Autonomous Operation in Contested Environments

Saurabh Bagchi
ECE & CS, Purdue University
Center for Resilient Infrastructures, Systems, and Processes (CRISP)

Joint work with:

ECE: Mustafa Abdallah, Parinaz Naghizadeh, Shreyas Sundaram
Economics: Tim Cason, Daniel Woods

Supported by NSF SaTC grant CNS-1718637 (2018-21), Army Research Lab A2I2 Institute, Sandia National Lab

CRISP

PURDUE
UNIVERSITY

ML in Security

1. ML algorithms used in security tasks: common case
   - Spam detection, credit card fraud detection, …
2. Security of ML algorithms themselves: more recent but intense activity
   - Categorization based on temporal characteristic of attack or attacker knowledge
   - Categorization 1: Training time[1,2] versus test time[3]
   - Categorization 2: Model knowledge by attacker

Bibliography at the end of the slide deck

Some Types of ML Attacks

- Evasion attacks
- Poisoning attacks
- AML in Deep Neural Networks

Evasion Attacks

- **Adversary who previously chose instance x (which is now classifie… …oses another instance x' w… …gn**

1.

From: spammer@example.com
Cheap mortgage now!!!

Feature Weights

2. cheap = 1.0
mortgage = 1.5

3. Total score = 2.5 > 1.0 (threshold)

Spam

Evasion Attacks

1. From: spammer@example.com
Cheap mortgage now!!!
Mt. Hood Oregon

Feature Weights

2. cheap = 1.0
mortgage = 1.5
Mt. Hood = -1.0
Oregon = -1.0

3. Total score = 0.5 < 1.0 (threshold)

OK

Modeling Evasion Attacks

- Attacker has an "ideal" feature vector x_{ideal}
  - These are the original malicious feature vectors in training data
- Modifying x into another feature vector x' incurs a cost $C(x_{ideal}, x')$
- The attacker's goal is to appear "benign" to the classifier
- Observation: **Feature space modeling**
  - Attacker can make arbitrary changes to features
  - Cost is meant to capture constraints faced by the attacker

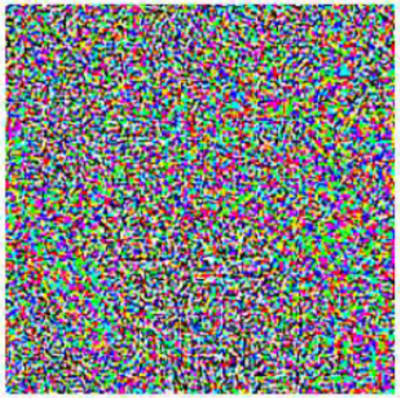
Slide from Yevgeniy Vorobeychik, AAAI 2018

Attacker Knowledge

- **Black-box attacks:** Attacks that fool a target model by adversarial examples made on a substitute model.
  - Adversaries do not know internal parameters of target model
  - However, using the same training data set, they can train their own DNN model; Can construct gradients of the target model with high similarity
- **White-box attacks:** Attacks that attempt to mislead the target model using the adversarial examples crafted on the target model itself
  - Adversaries are assumed to have access to the target model
  - Can compute the gradients of the target.

Adversarial Examples

$+ .007 \times$



$=$

NN prediction: Panda (70%)

NN prediction: Gibbon (99%)

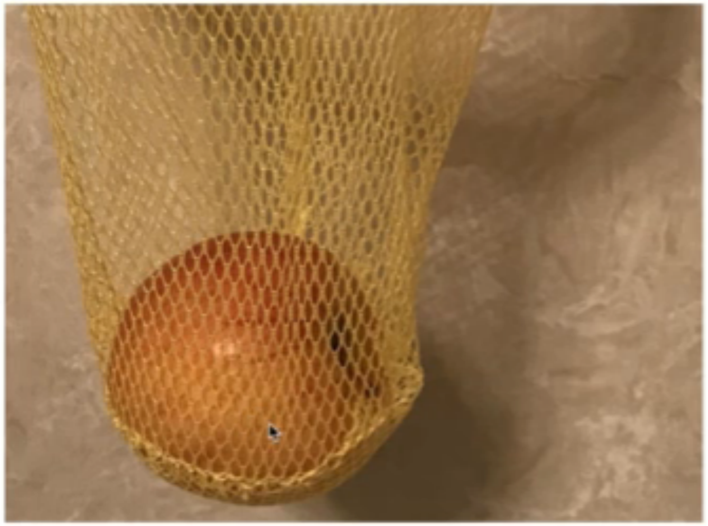
Training: $X \rightarrow \theta$

Inference: $\theta_x \rightarrow y$

Inference under attack: $\theta_{x'} \rightarrow y'$

Adversarial Examples in the Physical World

(Eykholt et al, 2017)



(Goodfellow 2018)

- **AE Transferability:** It was shown in [Goodfellow-NIPS14] that AEs crafted to mislead a DNN often also mislead a substitute model of the DNN

CRISP

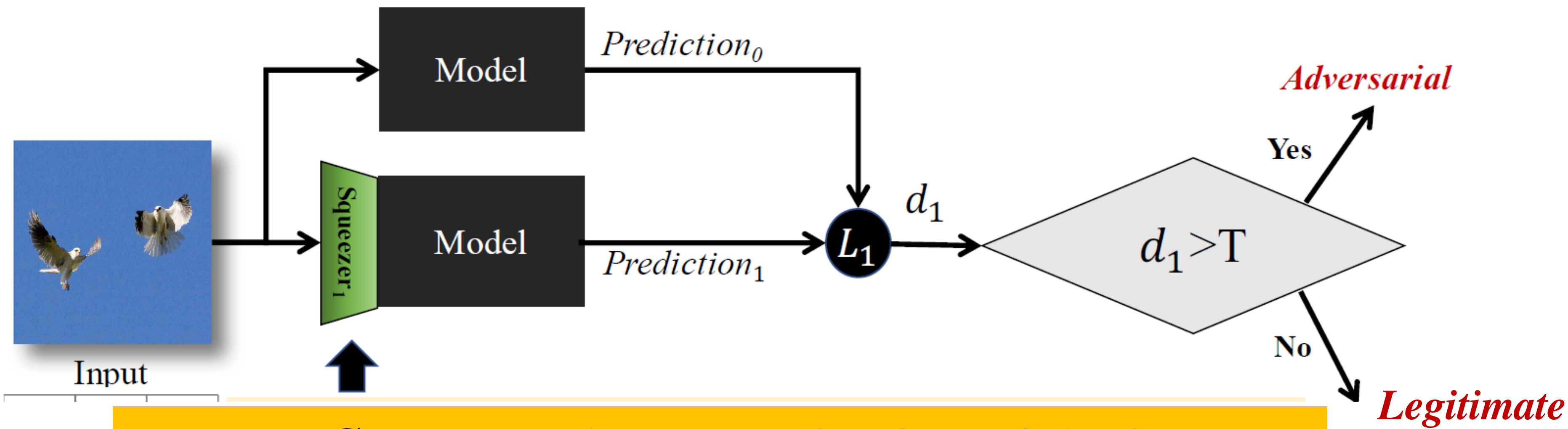
PURDUE UNIVERSITY

Some Ideas for Defense

1. **Adversarial training:** Proactively generating adversarial examples as part of the training procedure
   - Activity in efficiently generating lots of adversarial examples by perturbing actual data points
   - Model is then trained to assign the same label to the adversarial example as to the original example
2. **Defensive distillation:** Smooths the model's decision surface in adversarial directions exploited by the adversary
   - Distillation is a training method where one model is trained to predict probabilities output by another model that was trained earlier
   - First model is trained with "hard" labels (100% probability that an image is a dog rather than a cat) and then provides "soft" labels (95% probability that an image is a dog rather than a cat) used to train the second model
   - The second "distilled" model is more robust to attacks

Latest Defense against Adversarial Examples

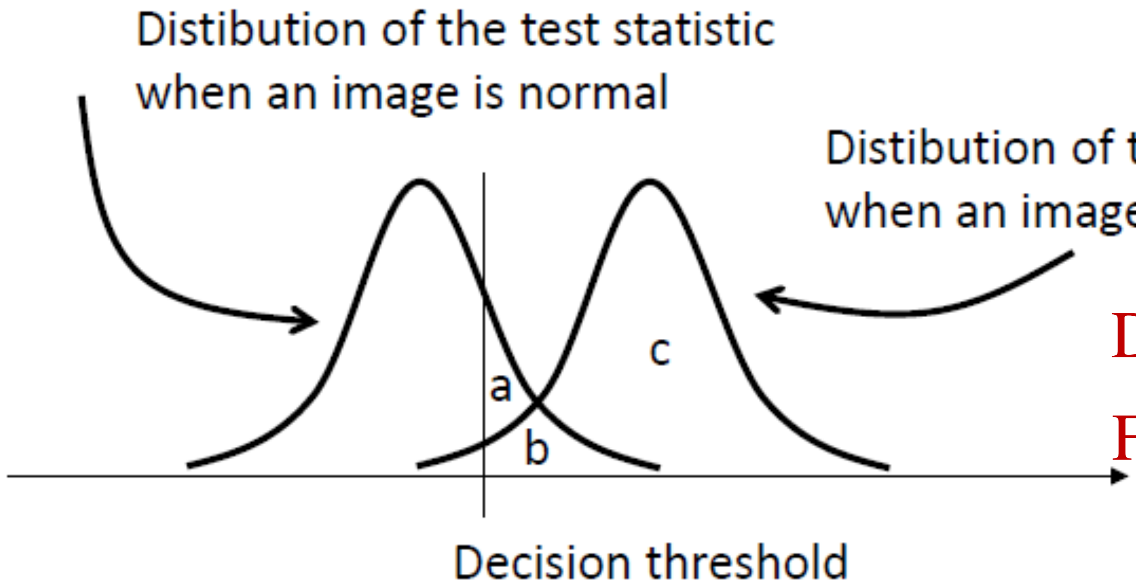
- **Feature Squeezing**: [Xu-Evans-Qi-NDSS18]
- Detect AEs rather than making model robust to AEs



Feature Squeezer does quantization of the image.
- Barely change legitimate input.
- Destruct adversarial perturbations

But the Arms Race Goes On

- Feature Squeezing's decision threshold needs to be fixed targeting a particular perturbation level
  - It performs poorly for perturbation levels that the threshold is not targeted for

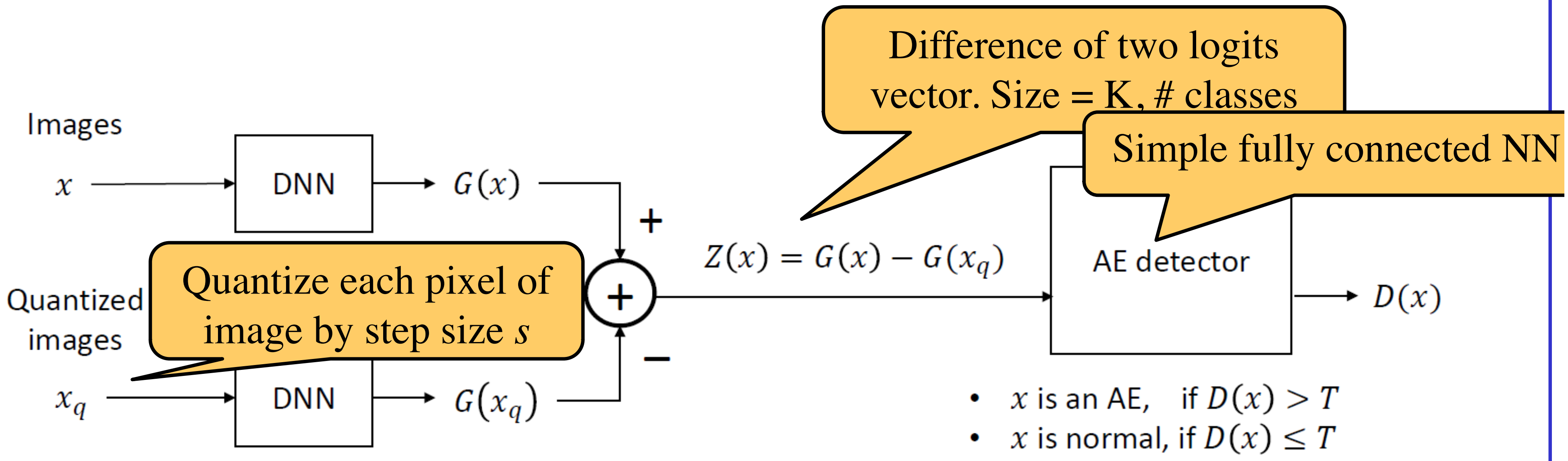


Detection Rate (DR) = $b+c$

False Positive Rate (FPR) = $a+b$

One Possible Solution

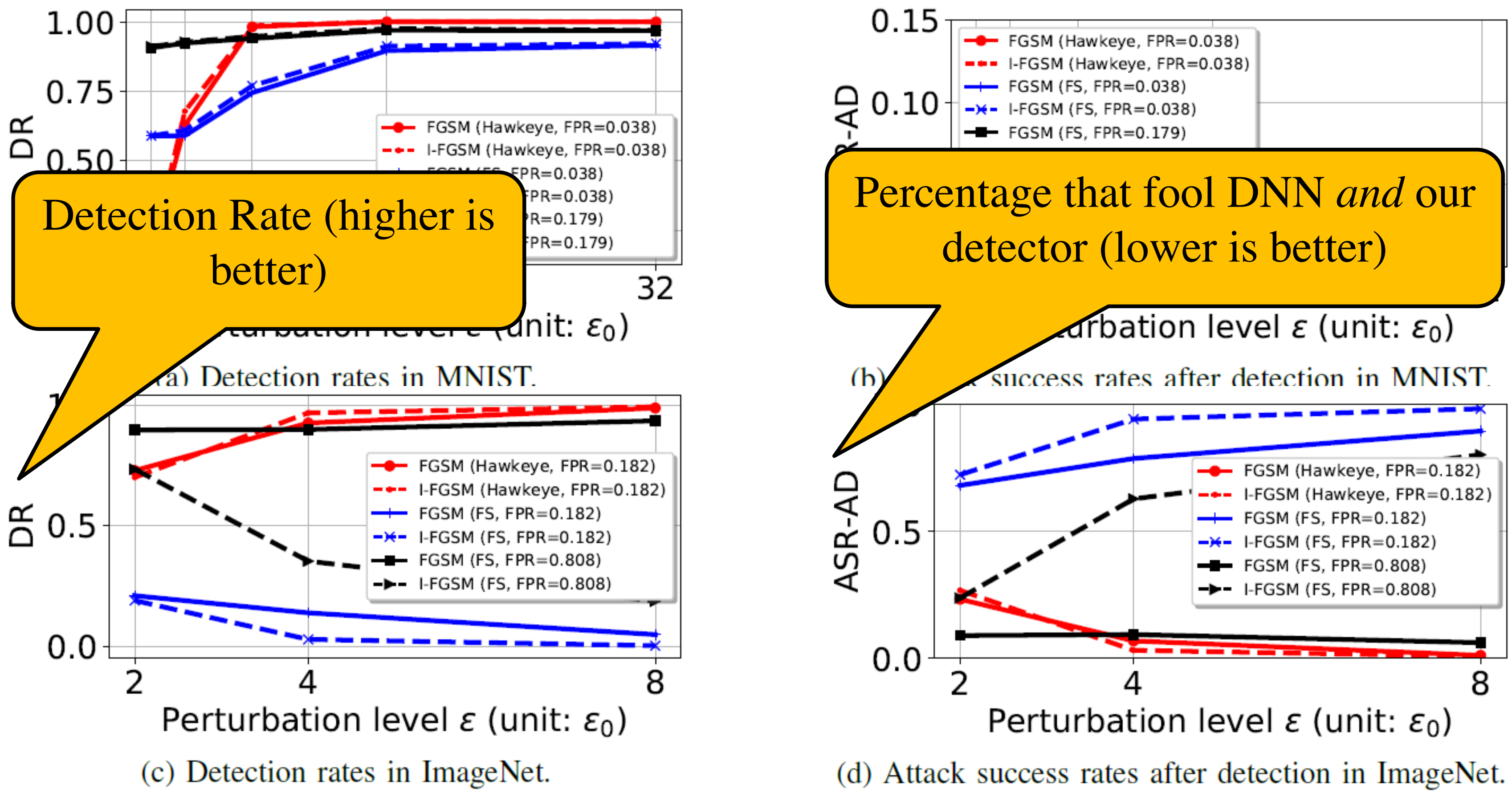
- Fundamentally, the drawback of FS is that there is a rigid mapping of perturbation level used to generate AE and L1 norm threshold
- We show that using a richer detector can lead to more precise detection across a wide range of perturbation levels
- For a given image x, we consider a quantized image x_q, which is made by quantizing each pixel of x with step size s



CRISP

PURDUE UNIVERSITY

Preliminary Evaluation



(a) Detection rates in MNIST.

(b) Attack success rates after detection in MNIST.

(c) Detection rates in ImageNet.

(d) Attack success rates after detection in ImageNet.

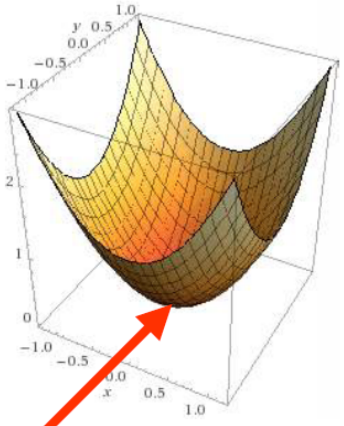
- HAWKEYE achieves a much lower ASR-AD than FS
- Even though DR at low perturbation level is not high, but it is not a big issue in terms of ASR-AD

Open Research Problems

- How is performance to “natural faults”
  - Examples: Brightness-reduced images *(simulating images taken at night time)*, occlusion by a noise box *(simulating an attacker or a water drop potentially blocking some parts of a camera)*, and occlusion by multiple tiny black dots *(simulating dirt on camera lens)*
- How can this class of techniques be used together with gradient-masking defenses that have been discredited in general, but often work well for low perturbation level attacks?
- Fundamentally, it is hard to defend against Adversarial Examples because it is hard to construct a theoretical model of the AE crafting process
  - AEs are solutions to an optimization problem that is non-linear and non-convex for many ML models
  - Because we don’t have good theoretical tools for describing the solutions to these complicated optimization problems, it is very hard to make any kind of theoretical argument that a defense will rule out a set of AEs

Game Theoretic View of Adversarial ML

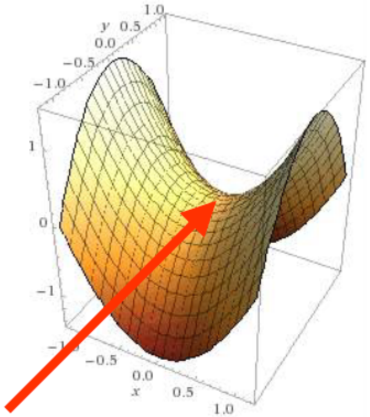
Traditional ML:
optimization



Minimum

One player,
one cost

Adversarial ML:
game theory



Equilibrium

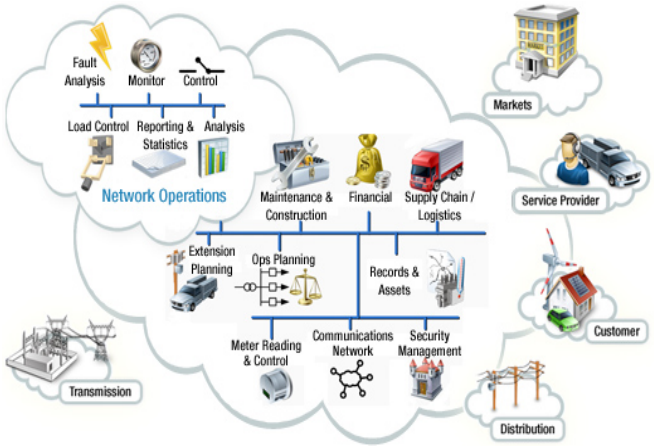
More than one player,
more than one cost

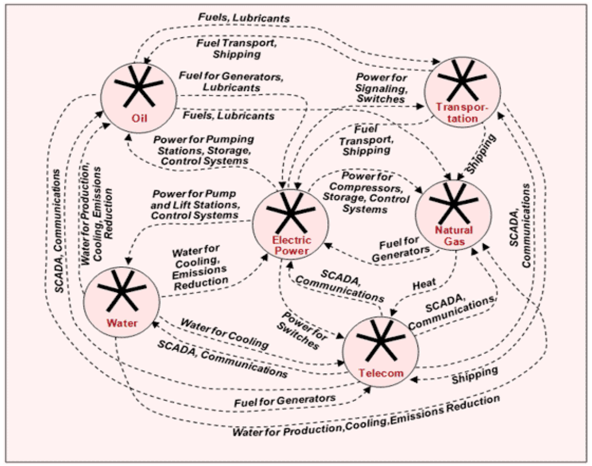
Defender: Minimize the maximum damage that can be inflicted by an adversary

Slide from Ian Goodfellow, 2018

Real-world Problem Context

- Modern critical infrastructures have a large number of assets, managed by multiple stakeholders.
- The security of these complex systems depends critically on the interdependencies between these assets.





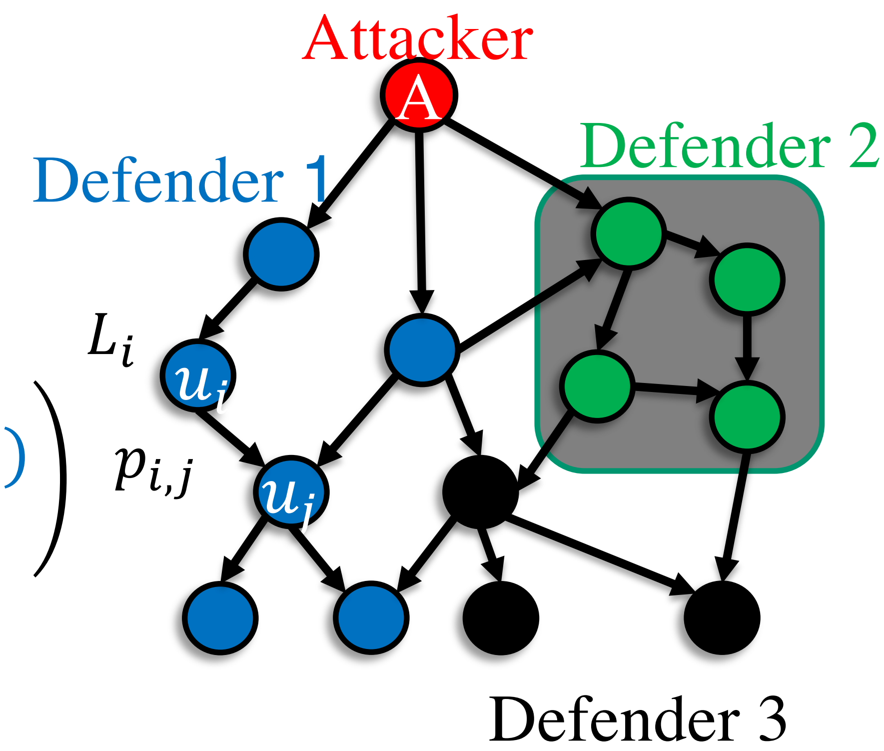
Goal: Create optimal and strategic allocation of defense resources in interdependent large-scale networks.
Tools: Machine Learning and Game Theory

Our Research Direction

- **Game-theoretic framework** involving attack graph models of large-scale interdependent systems and multiple defenders
- Each **human** defender misperceives the probabilities of successful attack in the attack graph
- We characterize impacts of such misperceptions on the security investments made by each defender

- The cost of a defender D_k is:

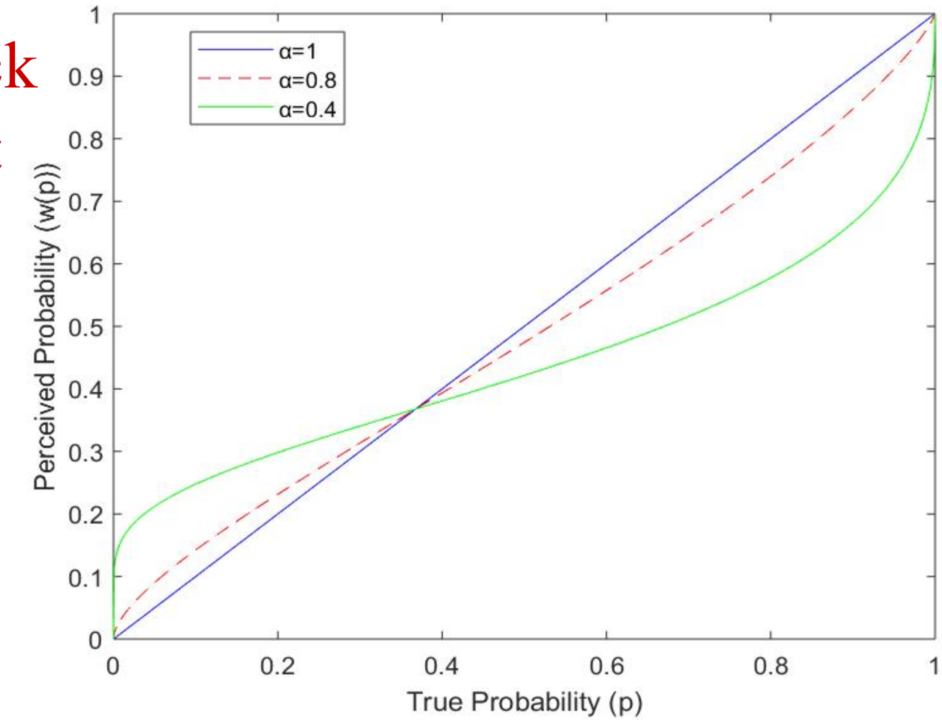
$$C_k(\mathrm{x}) \triangleq \sum_{u_m \in V_k} L_m \left(\max_{P \in \mathbb{P}_m} \prod_{(u_i,u_j) \in P} \boldsymbol{w}(p_{i,j}(\mathrm{x})) \right)$$



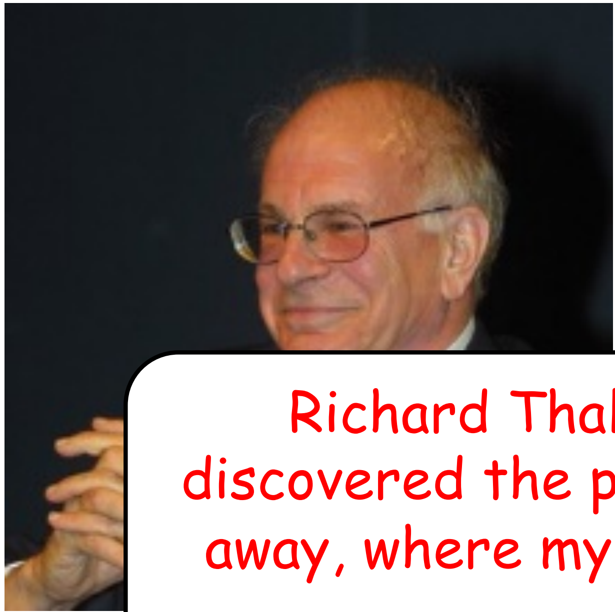
Behavioral Weighting Function

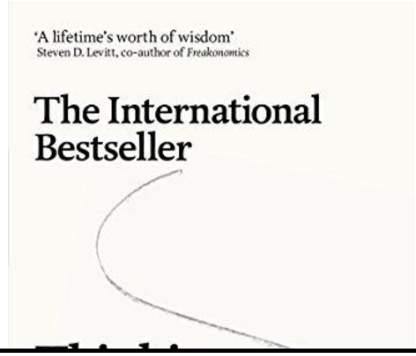
- Human perceptions of rewards and losses can differ substantially from their true values
- These perceptions can have a significant impact on the investments made to protect the systems that the individuals are managing.

- Humans overweight low attack probabilities and underweight large attack probabilities.
- Example: Prelec [1998] weighting function:
- $w(x) = \exp(-(-\ln(x))^{\alpha})$
- where parameter $\alpha \in (0,1]$.



What's Nobel Got to Do With It?





Richard Thaler (2017 Economics Nobel Laureate): "I discovered the presence of human life in a place not far, far away, where my fellow economists thought it did not exist: the economy."

a counterpoint to expected utility theory

Some Definitions

- Behavioral defender (colloquially "biased defender"): Makes security investment decisions under cognitive biases
  - Using prospect-theoretic, non-linear probability weighting models, they misperceive probabilities of a successful attack on edges of the attack graph
- Non-behavioral (colloquially "rational defender"): Makes security investment decisions based on the classical models of fully rational decision making
  - Correctly perceives the risk on each edge within the attack graph of the CPS network, and chooses investments accordingly
- Why do we need to consider human cognitive biases in security decision making?
  - Significant investments in security controls, security policies, or changes in the system architecture involve human decision making
  - One player may have partial observability of other player's actions
  - Deception may be used to create mis-perception of attack-defense successes

Optimization Problem Formulation

- The probability of successfully compromising v_j, starting from v_i, is given by

$$p_{i,j}(x_{i,j}) = p_{i,j}^0 \exp\left(-s_{i,j} \sum_{D_k \in D \text{ s.t. } (v_i, v_j) \in \mathcal{E}_k} x_{i,j}^k\right)$$

- A behavioral defender D_k chooses her investments $x_{i,j}^k$ to minimize her *perceived* loss

$$C_k(\mathbf{x}) = \sum_{v_m \in V_k} L_m \left(\max_{P \in P_m} \prod_{(v_i, v_j) \in P} w\left(p_{i,j}(x_{i,j})\right) \right)$$

- The probability weighting function $w(p)$ gives how humans mis-perceive true probability p
  - For example: a commonly believed functional form is the Prelec form where $\alpha \in (0, 1]$ determines the degree of mis-perception

$$w(p) = \exp\left[-(-\log(p))^{\alpha}\right].$$

Break for Games

http://ifipdemo.herokuapp.com/

Network Red

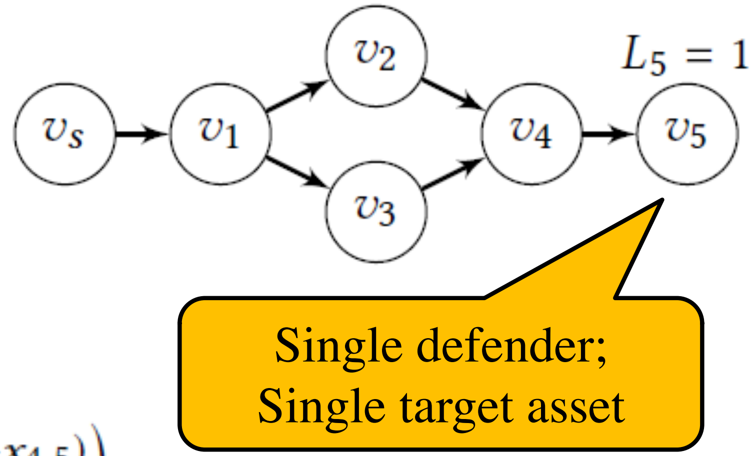
Network Blue

Follow: "Session-wide link

Open the below link in up to 1 browser tabs"

Intuition for Behavioral vs. Non-behavioral Decisions

- **Min-cut of a graph:** Given two assets *s* and *t* in the graph, an edge-cut is a set of edges E_c such that removing them from the graph removes *all* paths from *s* to *t;* A min-cut is an edge-cut of smallest cardinality over all possible edge-cuts
- Two possible min-cuts: (v_s, v_1), (v_4, v_5)
- Total loss function for the defender



$$C(x) = \max\left(e^{-(x_{s,1}+x_{1,2}+x_{2,4}+x_{4,5})}, e^{-(x_{s,1}+x_{1,3}+x_{3,4}+x_{4,5})}\right)$$

- **Theorem:** One can prove (using the KKT conditions of non-linear programming) that it is optimal for a non-behavioral defender to put all of her budget only on the min-cut edges, i.e., any solution satisfying $x_{s,1} + x_{4,5} = B$
  - Optimal investment leads to a loss of e^{-B}
- For the behavioral defender total loss function is:

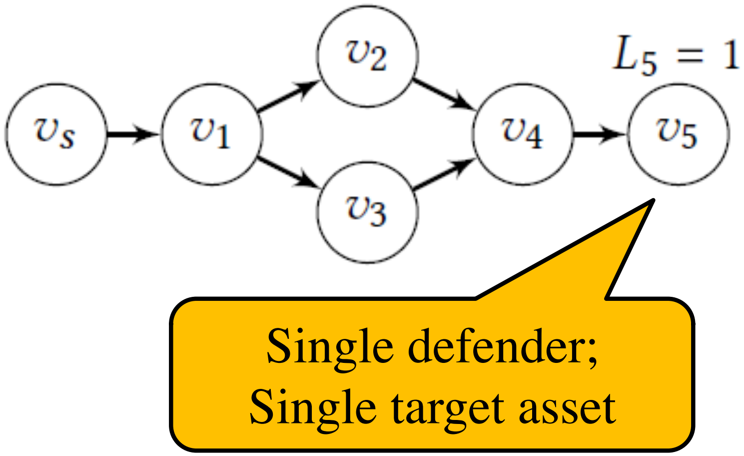
$$\min_{x} \max\left(e^{-x_{s,1}^{\alpha}-x_{1,2}^{\alpha}-x_{2,4}^{\alpha}-x_{4,5}^{\alpha}}, \ e^{-x_{s,1}^{\alpha}-x_{1,3}^{\alpha}-x_{3,4}^{\alpha}-x_{4,5}^{\alpha}}\right)$$

Intuition for Behavioral vs. Non-behavioral Decisions

- Optimal investment by behavioral defender:

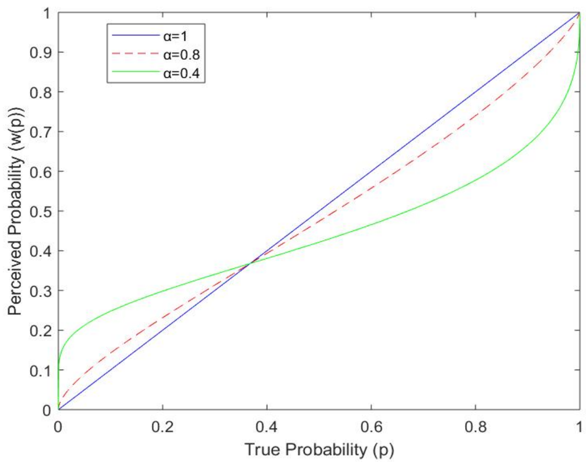
$$x_{1,2} = x_{2,4} = x_{1,3} = x_{3,4} = 2^{\frac{1}{\alpha-1}} x_{s,1}.$$

$$x_{s,1} = x_{4,5} = \frac{B-4x_{1,2}}{2} = \frac{B}{2+4(2^{\frac{1}{\alpha-1}})}.$$



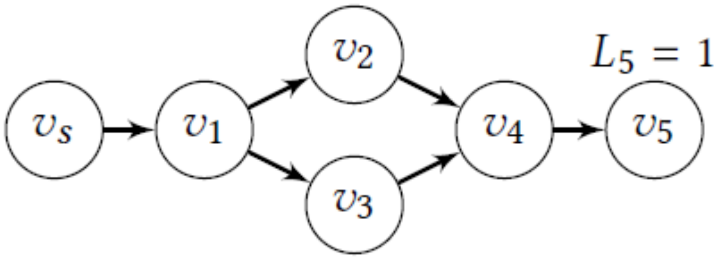
- There are investments on non-min-cut edges

  - **Loss for behavioral defender > Loss for non-behavioral defender**
  - Why this behavior?
    - When considering an undefended edge, the marginal reduction of attack probability on that edge as *perceived* by a behavioral defender is much **larger** than the marginal reduction of true attack probability
    - Thus the behavioral defender is incentivized to invest some non-zero amount on that edge

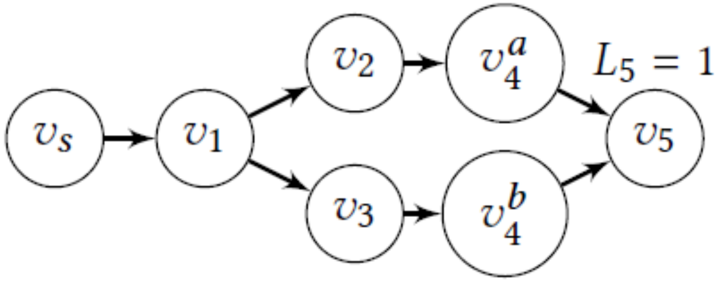


Other Modeling Factors

- Multi-hop dependence



(a) A baseline attack graph.

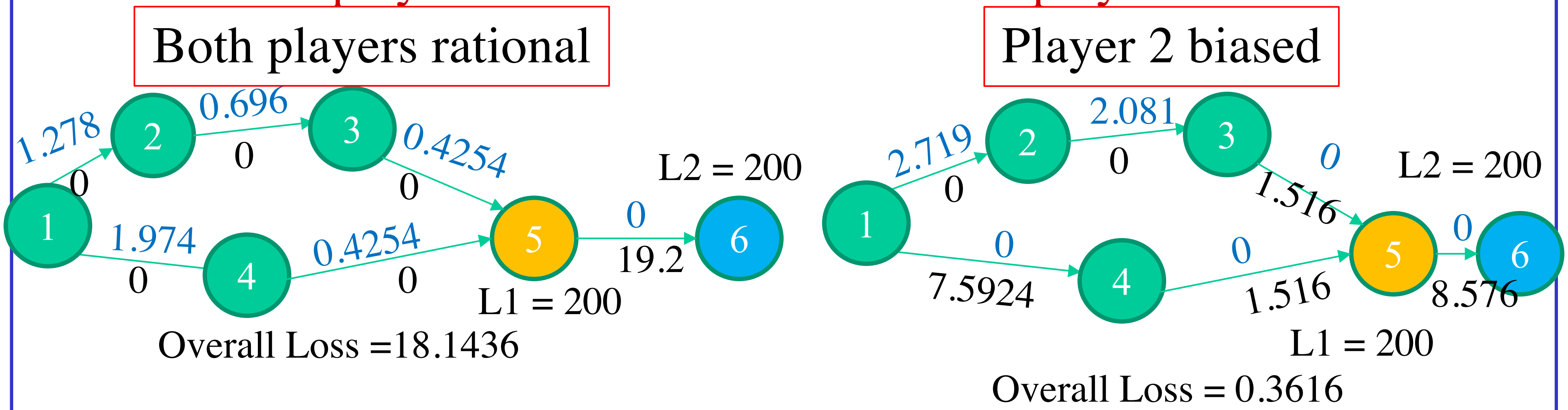


(b) An attack graph created from (a) if the nodes have two-hop dependencies.

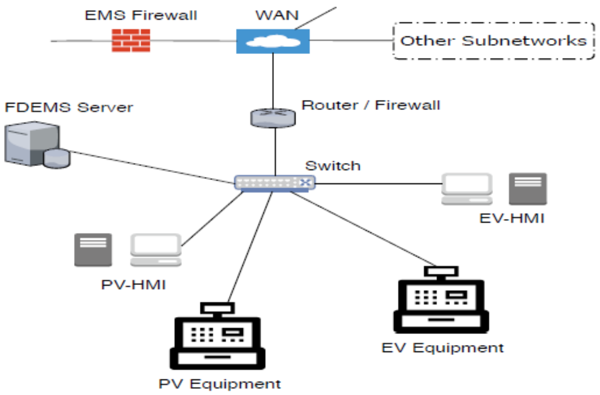
- Spreading behavior of security investments
  - Behavioral defender spreads her defensive investments on all edges throughout the attack graph
  - Solution approach: For each defender D_k, we set $x_{i,j}^k \geq \eta_k$
- Misperception due to information asymmetry or deception
  - **Hypergames** extend the classical game theory model by incorporating the *perception* of each player in the game analysis
  - Solution approach: We show hypergames is a valuable game-theoretic model to analyze how to use deception to increase security of inter-dependent systems

Initial Observations

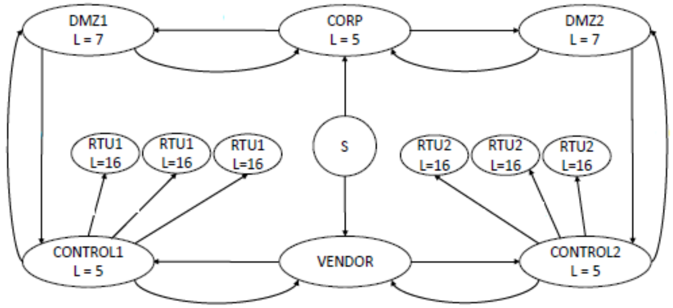
- Both games (vertex based and path based) have **Convex cost function** given a convex decreasing probability function
- Both games have a **Pure Nash Equilibrium (PNE) state**
- In each game, we can compute the best response by solving a convex optimization problem
- They have **different investment decisions** than standard security game which maximizes expected utility
- A rational player **can benefit** from a biased player



Sample System Applications



A network level systen source (DER) system within th the DER.1 failure scenario, wh physical access attack to the HM (adapted from [32]).



The attack graph for a SCADA-based con- trol network, adapted from [27]. The attacker's start- e has an associated loss).

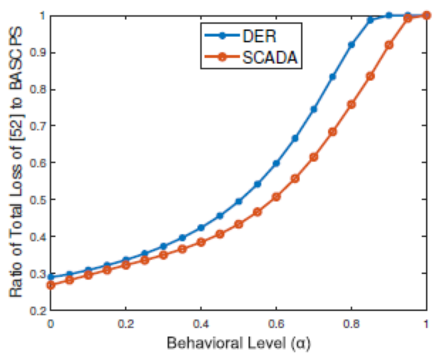


Figure 13: The ratio of loss estimated by [57] to the (true) loss estimated by BASCPS *for different behavioral levels, with* $\eta = 0$.

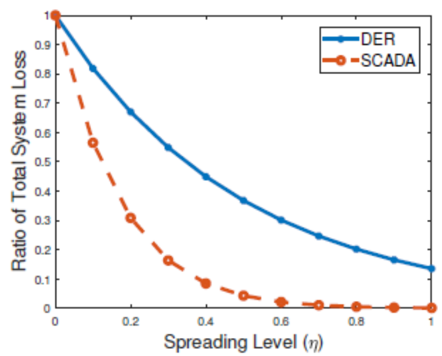
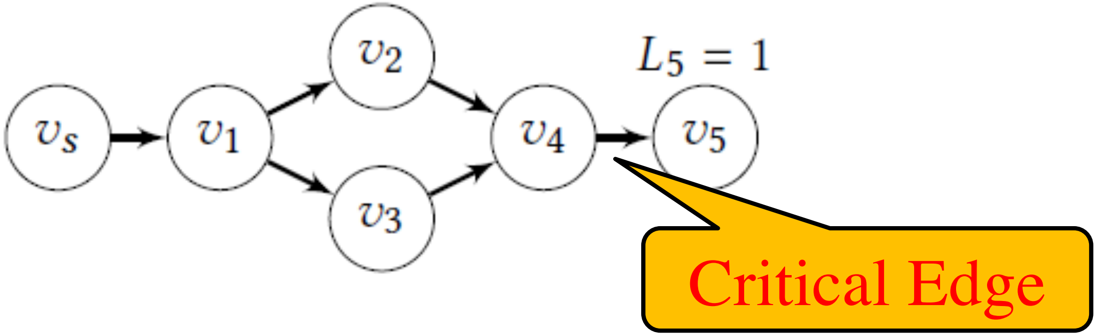


Figure 14: The ratio of loss estimated by [57] to the (true) loss estimated by BASCPS *for different spreading levels, with* $\alpha = 1$.

Insights about Behavioral Decision Making

System Parameter	Insights from Behavioral Decision Making
Defense Budget	The adverse effects of behavioral decision making are most severe with moderate defense budgets (Figure 10). In particular, at either extreme of sufficiently large or extremely limited budgets, the amount of the budget, rather than its allocation, is most crucial in determining the system's security, so the effects of behavioral decision making become secondary.
Interdependency	The impact of behavioral suboptimal decision making on the system is magnified as the degree of the interdependency between subnetworks belonging to different defenders increases (Figures 15, 19).
CPS Size	The impact of behavioral suboptimal decision making is magnified as the number of nodes in the CPS grows (Figures 11, 20).
Budget distribution	The negative effect of behavioral decision-making is more pronounced with asymmetric budgets among the defenders (Figures 12, 25).
Defense Mechanism	Selfish defense decisions together with behavioral decisions significantly increase security risk. Cooperative (or joint) defense among the defenders has the potential of overcoming the effects of suboptimal behavioral decision making. This even improves security outcomes over rational but selfish decision making (Figures 12, 21).
Central Planning	We compare the outcomes of decentralized decision making by individual defenders with those of investment decisions by a central planner, such as through a federal regulatory authority, tasked with minimizing social loss of the whole system. Central planning is most beneficial for improving CPS security when the defenders have a higher degree of behavioral bias and when the security budget is high (Figure 26).
Sensitivity	Behavioral decision making leads to investing less security resources on the parts of the network that are more sensitive to investments (i.e., probability of attack comes down faster with additional security investment) when there are few critical assets to be protected (Figure 16).

Human Subject Experiments



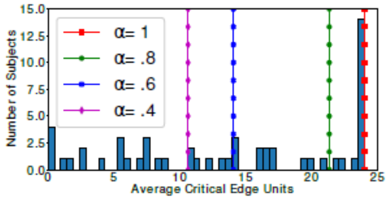


Figure 4: Histogram of human subjects' investments on the critical edge. The vertical red lines show the optimal allocations at specific behavioral levels (α).

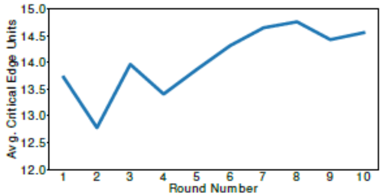


Figure 5: Average of all subjects' investments on the critical edge vs experiment rounds. The upward trend indicates that on average, subjects are learning.

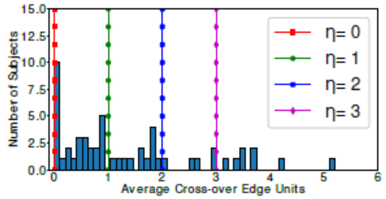


Figure 6: Histogram of human subjects' investments on the cross-over edge. The vertical red lines show the optimal allocations at specific spreading levels (η).

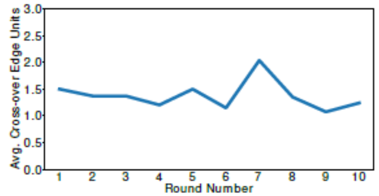


Figure 7: Average of all subjects' investments on the cross-over edge vs experiment rounds. There is only a weak downward trend in subjects' spreading behavior.

- Fully rational players tend to invest in min-cut edges
- Behavioral players also invest in non critical edges and have a spreading behavior

Take Aways and Open Challenges

- Adversarial ML algorithms need to be considered
  - To defend against malicious tampering of the model or the data
  - To protect against natural failures for high reliability scenarios: Autonomous vehicles, Air traffic control, Surgery robots, …
- Game theory can be applied to understand the effects of misperceptions, whether natural or maliciously induced
  - For inter-dependent systems, possibly with multiple defenders
  - Extensions to classical models needed
  - Behavioral game theory for handling misperceptions
  - Hypergame theory for handling different degrees of misinformation among players
- Open Challenges
  1. Laws of secure ML algorithms? Even under highly specific conditions
  2. Game theory being used to analyze dynamic scenarios. Respond in real-time.
  3. Induce beneficial misperception to lead to secure deployments.

Bibliography

1. Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. "Systematization of Knowledge: Towards the science of security and privacy in machine learning." IEEE Security and Privacy, 2016 (special category paper called Systematization of Knowledge). A great overview paper with a rich set of references and explanatory text for these references.
2. Kloft, Marius, and Pavel Laskov. "Online anomaly detection under adversarial impact." In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 405-412. 2010. Shows the impact of false data in training.
3. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proceedings of the 2014 International Conference on Learning Representations (ICLR). A systematic way to craft adversarial examples at runtime and reasoning about their effectiveness.

Our Papers

4. Gutierrez, Christopher N., Mohammed H. Almeshekah, Saurabh Bagchi, and Eugene H. Spafford. "A Hypergame Analysis for ErsatzPasswords." In IFIP International Conference on ICT Systems Security and Privacy Protection, pp. 47-61. Springer, Cham, 2018. How to apply game theory under misperception due to deceptive security practices.