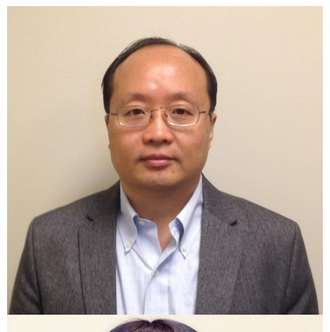
NSF SaTC: CORE: Small: Collaborative: A Framework for Enhancing the Resilience of Cyber Attack Classification and Clustering Mechanisms



NSF SaTC #2122631

PI: Prof. Shouhuai Xu (University of Colorado Colorado Springs; sxu@uccs.edu; https://xu-lab.org/)

Co-PI: Prof. Yanfang Ye (University of Notre Dame)

Problem Statement:

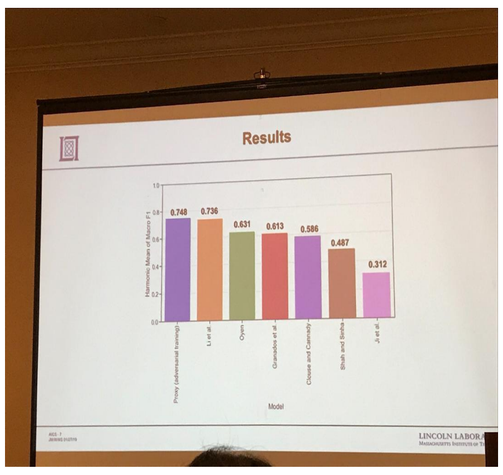
- Adversarial intelligent malware detection is an important problem that is little understood.
- How can we detect adversarial intelligent malware examples which are designed to evade AI/Machine Learning-based malware detectors?
- Can the resulting concepts and solutions be adapted to other AI/Machine Learning-based cybersecurity applications?

Challenges:

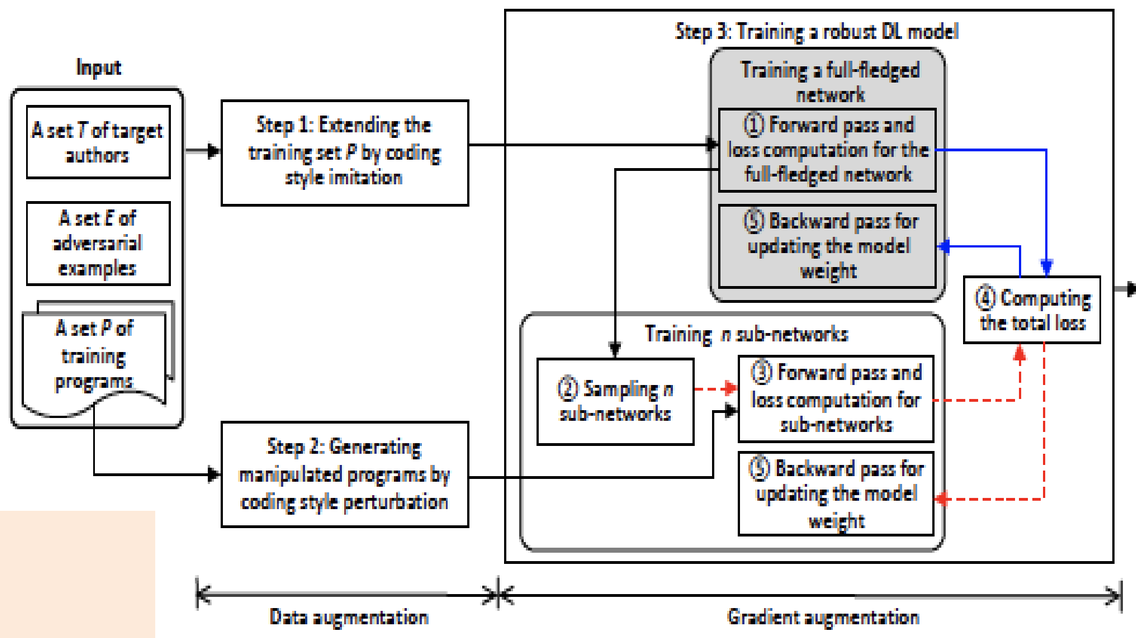
- How can we **quantify** the vulnerability and resilience of classification and clustering mechanisms against adversarial intelligent cyber evasion attacks, especially Adversarial Malware Detection?
- How can we **enhance** the resilience of classification and clustering mechanisms against adversarial intelligent cyber evasion attacks, especially adversarial malware?

Scientific Impacts:

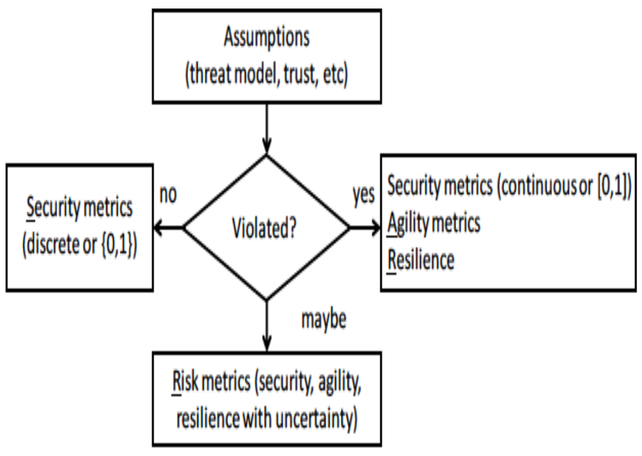
- Key innovations: Systematic (black-, gray-, white-box) threat model quantification → deeper understanding of the enemy → better defense
- A systematic defense approach: Defense principles → Framework → Metrics → Effective Mechanisms
- These results were investigated mainly in the context of Adversarial Malware Detection, but can be adapted to other cybersecurity contexts of Adversarial Machine Learning



Example Result 1: An adversarial malware classification framework won world-wide challenge organized by MIT Lincoln Lab [AICS'2019, IEEE TNSE 2021]



Example Result 2: Using data & gradient augmentations to learn robust models to defend against adversarial code authorship attribution [ICSE'2022]



Example Result 3: Unified metrics framework for quantifying cybersecurity (security, agility, resilience, risk) [SciSec'2021 Keynote]

Example Result 4: [ACSAC'2021]

- Q: Can we leverage predictive uncertainty to detect (dataset shift and) adversarial examples in Android malware detection?
- Idea: Quantify confidence (or uncertainty) associated with labels via model calibration
- A: Negative (adversarial examples can render calibration useless and so predictive uncertainty)
- Conjecture: Identify robust features

Broader Impacts:

- Safer AI/Machine Learning to make cyberspace more secure and trustworthy
- Potential transition to practice
- 30+ publications (including ICSE'22, ACSAC'21 IJCAI'19, AAAI'19, WWW'19, ACSAC'18)
- 3 outreach talks
- 3+ PhD Dissertations

Broad Participation:

- 7 PhD students participated (3 graduated so far), including 2 students from underrepresented groups
- 5 undergraduate participants
- 1 High School participant
- 10+ seminar/invited presentations
- 3 courses leveraged research materials