathcal{T}. \quad (15)$$

where $\mathcal{U}$ is the solution of the Lyapunov equation

$$\mathcal{U} - B\mathcal{Q}B^T = \mathcal{A}\mathcal{U}\mathcal{A}^T. \quad (16)$$

However, due to the authentication signal, the control input is not optimal any more. Mo and Sinopoli (2009) are able to prove that the increase in LQG cost (ΔJ) is trace $\left(\left(U + B^T S B\right)\mathcal{Q}\right)$.

2.2 Optimization of the Authentication Signal

There are two ways to solve the design problem behind the optimization. Firstly, the LQG performance loss (ΔJ) can be constrained to be less than some value, and the increase in the expected value of g_k in case of an attack is maximized. In this case, the optimal $\mathcal{Q}^*$ is the solution of the following optimization problem:

$$\begin{aligned} \underset{\mathcal{Q}}{\text{maximize}} \quad & trace(C^T\mathcal{P}^{-1}C\mathcal{U}) \\ \text{subject to} \quad & \mathcal{U} - B\mathcal{Q}B^T = \mathcal{A}\mathcal{U}\mathcal{A}^T \\ & trace[(U + B^T S B)\mathcal{Q}] \leq \Delta J. \end{aligned}$$

Remark 1. Ideally one would try to optimize the detection rate while maintaining the LQG lost. However, it can be shown that g_k under the attack follows a generalized χ^2 distribution and hence there is no analytical form of the detection rate. As a result, only maximization of the difference in the expectation is attempted.

Remark 2. It can be seen that the optimization problem is a linear programming problem and hence can be solved efficiently. Furthermore, it can be easily seen that if the last constraints are changed from ΔJ to $\alpha\Delta J$, then the optimal $\mathcal{Q}^*$ will be changed to $\alpha\mathcal{Q}^*$ instead.

Secondly, the expected increase in g_k in case of an attack can be constrained to be above a certain value, (thereby guaranteeing a fixed rate of detection). In this case, the optimal $\mathcal{Q}^*$ is the solution of the following optimization problem:

$$\begin{aligned} \underset{\mathcal{Q}}{\text{minimize}} \quad & trace[(U + B^T S B)\mathcal{Q}] \\ \text{subject to} \quad & \mathcal{U} - B\mathcal{Q}B^T = \mathcal{A}\mathcal{U}\mathcal{A}^T \\ & trace(C^T\mathcal{P}^{-1}C\mathcal{U}) \geq E\left[\Delta g_k\right] \end{aligned}$$

where Δg_k is defined as

$$\Delta g_k = g_k - \bar{g_k}.$$

Remark 3. It is easy to see that the two $\mathcal{Q}^*$ obtained from both the above optimizations will be scalar multiples of each other, thus solving either problem guarantees same performance. The choice of the problem to be solved can be done considering the control objectives. An intuitive way to see this, is that $\mathcal{Q}$ measures the sensitivity of the system to the different forms of the authentication signal. Thus, the $\mathcal{Q}^*$ should be a property of the system, indicating the optimal form of signal.

3. SIMULATION

3.1 System Model

We want to apply the above methodology to a simplified version of the Tennessee Eastman Control Challenge Problem (Downs and Vogel (1993)). The original problem requires co-ordination of three unit operations, with 41 measured output variables (with added measurement noise) and 12 manipulated variables. The control challenge presented by this case study is quite complex. However, a simplified version was proposed by N. Lawrence Ricker in 1993 (Ricker (1993)), which is the model we adopt. In this paper, Ricker derives a linear time-invariant dynamic model of the plant in its base-state, and a corresponding robust controller, with four outputs and four inputs.

$$\mathbf{y} = \begin{pmatrix} F_4 \\ P \\ y_{A3} \\ V_L \end{pmatrix} = \mathbf{Gu} = \begin{pmatrix} g_{11} & 0 & 0 & g_{14} \\ g_{21} & 0 & g_{23} & 0 \\ 0 & g_{32} & 0 & 0 \\ 0 & 0 & 0 & g_{44} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}. \quad (17)$$

The individual transfer functions are given in Equations 18–23 (the unit of s is assumed to be hr^{-1}):

$$g_{11} = \frac{1.7}{0.75s+1}, \quad (18)$$

$$g_{21} = \frac{45\,(5.667s+1)}{2.5s^2+10.25s+1}, \quad (19)$$

$$g_{23} = \frac{-15s-11.25}{2.5s^2+10.25s+1}, \quad (20)$$

$$g_{32} = \frac{1.5}{10s+1}e^{-0.1s}, \quad (21)$$

$$g_{14} = \frac{-3.4s}{0.1s^2+1.1s+1}, \quad (22)$$

$$g_{44} = \frac{1}{s+1}. \quad (23)$$

The transfer function g_{23} is not given in Ricker (1993). It was estimated using the method described in the paper. The system is sampled at 100 samples per minute. Also, we use $Q = 0.01I$, $R, W, U = I$.

3.2 Attack Model

We consider an attacker who knows all [2] the sensors and can hijack and modify their readings, but does not know the dynamics of the system. He only knows that the system is expected to be in steady state for the duration of the attack. Of the 30 minutes for which the system is simulates, the attacker records the sensor readings for the first 15 minutes, and then replays them to the controller for the next 15 minutes, while attacking the system undetected. The attack consists of varying the control inputs of the plant, to try and evolve it into a potentially dangerous state, like increasing the pressure beyond the rated value to cause the boiler to explode. Since the controller cannot get any information from the sensors, the system becomes open loop and no control performance can be guaranteed. The only hope to counter such an attack being to detect it, the focus of the next subsection will be intrusion detection.

[2] The requirement of control over all the sensors may be weakened if the system can be decomposed into several weakly coupled subsystems. In that case, compromising the sensors for one subsystem may be enough.

3.3 Authentication Signal

It can be ascertained that, for this system, $\mathcal{A}$ is stable. Thus, an authentication control input is necessart for detection of an integrity attack. The results of Remarks 2 and 3 can be applied to decouple the design of the signal into two steps. Since there is a linear relationship between performance loss and the amplitude of the signal, we can first identify the form of optimal $\mathcal{Q}$, and then design the norm based on the requirement of detection performance.

Form of $\mathcal{Q}$ For the first step two authentication signals are considered. The first one is not optimized. The second is designed to maximize the difference in the asymptotic expected value of g_k while maintaining LQG lost. In this case, the original LQG performance without authentication signal is $J = 0.62$ and we constrain ΔJ to be less than 0.1, which will results in no more than 16% LQG lost with respect to the original system. We use a χ^2 detector with a window size of 100 (averaging over 1 min). Each plot is the average performance of 500 simulations. The y-axis indicates $\Delta g_k/\bar{g}_k$, where Δg_k is defined as

$$\Delta g_k = g_k - \bar{g}_k.$$

As a result, the expected Δg_k is zero in the absence of an attack. During the attack, $E[\Delta g_k]$ becomes $2trace(C^T\mathcal{P}^{-1}C\mathcal{U})\mathcal{T}$.

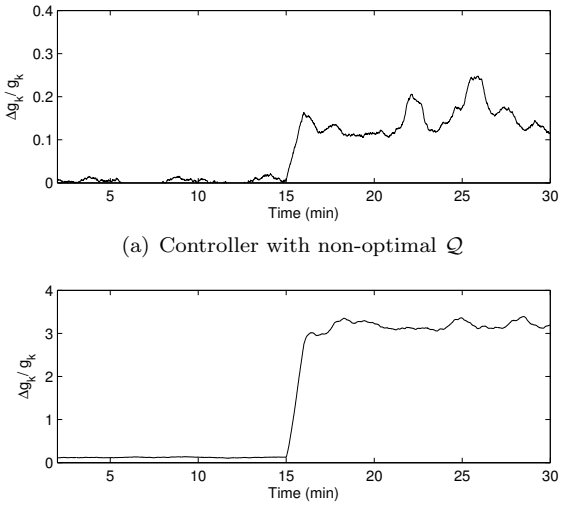
The results of the simulation using the non-optimal authentication signal are shown in Figure 1(a). It can be seen that while the controller is able to distinguish the start of an attack on average, the change in the value of g_k over normal operation is low enough, that it can be swamped by noise. The results of the simulation using the optimal authentication signal are shown in Figure 1(b). The importance of optimizing $\mathcal{Q}$ can be seen by the performance difference in the two simulations. When $\mathcal{Q}$ is optimized, the increase in g_k in case of attack is almost 15 times as much as the non-optimal case, for the same loss of performance (16%).

Norm of $\mathcal{Q}$ In the next step, $\mathcal{Q}^*$ is scaled to make Δg_k $0.2\bar{g}_k$, $0.4\bar{g}_k$, $0.6\bar{g}_k$, $0.8\bar{g}_k$ and $\bar{g}_k$, respectively, which corresponds increasing the strength of the signal. Figure 2 shows the Receiver Operating Characteristic (ROC) curves for the detector in each case, where the probability of detection 1 second after the attack begins is considered. The performance of the detector changes with an increase in $\|\mathcal{Q}^*\|$, so an appropriate signal strength can be designed taking into consideration the desired ROC and allowed performance loss. Figure 3 shows the relationship between the detection rate at a certain time (β) and $\|\mathcal{Q}^*\|$, which validates our assumption of Remark 1.

4. AN INTEGRITY ATTACK ON THE AUTOMATIC GENERATION CONTROL IN POWER SYSTEMS

The second system we consider is a power generator, or rather, a micro-grid with only one generator.

In a regular power grid, utility companies can predict the daily load pattern to within 1%. Any remaining imbalance, due either to inaccurate prediction of load (on a micro-scale, loads are switched on and off at random), or unexpected changes in supply and/or demand, is modeled



(a) Controller with non-optimal $\mathcal{Q}$

(b) Controller with optimal $\mathcal{Q}$

Fig. 1. A comparison of $\frac{\Delta g_k}{g_k}$ of the two detectors over time. In each case, the attack begins after 15 minutes of simulation time. Here $\bar{g}_k$ was estimated theoretically, and is less than the practical value, this the value of Δg_k is not exactly zero.

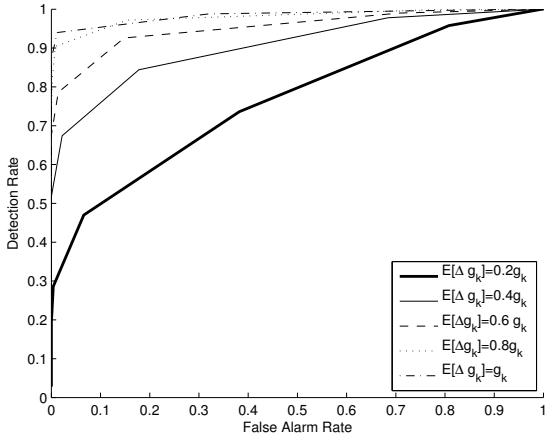


Fig. 2. ROC curves for detector, when Δg_k is $0.2g_k$ (dark solid line), $0.4g_k$ (thin solid line), $0.6g_k$ (dashed line), $0.8g_k$ (dotted line) and g_k (dash-dot line). Detection up to 1 second after attack is considered.

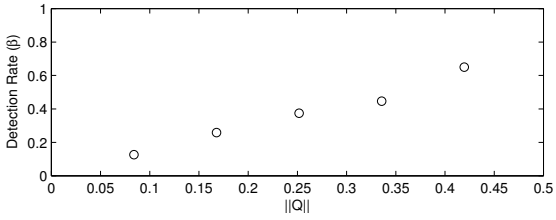


Fig. 3. β vs. $\|\mathcal{Q}^*\|$.

as noise, and handled by operating generators in frequency control mode by using a mechanical speed governor. The grid frequency is a critical indicator of system health. In nominal conditions, such imbalance is within 1% of the expected load (although the statistical distribution is hard to predict), and the generator is able to alter its output continuously to keep the frequency constant. The frequency is uniform over all of the grid, and a significant change in the value can cause transient instabilities leading to cascading failures. The target value for the frequency is thus within 1% of the nominal value. This necessitates that the closed-loop system involving the speed governor, the turbine and the rotor be a stable system.

The power grid also has a load shedder, which checks if the demand load is within the capabilities of the generator. If the load exceeds the maximum rating of the generator, this load-shedding control senses a drop in the frequency more than the rated value, and starts cutting of supply to lower priority loads till the demand equals the supply. A sudden problem in the load does not necessarily indicate an attack or a faulty sensor, and should be handled without raising an alarm. Since the Gaussian noise model is not guaranteed for this scenario, the method proposed by Mo and Sinopoli (2009) could possibly fail.

A linear model of the generator described in Bergen and Vittal (2000) is adopted. The block diagram of the model is shown in Figure 4. ΔP_C denotes the change in demand by increasing load, and $\Delta\omega$ denotes the change in grid frequency.

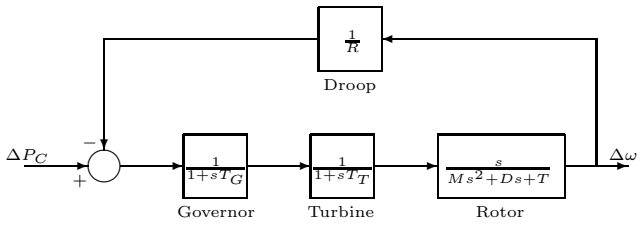


Fig. 4. Generator Block Diagram

4.1 Attack Model

For this plant, a slightly different kind of attack is considered. The system is attached to five loads numbered 1 through 5, where a lower load number indicates lower priority. It is assumed that the goal of the attacker is to cause the power to be shut down for some time in an area denoted by one of the loads, say load 3. We again consider an attacker who knows all the sensor readings and has the ability to hijack and modify them. The attacker also knows when the power to load 3 has been cut off. To achieve his goal, the attacker hijacks the frequency sensor(s), and makes it seem to the load shedding logic that the system has more load than can be supplied. The controller thus shuts off power to load 1. The attacker maintains the low frequency value, while the controller continues to shut off power to loads 2 and 3. The attacker releases the hold on the frequency sensor. During this time, the speed governor has reduced the steam intake to reduce the supplied power down to the demand value (of only loads 4 and 5), thus maintaining the grid frequency.

4.2 Countermeasure

As mentioned before, the white Gaussian noise assumptions fails in this case. However, the idea as used by Mo and Sinopoli (2009) can still be employed, by adding a wide sense stationary (WSS) Gaussian authentication signal $\Delta P_c \sim \mathcal{N}\left(0, \sigma^2\right)$ to the set point of the generator. In this case, a model of the generator is simulated inside the controller, which calculates the effect $\widehat{\Delta\omega}$ of this authentication signal on the grid frequency ω. The detector then observes the actual grid frequency, and verifies whether the predicted effect is present. A high-pass filter can remove the slow changes caused by a system fault, to obtain the actual variation $\Delta\omega$.

The quadratic values considered will be $g_k = \widehat{\Delta\omega}_k \Delta\omega_k$. Since the generator is a LTI system and the input is a WSS Gaussian process,

$$\Delta\omega_k \sim \mathcal{N}\left(0, \sigma'^2 + \sigma_n^2\right), \tag{24}$$

where σ'^2 is related to σ^2 based on the system characteristics and σ_n^2 is variance of the process noise. It can be assumed that this authentication signal is uncorrelated to the rest of the noise in the system, or any noise generated by the attacker. Thus, if the authentication signal is present in the grid frequency,

$$E\left[\Delta\omega_k \widehat{\Delta\omega}_k\right] = E\left[g_k\right] = \sigma'^2. \tag{25}$$

If the signal is not present, the output is completely uncorrelated to the authentication signal, and

$$E\left[\Delta\omega_k \widehat{\Delta\omega}_k\right] = E\left[g_k\right] = 0. \tag{26}$$

We can thus detect the absence of the authentication signal in the output and hence, the attack.

σ'^2 should be fixed to a value which does not cause more variation in the generator frequency than is allowed in the system specifications.

4.3 Simulation

We use a window $\mathcal{T} = 10s$. The five loads are set up to vary randomly by a small amount every 10 seconds. The load shedding logic, which is usually manually controlled, has been given a response time of 10 seconds, to simulate manual intervention. The system is simulated for 50 seconds. In the first case, we consider that load 3 has some problem, which causes the output frequency to drop steadily. This frequency drop is removed when load 3 is shed. In the second case, an attack begins after 10 seconds, and continues till the power has been shut off to load 3.

The change in grid frequency during simulation of a system fault is shown in Figure 5. The attacker just follows this profile while attacking. Figure 6 shows the values of g_k during the simulations. It can be seen that the expected value of g_k does fall to 0 for the duration of the attack, whereas in the case of a system fault, the expectation remains above 0. All plots show an average of 500 simulations.

5. CONCLUSION AND FUTURE DIRECTIONS

In this paper, the problem of detecting integrity attacks on control systems was investigated. In the first part of the paper, the results of Mo and Sinopoli (2009) regarding sensor

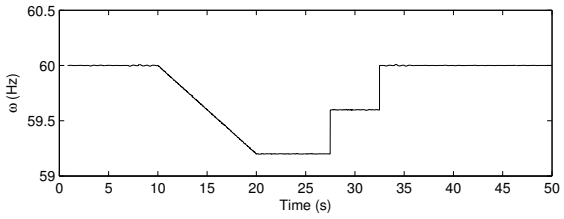
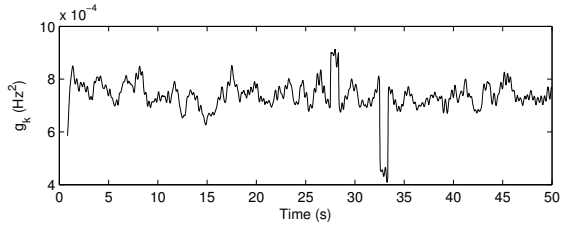
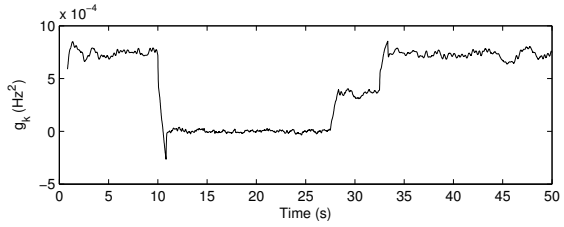


Fig. 5. The variation in grid frequency as seen by the controller during a system fault. The attacker just follows this profile while attacking.



(a) System Fault



(b) Attack

Fig. 6. g_k as a function of time during a system fault, and an attack. This shows that the detector is indeed able to detect the fall in g_k due to an attack.

replay attacks were extended by providing a procedure to optimally design an authentication control input capable to maximize the detection rate, while maintaining a specified bound on the performance loss due to the introduction of such signal. The relationships between performance loss, detection rate and the covariance of the Gaussian authentication input were characterized. Such relationship allows the designer to achieve the desired ROC by scaling the norm of the covariance of the authentication signal appropriately. Simulations on a reduced order model of a chemical plant were used to illustrate the results. In real world scenario, several engineering considerations can be employed to improve upon the proposed design. For example, the authentication signal can be injected at scheduled intervals rather than continuously, thus reducing the loss in performance. Also, using an authentication signal should not interfere with safe operations of the plant.

In the second part we consider more general integrity attacks, where the attacker is not limited to strictly replaying past observations and the process noise is not restricted to be Gaussian. In this case we show how to detect the presence of an authentication signal using autocorrelation. While these results pertain to replay attacks as defined by Mo and Sinopoli (2009), since the system checks for the presence of a time dependent authentication signal, it is a reasonable assumption that the methodology will also work against general integrity attacks which cannot precisely reproduce this authentication signal. We validated the method on a simplified model of a micro-grid, by showing how the proposed approach can detect integrity attacks on the frequency sensor of the load-shedding control which would otherwise cause the system to unnecessarily shed load. Future work will concentrate on extending these techniques to more sophisticated attack models and to distributed control systems.

REFERENCES

Amin, S., Cárdenas, A., and Sastry, S. (2009). Safe and secure networked control systems under denial-of-service attacks. *Hybrid Systems: Computation and Control*, 31–45.

Bergen, A.R. and Vittal, V. (2000). *Power Systems Analysis*. Prentice Hall, 2 edition.

Broad, W.J., Markoff, J., and Sanger, D.E. (2011). Israeli test on worm called crucial in iran nuclear delay.

Byres, E. and Lowe, J. (2004). The myths and facts behind cyber security risks for industrial control systems. *Proceedings of the VDE Kongress*.

Cárdenas, A., Amin, S., and Sastry, S. (2008a). Secure control: Towards survivable cyber-physical systems. *28th International Conference on Distributed Computing Systems Workshops, 2008. ICDCS'08*, 495–500.

Cárdenas, A.A., Amin, S., and Sastry, S. (2008b). Research challenges for the security of control systems. *Proceedings of the 3rd conference on Hot topics in security*.

Carlin, J. (1997). A farewell to arms. URL `http://www.wired.com/wired/archive/5.05/netizen.html`.

Downs, J. and Vogel, E. (1993). A plant-wide industrial process control problem. *Computers & Chemical Engineering*, 17(3), 245–255.

Giani, A., Sastry, S., Sandberg, H., and Johansson, K. (2009). The viking project: An initiative on resilient control of power networks. *Resilient Control Systems, 2009. ISRCS '09. 2nd International Symposium on*, 31 – 35. doi:10.1109/ISRCS.2009.5251361.

Greenwood, P.E. and Nikulin, M.S. (1996). *A guide to chi-squared testing*. John Wiley and Sons, Inc.

Markoff, J. (2010). A silent attack, but not a subtle one. URL `http://www.nytimes.com/2010/09/27/technology/27virus.html`.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. *Proceedings of the 47th annual Allerton conference on Communication, control, and computing*, 911–918.

Ricker, L. (1993). Model predictive control of a continuous, nonlinear, two-phase reactor. *Journal of Process Control*, 3(2), 109–123.

Sandberg, H., Teixeira, A., and Johansson, K.H. (2010). On security indices for state estimators in power networks. *First Workshop on Secure Control Systems, CPSWeek 2010*.

Willsky, A. (1976). A survey of design methods for failure detection in dynamic systems. *Automatica*.