Malicious Agent Classification In Multi-Agent Formation Control

William Chong: wchong@stanford.edu
Department of Mechanical Engineering: Stanford, CA. USA

**Introduction**

Multi-agent formation control laws typically depend on the graph state of the robot formation to calculate control inputs for each robot in a decentralized method to reach a goal formation [1]. If malicious agents are present, then those actors can impede the formation control. This work attempts to train a malicious agent classifier that has no knowledge of the formation control law as well as the connectivity of the graph. Feature inputs to the network consist of observed robot positions, velocities, and position FFT data along the trajectories. The network is designed with two layers where the first layer is a parallel structure of 2 LSTMs (position, velocity) and 1 MLP (position FFT), which output is then concatenated to an MLP for the second layer. The output is a 1x1 binary vector indicating whether each robot is a good or bad agent. The training and testing data contain 0-5 malicious agents for each individual simulation. Results show high precision and recall, as well as a high balanced accuracy given no knowledge of the graph connectivity or the control law.

**Figure 1 (next column):** Desired robot formation shape. Robot positions are randomly initialized at the start of every simulation and assigned random indices, and malicious agents are randomly chosen. Arrows indicate edge connections for the control law.

**Data and Simulation Plots**

**Figure 3-6:** Simulation plots of robot trajectories for 0-3 malicious agents (5 sec, 100 Hz). Malicious agent strategy is a Brownian motion with velocity that is similar to the average of good agent velocities with the same noise covariance.

**Results**

**Classifier Performance**
- Precision: 0.959
- Recall: 0.961
- Specificity: 0.878
- Balanced Accuracy: 0.920
- Average Precision: 0.949

**Conclusion and Future Work**

The malicious agent classifier yielded fairly good results in terms of precision and recall with the imbalanced dataset. This is expected since LSTMs are designed for learning temporal relationships, and the FFT position information was extremely useful in showing correlation among good agents and among bad agents. Future work involves extending the malicious agent strategies and classifying malicious agents with their respective strategies.