TWC SBE : Small : From Threat to Boon : Understanding and Controlling Strategic Information Transmission in Cyber-Socio-Physical Systems

Cedric Langbort, University of Illinois at Urbana-Champaign, langbort@illinois.edu

Overarching Theme:

The ever-increasing reliance of cyber-physical systems, computer networks, and infrastructures on data, coupled with their integration of new data sources, filters, and sensors have opened the door to new vulnerabilities. In particular, it is now possible to envision bringing down a system by subverting the data that feeds it.
This kind of move, which we call Strategic Information Transmission (SIT) because of the aspects it shares with a subfield of Information Economics of the same name, constitutes an indirect attack in the sense that it is not the attacker's action that is problematic in and of itself but, rather the response it triggers in the system under attack. Data sources engaged in strategic information transmission are fundamentally different than faulty ones, because they are not merely failing or dysfunctioning, but actively trying to mislead the system for their own benefit. This project focuses on understanding the consequences of SIT as a mode of attack.

Recent Contribution #1: SIT under human biases

We considered situations where a prospect theoretic malicious sender aims at misleading at prospect theoretic receiver about the value of a variable of interest.
Such agents deform probabilities by overweighing small probability events and under-weighing large probability ones. There is thus an opportunity for the sender to further manipulate the receiver's beliefs, a priori.

After extending the prospect-theoretic setting to allow for continuous choice spaces, we showed that, in the linear quadratic gaussian case, equilibrium strategies are the same for "fully rational" and prospect theoretic agents, even though the latter experience a varying degree of degraded performance depending on how they interpret small probabilities.

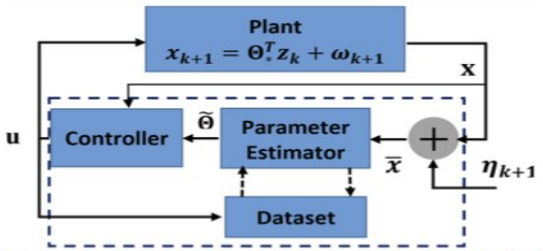

Fig. 1: Diagram of a closed-loop system with the learning block, attacked in the loop

Recent Contribution #2: Attack in the loop of learning -based control

Although malicious agents usually target sensors and actuators, the increased complexity of CPS systems is constantly opening new angles of attack for the adversarial agents bringing about dramatically increasing vulnerability. One possible angle of attack can be the learning algorithm embedded in a control unit.

As a first approach towards studying this new attack modality, we considered a situation where an unknown linear system controlled by an adaptive control algorithm undergoes a learning algorithm attack. This situation can be seen as a special case of data poisoning attacks on machine learning algorithms. An additional complication in the context of adaptive control, however, is that the learning algorithm is in the loop, driving the process to be controlled.

The effects of this additional closed loop are not intuitively clear. On the one hand, one may reason that the attack will be more damaging, since false data not only triggers mis-estimation of the system's model, but this mis-estimation may itself result in the computation and injection of an incorrect control input signal, further driving the system's state away from its desired optimal value. On the other hand, one could argue that because the controller constantly adapts itself and receives new measurements, it might be able to correct the effects of an attack if it is limited in space and time.

Focusing on LQ adaptive control and considering a learning-based algorithm for which rigorous regret bounds are available, we establish new regret bounds in the presence of learning algorithm attack.

Grant Number:#1619339