Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Security of Control Systems with Erroneous Observations

Jaewon Kim * **P. R. Kumar** **

* *Texas A&M University, College Station, TX 77843 USA (e-mail: jwkim8804@tamu.edu).*
** *Texas A&M University, College Station, TX 77843 USA (e-mail: prk@tamu.edu).*

Abstract: We address the problem of security of stochastic control systems when observation measurements used to close the control loop may be erroneous, due to a malicious adversary who has intercepted the associated sensors or the communication network. We show how the method of dynamic watermarking can be employed to secure such a system. This is a method of defense based on stochastic considerations, relying on the inability of the attacker to separate the ambient noise present in the system from a deliberatively superimposed random watermark. We present the results of experiments against several attacks, and show the capability of this method to detect attacks in all the tested cases. The experiments are conducted on a prototypical process control system consisting of two coupled water tanks.

Keywords: Security, Malicious Sensors, Cyber Physical Systems, Dynamic Watermarking.

1. INTRODUCTION

Advanced control systems including robotics systems, power grid, unmanned aerial transportation systems, autonomous transportation systems, process control systems, etc., rely on measurements reported by sensors to provide state information and situational awareness to the control logic so that it can implement the feedback control law. However, such systems are vulnerable to cyber attacks if the sensor measurements are deliberately corrupted, thereby resulting in inappropriate feedback. There have been several such reported attacks on control systems. A prominent one is the Stuxnet replay attack on a system of centrifuges, Kerr et al. [2010], where fictitious measurements were provided to the control system. The problem of susceptibility to malicious attacks against control systems becomes even more acute when sensor measurements are conveyed over a network and thus liable to interception by malicious adversaries. Adversarial trespassers can hack into the system through open networks and conduct malicious attacks to cause critical damage to the entire system.

In this paper, we address the problem of cyber-security for process control systems with erroneous observations.

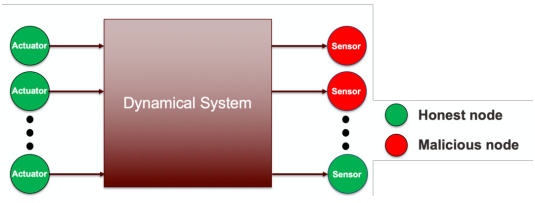

Fig. 1. Dynamical System with some malicious sensors.

We specifically address the problem of attacks on sensor measurements. When erroneous sensor measurements are used in feedback control loops, it can lead to erroneous actuation that can cause poor performance or instabilities. Malicious sensors can distort the measurements to achieve an undesirable objective that they may have, such as degrading the system performance or destabilizing the closed loop.

We explore a general-purpose solution called "Dynamic Watermarking (DW)" consisting of two tests that has been proposed in Satchidanandan and Kumar [2016a] to detect tampering with sensor measurements. It is shown there that this technique can be used to secure linear stochastic systems in the presence of malicious sensor nodes or compromised sensor measurements. In Ko et al. [2016], it has been shown that DW can detect attacks on autonomous vehicles. It is shown in Kim et al. [2019] that the DW method can be applied to process control systems. In earlier work, Physical Watermarking was introduced in Mo and Sinopoli [2009] to detect replay attacks, and extended in Weerakkody et al. [2014].

In this paper, we conduct a thorough study of the ability of the DW methodology to secure a prototypical example of a process control system, a system of coupled water tanks. In order to evaluate the technique it is imperative to conduct experimental results. The first reason that experimentation is imperative is that DW is fundamentally a stochastic methodology. It relies on the ability to detect correlations in signals to detect attacks. One may wonder why this cannot be done entirely thorough simulations. The reason is that simulation models, whether of IEEE many bus power systems TEES [2016], ICSEG [2013a], ICSEG [2013b], or process control systems, such as the Tennessee Eastman process, Chiang et al. [2001], only provide the model of the *deterministic* part of the system. They do not provide the model of the *system noise or observation noise* in the system. Thus these models cannot be used to test a fundamentally stochastic methodology

that camouflages a watermark in system noise to detect the attack. In fact, as will be seen in the process control system tested in this paper, the "froth" in the water flow into the bottom tank (Fig. 2), which constitutes "noise", is quite substantial. Therefore one needs to experimentally confirm that the tiny watermark signal that is superimposed can indeed survive passage trough the system. Therefore, a defense methodology against cyber-attacks that fundamentally relies on stochastic considerations needs experimental validation, not validation by simulation.

There are three other reasons why critical experimental verification is necessary. The DW methodology calls for actively injecting noise into a control system. Such active injection will not be tolerated in practice if it results in any adverse performance of the system in normal operation. We therefore employ a signal so minute that it does not perceptibly have any effect on performance. However, the lower the variance of the signal employed, the longer potentially is the time to detect an attack on the control system when such an attack commences. Therefore one needs to demonstrate that the detection delay is so small that one can take corrective action if an attack is detected. To address both issues, it is critical to experimentally demonstrate that these conflicting objectives can be satisfied in practice. Moreover, as noted above, the performance of both these objectives depends on stochastic considerations and cannot therefore be verified through simulations of deterministic models.

Finally, there is one more important reason for experimental investigation. The DW methodology is designed to be a general purpose method for defending against attacks on control systems. In much of the security literature, defenses are developed against specific attacks. In fact one daily receives patches of operating systems to plug against discovered vulnerabilities. In contrast, the DW methodology is meant to be a methodology to defend against a range of attacks, i.e., a general purpose methodology. In order to verify this it is necessary to test against a range of attacks, and verify that it is indeed so.

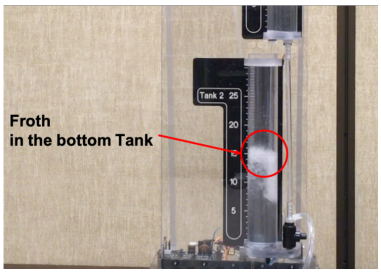

Fig. 2. The froth in the water flow into the bottom tank. The watermark signal needs to be detectable even in such noisy froth.

2. DYNAMIC WATERMARKING

In order to secure various control systems, we employ the technique of Dynamic Watermarking (DW); we call it a "watermark" because it is indelible like a watermark on a sheet of paper. DW's central idea is to have each actuator i inject $e_i[t]$, a random private signal superimposed on the control input $u_i(z^t)$. The actuator can check if the private excitation comes back appropriately transformed by using two specific DW Tests. The watermark $e_i[t]$'s statistics can even be disclosed to other nodes in the system; however, its actual realization is not revealed to any other node $i \neq j$ in the system.

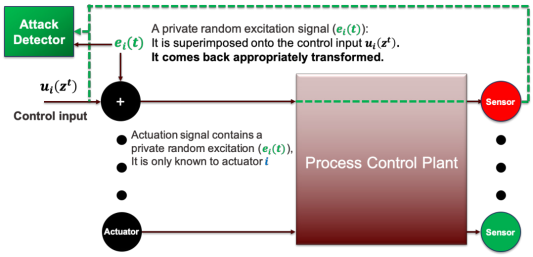

Fig. 3. Dynamic Watermarking method for securing CPS.

Here we will consider an independent and identically distributed (i.i.d.) $e[t] \sim \mathcal{N}(0, \sigma_e^2 I)$ vector of watermarks injected by the actuator nodes. The system with output vector $y[t]$, manipulated variable vector u_t, and white Gaussian noise $w[t]$ with zero mean and Covariance matrix $\sigma_w^2 I$, is

$$y[t+1] = Ay[t] + Bu_t(z^t) + Be[t] + w[t+1].$$

As a result, the sensor measurements y should satisfy the following equations if the sensor measurements are true:

$$(y[t+1] - Ay[t] - Bu_t(z^t)) \sim \mathcal{N}(0,\, BB^T\sigma_e^2 + \sigma_w^2 I)$$
$$E[e_i[t](y[t+1] - Ay[t] - Bu_t(z^t))]$$
$$= B_i\sigma_e^2 \sim \mathcal{N}(0,\, BB^T\sigma_e^2)$$

,where B_i = i-th column of B.
Based on these equations, each honest actuator $i \in 1, 2, ..., m$ subjects the reported sequence of sensor measurements $z(t)$ to the following two variance Tests:

TEST 1: The i-th node checks if the reported sequence of measurements satisfy

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (z[k+1] - Az[k] - Bg_k(z^k))$$
$$(z[k+1] - Az[k] - Bg_k(z^k))^T = \sigma_e^2 BB^T + \sigma_w^2 I_n$$

TEST 2: The i-th node also checks if the reported sequence of measurements satisfy

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} e_i[k](z[k+1] - Az[k] - Bg_k(z^k)) = B_i\sigma_e^2.$$

In Satchidanandan and Kumar [2016a], it is established that if the reported sequence of measurements passes both Test 1 and Test 2, then any malicious sensor present could not have distorted the actual measurement value beyond adding a zero power signal to the ambient noise. Specifically, the following Theorem is proved:

Theorem
In order to pass the two DW tests, the malicious can only report sensor measurements z[t] that satisfy

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (z[t+1] - Az[t] - Bu_t(z^t) - Be[t] - w[t+1])^2 = 0.$$

That is, the malicious agent can at best report sensor measurement that perturb the noise w[t] anyway present in the system by an additive signal of zero power.

3. THE EXPERIMENTAL SYSTEM

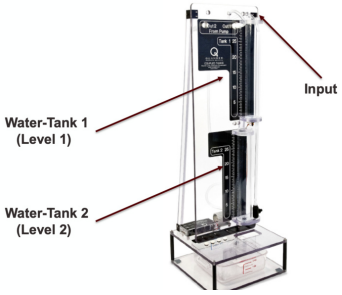

Fig. 4. Coupled Water-Tanks System.

The system tested consists of two coupled water tanks as shown in Fig. 4. The two tanks and a water basin are cascaded vertically, so that the upper tank feeds the lower tank whose water flows down into the basin. Pressure sensors at the bottom of each tank measure the water level, and provide it as a feedback signal to the controller. The actuation input to the system is the voltage of a motor that controls the rate of water flow into the upper tank. The main control objective is to maintain the water levels of both the upper and lower tanks at pre-specified set-points.

4. THE ATTACKS CONSIDERED

We consider the performance of the DW method for several attacks on the system of two coupled water tanks that may be launched by malicious sensors. We specifically consider the following range of attacks:

(1) Stealth Attack: In this attack, the adversary feeds back fictitious measurements obtained from a simulated system rather than the actual system.
(2) Replay Attack: The adversary stores the measurements obtained from the system at some time in the past, and plays them back as current measurement. This was used in the Stuxnet attack, Kerr et al. [2010].
(3) Toggling Sign Attack: The attacker flips the sign of certain measurements.
(4) Time Delay Attack: The attacker adds a time delay to the reported measurement.
(5) Average Value Attack: The attacker reports an average of a reference value, simulated measurements and actual measurements to the controller.
(6) Ramp-function Bias Injection Attack: This attack is designed to be deliberately different from other attacks. The attack is very slow, and ramps up only gradually. For about 200ms, the reported sensor measurements are approximately the same as the true measurements. After this initial period, the measurements begin to significantly diverge. Thus this attack does not cause any damage unless it remains undetected even after it reports significantly wrong measurements. The attacker superimposes a bias consisting of a ramp function on the actual sensor measurements.

We implement each attack and show how each leads to the control system failure, causing overflows in the water tanks. For each attack, we then employ the DW technique to secure the coupled water tanks system. The DW technique consists of superimposing a private excitation on the actuation input which is of sufficiently small magnitude that it does not perceptibly affect the performance of the system. The system then employs an attack detector that examines the measurements reported by the sensors to check if they contain the appropriately transformed version of the dynamic watermark by performing two tests of variance. The tests verify the veracity of the reported sensor measurements by checking if they are appropriately correlated with the injected private excitation signals (i.e., the watermark) as described above.

If the statistics of the reported measurements fail either of the two DW tests by going over the limit of pre-defined thresholds, then the attack detector concludes that the watermarks were distorted or removed, and generates a warning of sensor malfeasance.

What recourse to take once a threat has been detected generally depends on the particular system at hand. In the coupled water-tanks system, further inflow of water is simply stopped, thus preventing overflow from the tanks.

5. EXPERIMENTAL RESULTS FOR SEVERAL ATTACK MODELS

We now consider the performance of the DW method against several attacks on the system of two coupled water tanks that may be launched by malicious sensors. We comprehensively evaluate the performance of the DW tests against the above range of attacks on the process control system.

For each attack, we first implement the system under attack in the absence of the DW defense, i.e., without any watermarking signal. This is shown in Figs. 5, 7, 9, 11, 13, 15. The attack commences at time 70s for the Stealth Attack and the Toggling Sign Attack, at time 100s for the Replay Attack, and at time 60s for the Ramp-function bias injection Attack, Time Delay Attack, and Average Value Attack. Prior to the commencement of the attack the system is seen to be controlled properly with the levels of both tanks maintained at their desired set-points in all scenarios. In each case, following the commencement of the attack, there is tank overflow (both tank 1 and tank 2).

Subsequently, for each attack, we implement DW. In each case we see that the attack is quickly detected by the DW test signals crossing the threshold set for normal operation. In all these attacks we see that both tests detect the attack.

5.1 ***Stealth Attack***

In this attack, at attack time beginning at time 70 sec, the malicious sensors simulate the coupled water-tanks system based on arbitrary actuator and measurement noises, and send the resulting erroneous measurements $z(k+1)$ to the controller.

As can be seen in Fig. 6, the water-tanks system is controlled properly before the attack time 70 sec; however, it overflows soon after the controller begins receiving the erroneous measurements.

The two DW tests are implemented to secure the whole system against any form of erroneous measurements re-

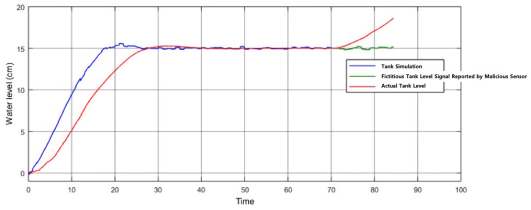

Fig. 5. Stealth Attack: The behavior of Tank 2 (Lower tank). The watermark signal has mean(μ_e) = 0, var(σ_e^2) = 0.01.

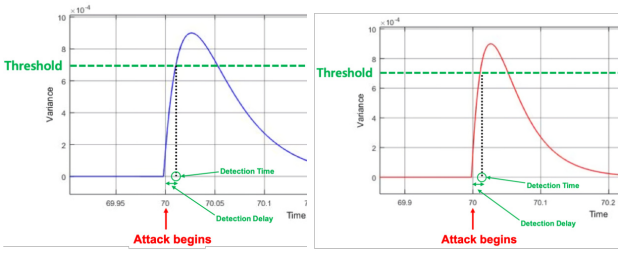

Fig. 6. Stealth Attack: Detection delay of the two DW Tests for Tank 2. The detection delay is less than 10ms.

ported by malicious nodes in the system. As we can see in the above graph, the attacks result in the tanks overflowing. However, as can be seen from Fig. 6, the DW tests detect the attack less than 0.01 seconds after the attack begins. One may note that the detection delay is very small even though the watermark signal is of very small variance, just 0.01, so that it does not affect the normal operation of the water-tanks system.

5.2 *Replay Attack*

In this attack, beginning at time 40 sec, the malicious sensor stores the series of actual true measurements of the system and then replays them at attack time beginning at time 100 sec, informing these erroneous measurements $z(k+1)$ to the controller.

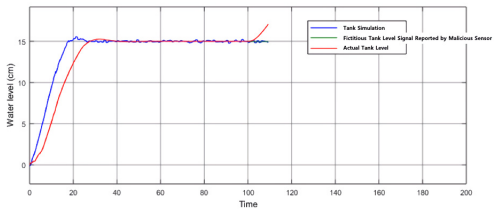

Fig. 7. Replay Attack: The behavior of Tank 2 (Lower tank). The watermark signal has mean(μ_e) = 0, var(σ_e^2) = 0.01.

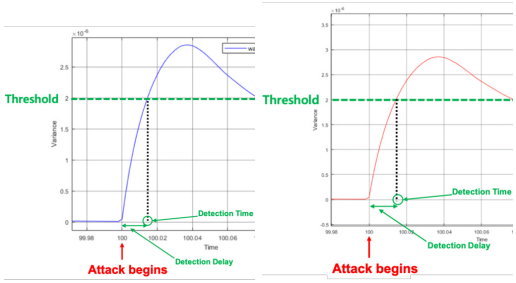

Fig. 8. Replay Attack: Detection delay of the two DW Tests for Tank 2. The detection delay is less than 15ms.

The behavior of the system without DW is shown in Fig. 7, with the tank overflowing soon after the attack commences. As can be seen from Fig 8, the DW tests detect the attack less than 0.015 seconds after the attack begins. As before, the detection delay is very small even though the watermark signal is of very small variance, just 0.01, so that it does not affect the normal operation of the water-tanks system.

5.3 *Toggling Sign Attack*

In this attack, at time beginning at time 70 sec, the malicious sensors flip the sign of actual tanks' measurement values, and send the resulting erroneous measurements $z(k+1)$ to the controller.

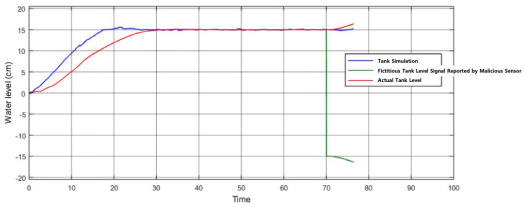

Fig. 9. Toggling Sign Attack: Attack based on toggling the sign of the true measurements: The behavior of Tank 2 (Lower tank). The watermark signal has mean(μ_e) = 0, var(σ_e^2) = 0.01.

The operation without watermarking is shown in Fig. 9, with the tank overflowing after the attack commences. As shown in the Figs. 10, the DW tests detect the attack less than 0.01 seconds after the attack begins.

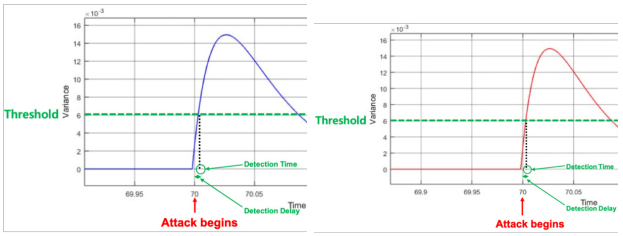

Fig. 10. Toggling Sign Attack: Detection delay of the two DW Tests for Tank 2. The detection delay is less than 10ms.

5.4 *Time Delay Attack*

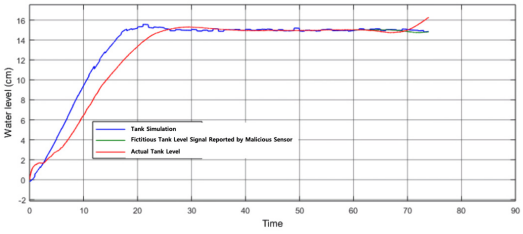

Fig. 11. Time Delay Attack: The behavior of Tank 2 (Lower tank). The watermark signal has mean(μ_e) = 0, var(σ_e^2) = 0.01.

In this attack, at time beginning at 60 sec, the malicious sensors add a time-delay to the actual tanks' measurement values, and inform the resulting erroneous measurements $z(k+1)$ to the controller. As seen in Fig. 12, the DW tests detect the attack less than 0.015 seconds after the attack begins.

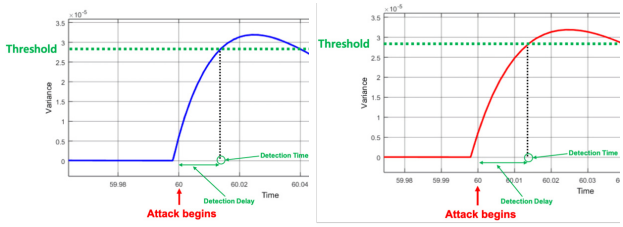

Fig. 12. Time Delay Attack: Detection delay of the two DW Tests for Tank 2. The detection delay is less than 15ms.

5.5 *Average Value Attack*

In this attack, beginning at time 60 sec, the malicious sensors take the average of three values: 1) reference input value, 2) simulation output value, and 3) actual tanks' measurement values, and inform the resulting erroneous measurements $z(k+1)$ to the controller.

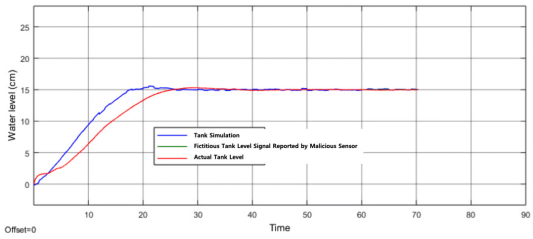

Fig. 13. Average Value Attack: The behavior of Tank 2 (Lower tank). The watermark signal has $\text{mean}(\mu_e) = 0$, $\text{var}(\sigma_e^2) = 0.01$.

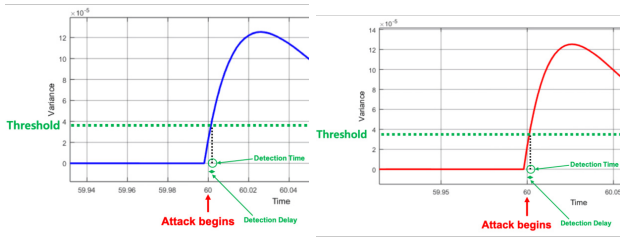

Fig. 14. Average Value Attack: Detection delay of the two DW Tests for Tank 2. The detection delay is less than 10ms.

As can be seen from the Fig. 14, the DW tests detect the attack less than 0.01 seconds after the attack begins.

5.6 *Ramp-function Bias Injection Attack*

This attack is different from the other attacks considered in that it is a very slow attack. In this attack, at attack time beginning at time 60 sec, the malicious sensors superimpose a ramp function bias on the actual tanks' measurement values, and informs the resulting erroneous measurements $z(k+1)$ to the controller.

As can be seen in Fig. 17, for about 200ms, the sensor measurements are approximately the same as the true measurements, after which they begin to significantly diverge from the true measurements. Hence the attack is not damaging to the system unless it is undetected even after it begins to report significantly wrong measurements.

As shown in the Fig. 16, the DW tests detect the attack less than 30ms after the significantly erroneous reporting of sensor measurements commences, which is about 0.23 seconds after the slow attack begins.

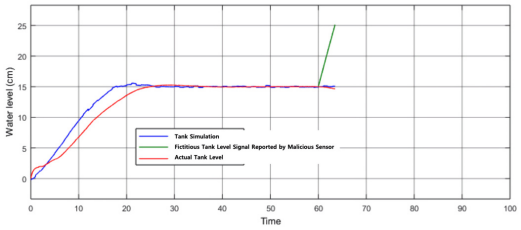

Fig. 15. Ramp-function Bias Injection Attack: The behavior of Tank 2 (Lower tank). The watermark signal has $\text{mean}(\mu_e) = 0$, $\text{var}(\sigma_e^2) = 0.01$.

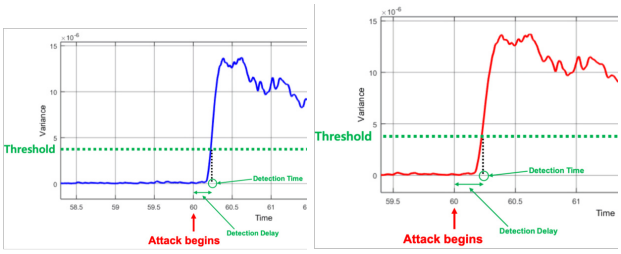

Fig. 16. Ramp-function Bias Injection Attack: Detection delay of the two DW Tests for Tank 2. The detection delay is less than 230ms.

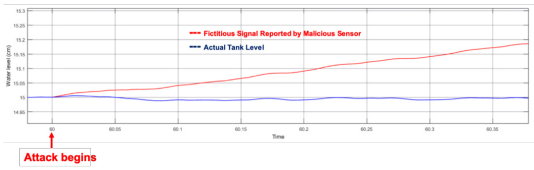

Fig. 17. Ramp-function Bias Injection Attack: The behavior of the slow attack.

From the experimental results on the system of coupled water-tanks, we see that the DW method reliably detects all the malicious sensor attacks considered. It does so employing a watermark level so minute that it does not adversely affect the normal performance of the system when it is not under attack. This is important since there is resistance to employing any methodology which degrades normal operating performance. In spite of this, the DW methodology detects the attack in all case in less than 15ms, with the exception of the very slow Ramp-function Bias Injection Attack, allowing ample time to prevent any adverse consequences.

With regard to the Ramp-function Bias Injection Attack, it is a very slow attack and can only damage performance very slowly. At the beginning of the attack time, the malicious sensors start superimposing a ramp-function onto the actual sensor measurements. Therefore, at the beginning of the attack, the injection value is minute. Thus, the reported measurements are very close to the true sensor measurements for about 200ms, as can be seen in Fig. 17. This means the malicious sensor acts as an approximately honest sensor for 200ms. The DW method detects the attack in less than 30ms after this time period of 200ms as the controller begins receiving sufficiently erroneous measurements from the malicious sensors. Since the Ramp-function Bias Injection Attack is a very slow attack compared to other types of attack strategies, the detection time is higher than for other

attacks; nevertheless, the DW method detects the attack successfully.

DW thereby acts successfully as a methodology to secure the control system against malicious attacks on the sensors or the information processing systems transporting the sensor values.

6. CONCLUSION

In this paper, we have addressed the problem of cyber-security of cyber-physical systems. Being a defense based on stochastic considerations, DW needs critical experimental evaluation to judge its capability since normal models of process control systems, or even other systems such as IEEE multi-bus power system models, model only the deterministic part of the system and do not provide any models of the noise in the system.

We have experimentally evaluated the performance of the DW method for detecting attacks on a prototype of a process control system, the two coupled water tanks system. We have tested a range of attacks on the system that are designed to lead to catastrophic results - tank overflow - if the system is not protected. We tested the DW method with a very small signal so that it does not affect the normal operation of the system. In each case, the DW method detects the attack in a short time, 10ms-15ms after the attack commences, for all except the slow-to-start Ramp-function Bias Injection attack, and is able to shut off the water inflow to prevent tank securing control systems with erroneous observations.

The advantage of the DW approach is that it is designed to secure the control system against arbitrary attack strategies. We experimentally demonstrate its effectiveness as a methodology to secure control systems by showing its effectiveness against a range of attacks on the sensor measurements.

ACKNOWLEDGEMENTS

This material is based upon work partially supported by NSF Science & Technology Center Grant CCF-0939370, the U.S. Army Research Office under Contract No. W911NF-18-10331, the U.S. Army Research Laboratory under Contract No. W911NF-19-2-0033, and U.S. ONR under Contract No. N00014-18-1-2048.

REFERENCES

Chiang, L.H., Russell, E.L., and Braatz, R.D. (2001). Tennessee eastman process. In *Fault Detection and Diagnosis in Industrial Systems*, 103–112. Springer.

Huang, T., Satchidanandan, B., Kumar, P.R., and Xie, L. (2018). An online detection framework for cyber attacks on automatic generation control. *IEEE Transactions on Power Systems*, 33(6), 6816–6827.

ICSEG (2013a). IEEE 118 bus system: https://icseg.iti.illinois.edu/ieee-118-bus-system/.

ICSEG (2013b). IEEE 14 bus system: https://icseg.iti.illinois.edu/ieee-14-bus-system/.

Kerr, P.K., Rollins, J., and Theohary, C.A. (2010). *The stuxnet computer worm: Harbinger of an emerging warfare capability*. Congressional Research Service Washington, DC.

Kim, J., Ko, W.H., and Kumar, P.R. (2019). Cyber-security with dynamic watermarking for process control systems. In *Proceedings of 2019 AIChE (American Institute of Chemical Engineers) Annual Meeting (Cyber Security Division)*. Orlando, Florida.

Ko, W.H., Satchidanandan, B., and Kumar, P.R. (2016). Theory and implementation of dynamic watermarking for cybersecurity of advanced transportation systems. In *2016 IEEE Conference on Communications and Network Security (CNS)*, 416–420. IEEE.

Ko, W.H., Satchidanandan, B., and Kumar, P.R. (2019). Dynamic watermarking-based defense of transportation cyber-physical systems. *ACM Transactions on Cyber-Physical Systems (TCPS): Special Issue on Transportation Cyber-Physical Systems*.

Kwon, J.S., Satchidanandan, B., Ko, W.H., Kim, J., Narasingam, A., and Kumar, P.R. (2018). Securing process control systems using dynamic watermarking. In *Proceedings of 2018 AIChE (American Institute of Chemical Engineers) Annual Meeting (Computing and Systems Technology Division)*. Pittsburgh, Pennsylvania.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 911–918. IEEE.

Satchidanandan, B. and Kumar, P.R. (2017a). Defending cyber-physical systems from sensor attacks. In *International Conference on Communication Systems and Networks*, 150–176. Springer.

Satchidanandan, B. and Kumar, P.R. (2019). On the design of security-guaranteeing dynamic watermarks. *IEEE Control Systems Letters*, 4(2), 307–312.

Satchidanandan, B. and Kumar, P.R. (2020). Towards characterizing the watermark-securable subspace of a linear stochastic system. In *Proceedings of 2020 12th International Conference on Communication Systems and Networks (COMSNETS)*. IEEE.

Satchidanandan, B. and Kumar, P.R. (2016a). Dynamic watermarking: Active defense of networked cyber–physical systems. *Proceedings of the IEEE*, 105(2), 219–240.

Satchidanandan, B. and Kumar, P.R. (2016b). Secure control of networked cyber-physical systems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 283–289. IEEE.

Satchidanandan, B. and Kumar, P.R. (2017b). On minimal tests of sensor veracity for dynamic watermarking-based defense of cyber-physical systems. In *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*, 23–30. IEEE.

TEES (2016). IEEE 300 bus system: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/ieee-300-bus-system/.

Weerakkody, S., Mo, Y., and Sinopoli, B. (2014). Detecting integrity attacks on control systems using robust physical watermarking. In *53rd IEEE Conference on Decision and Control*, 3757–3764. IEEE.