Mitigation of Industrial Control System Attacks Using A Boosted Filter Ensemble

Kelvin Ly*, Orlando Arias*, Jacob Hazelbaker*, Andrew Hughes*
Faculty Advisor: Yier Jin†
*Department of Electrical and Computer Engineering, University of Central Florida
†Department of Electrical and Computer Engineering, University of Florida
rangertime@knights.ucf.edu, oarias@knights.ucf.edu, yier.jin@ece.ufl.edu

***Abstract*—We are developing a solution that uses an ensemble of modelling methods to detect sensor and PLC compromises/failures. These different methods are combined together using gradient boosting to produce a highly robust filter. Various other novelties in our approach are mentioned as well.**

I. Introduction

The CSAW 2017 Embedded Security Challenge focuses on security industrial control systems against attacks on PLCs and sensors. The challenge is based around the problem of securing a system built on top of legacy PLCs. These PLCs will be running code that we will not have access to, and consequently, they effectively act as black boxes that may or may not be controlled by the attacker. Similarly, the sensors may or may not be compromised by the attacker. Following Shannon's maxim, we assume the attacker has an almost perfect knowledge of the system, in terms of its implementation, aside from not knowing some set of secrets. The goal overall is to deploy a system on top of the existing infrastructure to mitigate an attacker's ability to cause the system to break by allowing it to enter an unsafe state.

There are two different forms of attacks, attacks that alter the output of a PLC, and attacks that alter the outputs of sensors in the system. Our defense against both of these consists of providing predictors that try to predict the future state of the system, estimators that support the predictors by estimating unknown hidden variables, and classifiers that use the predictions and estimations to determine whether a given deviation should be classified as an attack or failure. The predictors and estimators here will be a combination of some of the ones commonly found in control theory and those used in machine learning. This ensemble of classifiers will be combined together using gradient boosting to provide a hopefully robust single classifier that can accurately detect attacks. Apart from sharing this basic framework, PLC outputs and sensor outputs will also have some additional security measures that take advantage of their individual attributes.

II. Proposal

Our method builds on top of the current standard for using redundancy to improve system reliability, enhancing it to be more robust. Our system shold be able to withstand some attacks where the attacker controls a majority of the controllers, and can detect some attacks on the sensors within the system as well. In an attempt to provide a robust, general solution that can reliably detect a large number of attacks than simple majority voting can't. The system will use an ensemble of approaches to determine the likelihood of error or attack originating from any of the PLCs. These of these approaches will individually calculate the likelihood of a failing or compromised PLC. These approaches will include modeling the PLCs as time delay neural networks, hidden Markov models, and low order control systems, using simple majority voting, and classifying their behavior using support vector machines. After some experimentation, some of the lower scoring approaches may be pruned to improve the performance of the system.

A. Security mechanisms exclusive to PLCs

Only one, if any, of the PLCs, will be virtually connected to the plant at any moment. All the PLCs will receive sensor data for each time step, but only the output of one of them will actually be fed into the actuators in the system. If the system predicts that an attacker has assumed complete control over all the PLCs, the system will temporarily switch into running using the predictors and estimators described above. Their control over the actuators will be time multiplexed, with the length of each PLC's time slot governed by the classifiers' belief in that given PLC's trustworthiness, preventing any attacker from exerting too much control over the system.

Additionally, the devices which are not at a given instant controlling the plant may be subject to various disturbances, which will provide additional data that can be used to characterize the PLCs. In particular, this will prevent some of the machine learning algorithms from overfitting by broadening the conditions they observe. It is expected that the systems are probably naturally fairly stable, and consequently most of the data collected will be of the system working at or near its set point. Emulating disturbances allows a more complex model that more accurately predicts system behavior to be developed.

B. Security mechanisms exclusive to sensors

Following prior work, watermarking signals are added to the actuator values (when the actuator is analog). Assuming

the signal is linear to some extent, it is expected that some amount of this signal will leak into the sensor data. This will prevent some classes of false data injection on the sensors, as the watermarking signals will be made to be spoof resistant, as will be explained below. The watermarking signals will be very weak in nature, with a power level near noise floor of the system. The use of spread spectrum coding will allow this to still be detected, only at the cost at requiring a longer time window to do so. The signal strength, window size, and coding will all be adaptively adjusted to individual control systems.

The signal will be generated using Gold codes, a class of pseudorandom sequences which tend to have very good autocorrelation and cross correlation properties. These codes are generated using two distinct maximum length sequences of the same period. The starting point of the Gold code and the amount the maximum length sequences are shifted relative to each other provide a pair of secret values that someone attempting to capture the signal would need to determine. In most systems, these values are shared prior to an attempt to use the channel to communicate. It would take $O(n^2)$ operations to determine these two parameters via brute force, where n is the period of the Gold code.

This system will use chip modulation, where a fast, pseudorandom code is phase shifts the information that is being sent. This spreads out the signal to cover a wider bandwidth, making it easier to recover at higher signal to noise rates, at the cost of lowering the information rate.

Let's examine the case where both the chip code and the information code are Gold codes. This would mean the controller would need to keep track of four parameters and perform twice as many XOR operations. However, the effect on the attacker is more drastic. The attacker, who is attempting to recover the parameters, would require $O(n^4)$ operations to brute force the parameters.

So let's examine if this chip modulated Gold code was used to modulate another Gold code. We'll use the simplifying assumption that this Gold code is of the same period as the first two. Then the controller would need to keep track of six parameters and perform three times as many XOR operations as in the first case. More interestingly, the attacker is faced with $O(n^6)$ operations to recover the parameters now.

This shows that a linear increase in work for the controller results in an exponentially large amount of work for the attacker, something that is very reminiscent of private keys and secrets in cryptography. A sufficiently large layering of sequences with a sufficiently large period should provide enough security to prevent an attacker from spoofing the signal. This kind of signal would normally be close to useless, as each additional layer causes an exponential slowdown of the information transmission, but this works perfectly for this use case, where all that matters is the presence of a signal. This unfortunately only provides some small amount of protection here, as the attacker still has plenty of ways of altering the sensor data without disrupting the watermarking signal. However, this does seem like a promising area for future research.

C. Filters to be implemented

This section will cover the different systems that will feed into the gradient boosting algorithm. Each of these will independently model, predict, and monitor the inputs and outputs of PLCs and sensors.

One filter will be a simple majority filter. Given an odd number of redundant PLCs, the filter will select the median value of the PLCs. The standard deviation of the output will be measured online, and this estimate will be continually improved as the system runs. Any PLCs deviating from the median value past a certain multiple of the standard deviation will be marked as faulty.

Another filter will be use least means square to estimate the control system, and each PLC using third order models. The filter will constantly accept new data to improve its estimates, and it will also maintain some confidence value for the accuracy of the parameters. A PLC whose parameters deviate noticeably from those of the others, or whose signal violates its own predicted signal by a noticeable margin, will be marked as faulty.

A third filter will attempt to use time delay neural networks to model the system, PLCs, and sensors. It will attempt to classify the signal at any moment as belonging either within stable operation, or as belonging to a compromised or failed device.

III. Related Works

False data injection detection is very much still an open research area. A lot of work has been done in it for the specific case of large power grid systems, beginning with the work by Yao which demonstrated that a family of attacks were possible which were undetectable from normal power grid operation from the standpoint of state estimation [1]. Various other papers devised methods to overcome this shortcoming [2].

IV. Discussions

This system uses many separate, fairly robust, well studied algorithms as a result of the very general nature of the problem it is attempting to solve. Industrial control systems are very diverse, and consequently any attempt to protect them in a general case must be very robust to adequately work in a large number of systems. It isn't expected that any one of these algorithms can adequately model all control systems, but it is expected that at least some of them will perform well enough for all control systems. The gradient boosting performed acts to select the set of algorithms that work well for each system.

V. Conclusion

TBD

REFERENCES

[1] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[2] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on dc state estimation," in *Preprints of the First Workshop on Secure Control Systems, CPSWEEK*, vol. 2010, 2010.