Entanglement of Learning and Security: Strategic Learning for Adaptive Defense

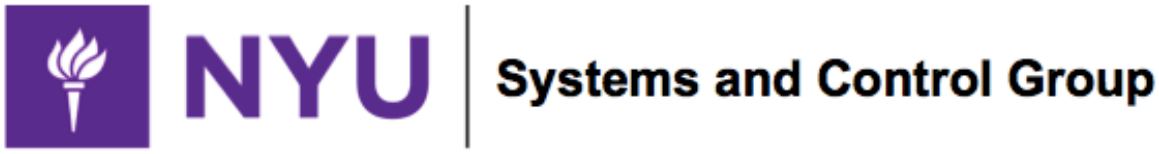
Quanyan Zhu

Tandon School of Engineering

New York University

CDC Workshop on Learning, Games, and Control for Security of Cyber-physical Systems

Dec. 10, 2019



Introduction

- Cyber defense needs to be proactive and adaptive to changing environment.
- There is need for scientific and engineering foundations for online learning for adaptive proactive defense.

- *Moving target defense:* Baseline learning algorithms
- *Proactive defense against APT:* learning of dynamic games with incomplete information
- *Attacker engagement problem:* reinforcement learning
- *Security of reinforcement learning*: deceptive RL

Game Theory Meets Network Security

A Tutorial

Stefan Rass and Quanyan Zhu

25th ACM Conference on Computer and Communications Security

October 15, 2018

UNIVERSITAET
KLAGENFURT

NYU

https://arxiv.org/abs/1808.08066

Game Theory Meets Network Security

A Tutorial

Stefan Rass and Quanyan Zhu

Game Theory for Cyber Deception

Quanyan Zhu

Symposium and Bootcamp on the Science of Security (HotSoS)

April 2-3, 2019

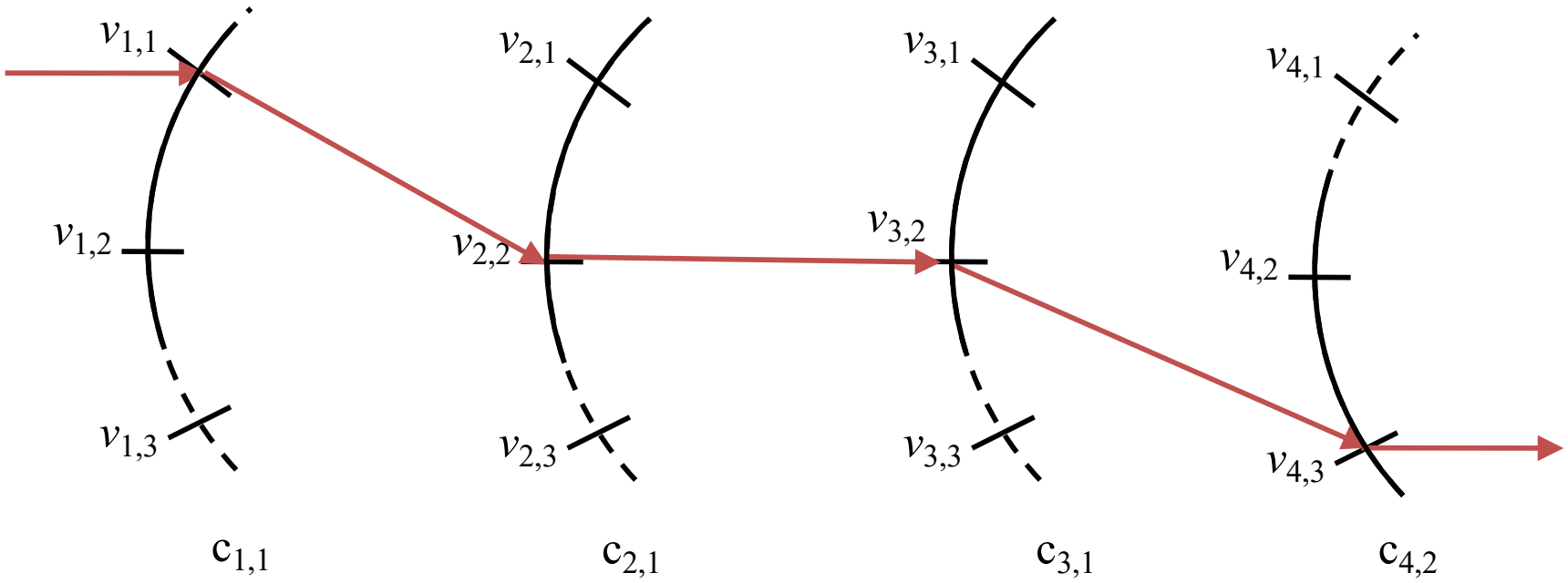
Nashville, Tennessee

NYU TANDON SCHOOL OF ENGINEERING

Laboratory for Agile and Resilient Complex Systems

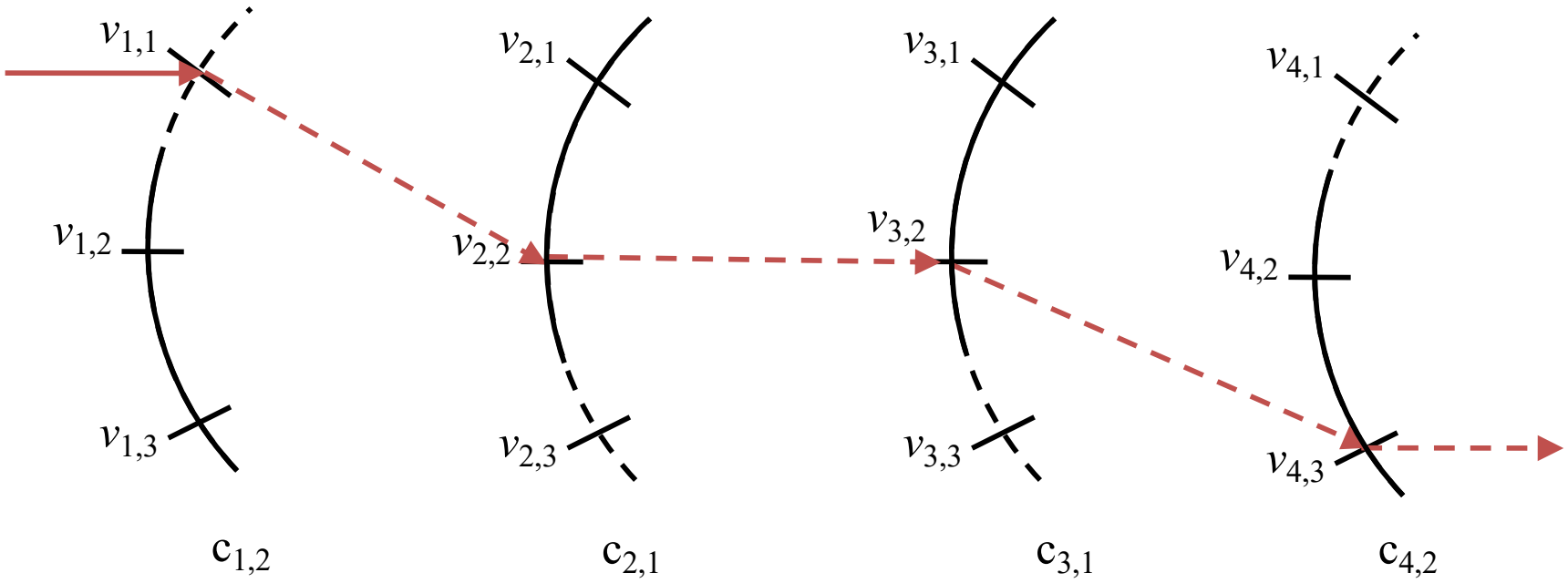
https://arxiv.org/abs/1903.01442

Abstraction



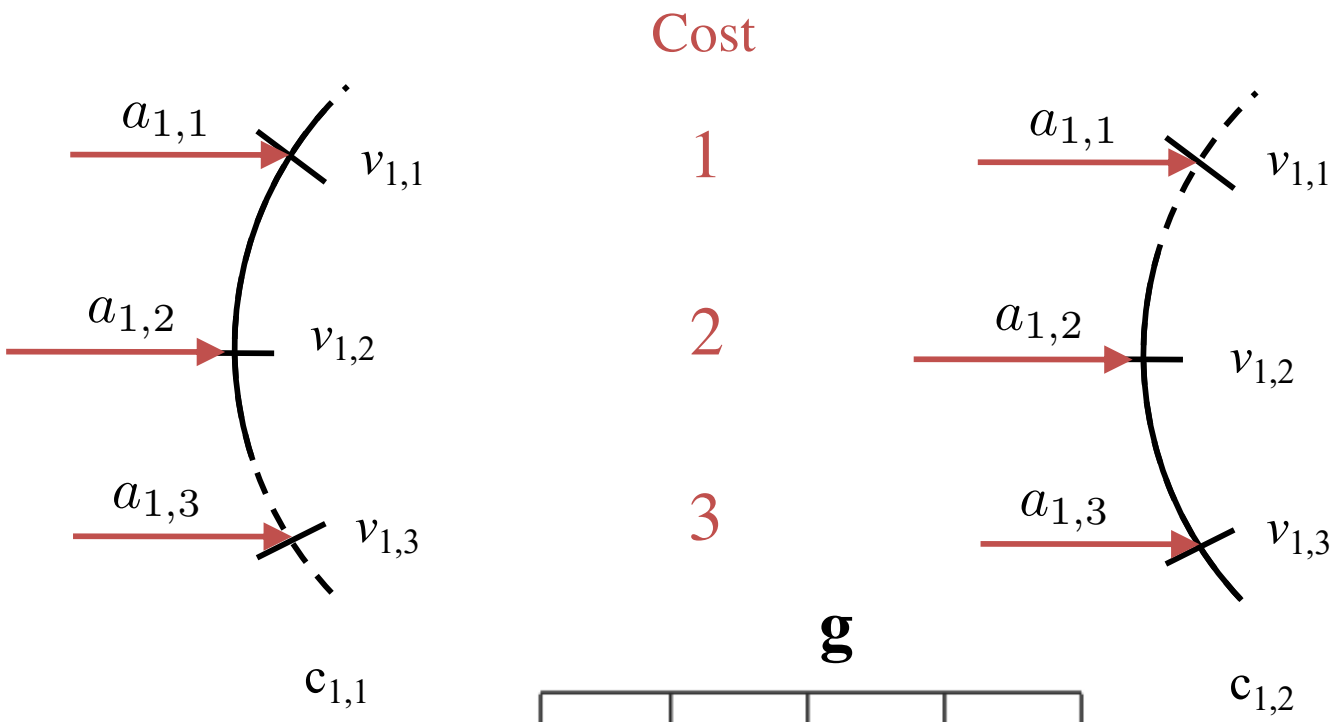
- The system has three vulnerabilities: $v_{1,1}$, $v_{1,2}$, $v_{1,3}$ at stage 1.
- The system is configured to $c_{1,1}$, which exhibits vulnerabilities $v_{1,1}$, $v_{1,2}$.
- The *attack surface* of the configuration is the set of vulnerabilities $\{v_{1,1}$, $v_{1,2}\}$.

Abstraction



- The system has three vulnerabilities: $v_{1,1}$, $v_{1,2}$, $v_{1,3}$ at stage 1.
- The system is configured to $c_{1,2}$, which exhibits vulnerabilities $v_{1,2}$, $v_{1,3}$.
- The *attack surface* of the configuration is the set of vulnerabilities $\{v_{1,2}$, $v_{1,3}\}$.

Complete Information Game-Theoretic Model



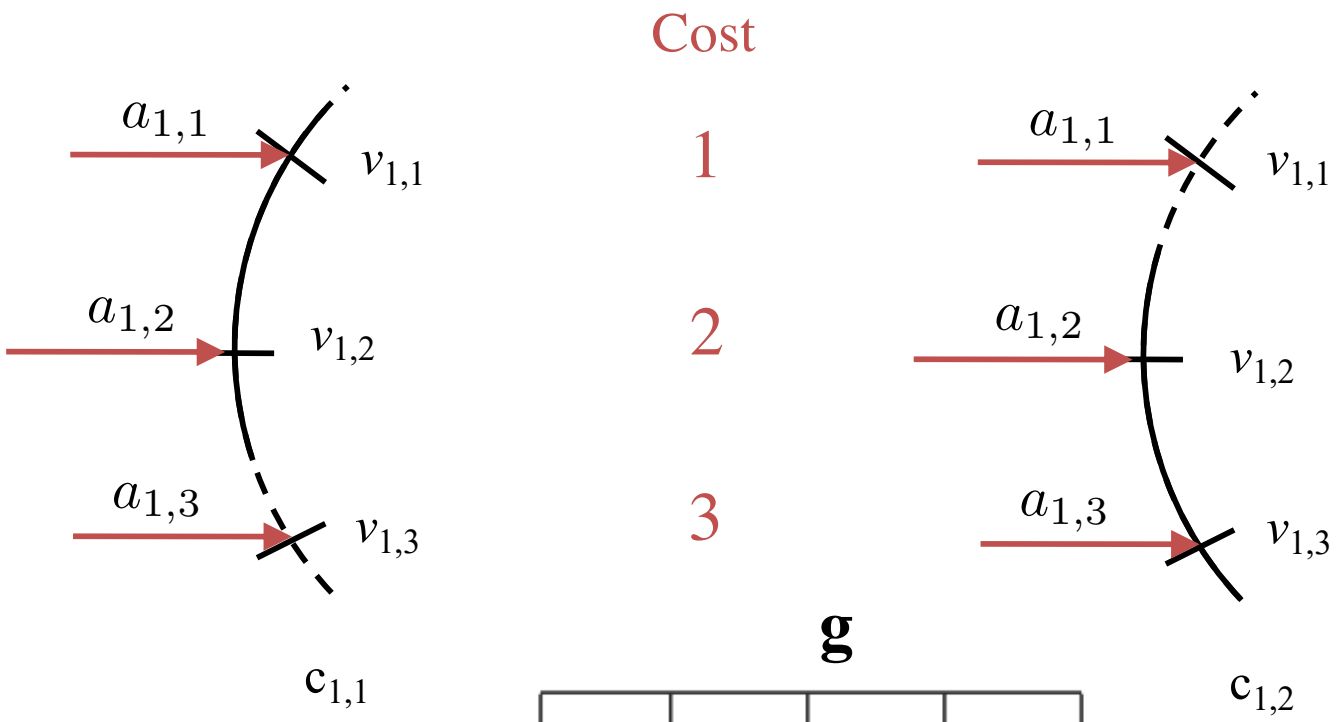
Ξ_1 : **f**

g

	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$
$c_{1,1}$	1	2	0
$c_{1,2}$	0	2	3

- **f** and **g** are mixed strategies. Mixed strategies are *randomizing* strategies.
- Nash equilibrium to the zero-sum game exists, and yields the worst-case defense strategy.

Not Sufficient Yet!



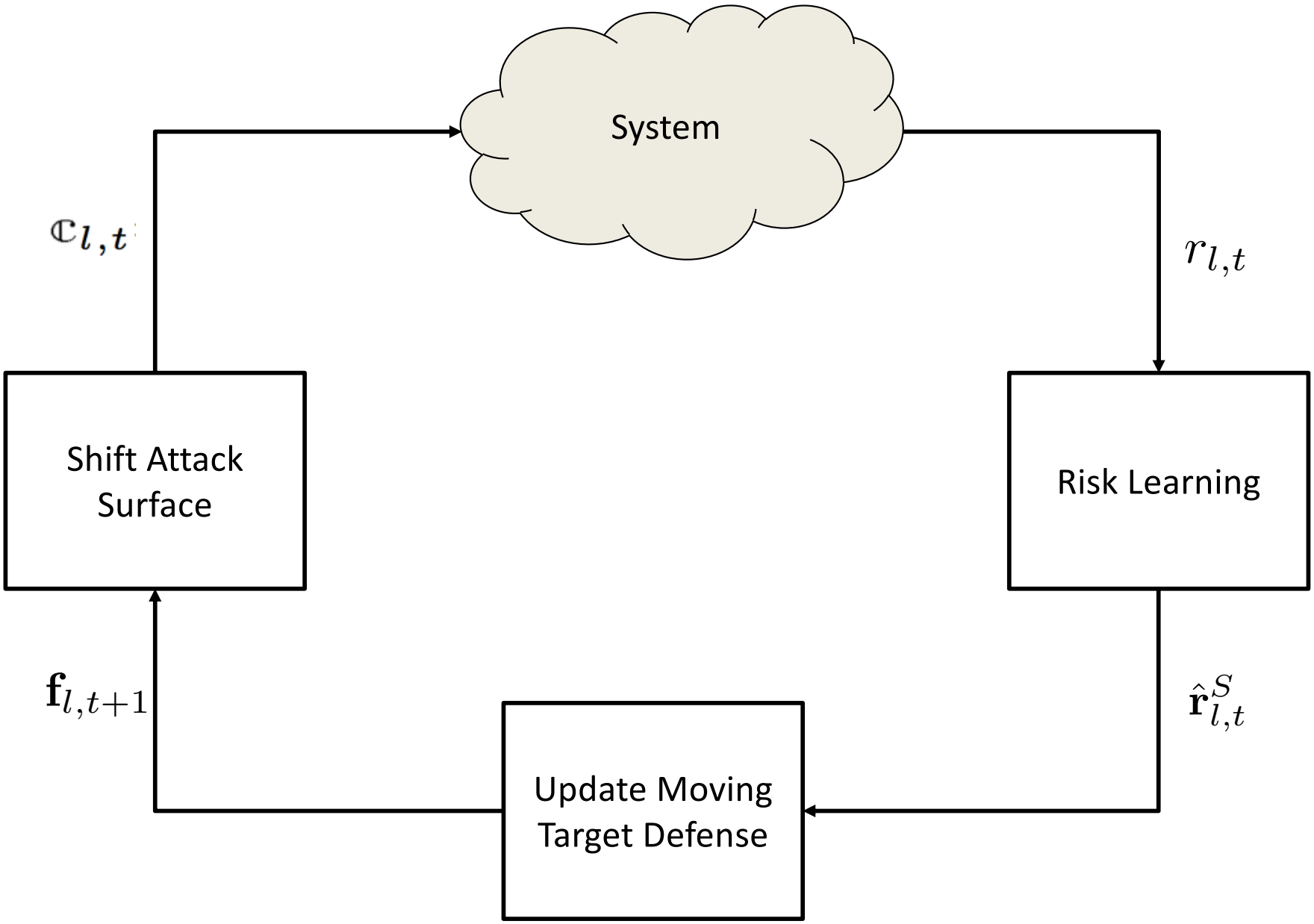
Ξ_1 : **f**

g

	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$
$c_{1,1}$	1	2	0
$c_{1,2}$	0	2	3

- **f** and **g** are mixed strategies. Mixed strategies are *randomizing* strategies.
- Nash equilibrium to the zero-sum game exists, and yields the worst-case defense strategy.

A Feedback System Model



Observed payoff of a chosen action

Estimated average payoff of a chosen action

$$\hat{r}^S_{l,t+1}(c_{l,h}) = \hat{r}^S_{l,t}(c_{l,h}) + \mu^S_t \mathbb{1}_{\{\mathbb{c}_{l,t}=c_{l,h}\}}(r_{l,t} - \hat{r}^S_{l,t}(c_{l,h})),$$

$$\hat{r}^A_{l,t+1}(a_{l,h}) = \hat{r}^A_{l,t}(a_{l,h}) + \mu^A_t \mathbb{1}_{\{\mathbb{a}_{l,t}=a_{l,h}\}}(r_{l,t} - \hat{r}^A_{l,t}(a_{l,h})).$$

- μ_t^S and μ_t^A are learning rates.
- Averaging over observed payoffs.

$$(\mathrm{SP}) \quad \sup_{\mathbf{f}_{l,t+1} \in \mathcal{F}_l} \underbrace{\langle \mathbf{f}_{l,t+1}, -\hat{\mathbf{r}}^S_{l,t} \rangle}_{} - \underbrace{\epsilon^S_{l,t} \sum_{h=1}^{m_l} f_{l,h,t+1} \ln \left(\frac{f_{l,h,t+1}}{f_{l,h,t}} \right)}_{}.$$

Average risk to be minimized

Relative entropy: Distance between two distributions

Cost on changing the strategy (usability)

$$(\mathrm{SP}) \qquad \sup_{\mathbf{f}_{l,t+1} \in \mathcal{F}_l} \underbrace{\langle \mathbf{f}_{l,t+1}, -\hat{\mathbf{r}}^S_{l,t} \rangle}_{} - \underbrace{\epsilon^S_{l,t} \sum_{h=1}^{m_l} f_{l,h,t+1} \ln \left(\frac{f_{l,h,t+1}}{f_{l,h,t}} \right)}_{}.$$

Average risk to be minimized

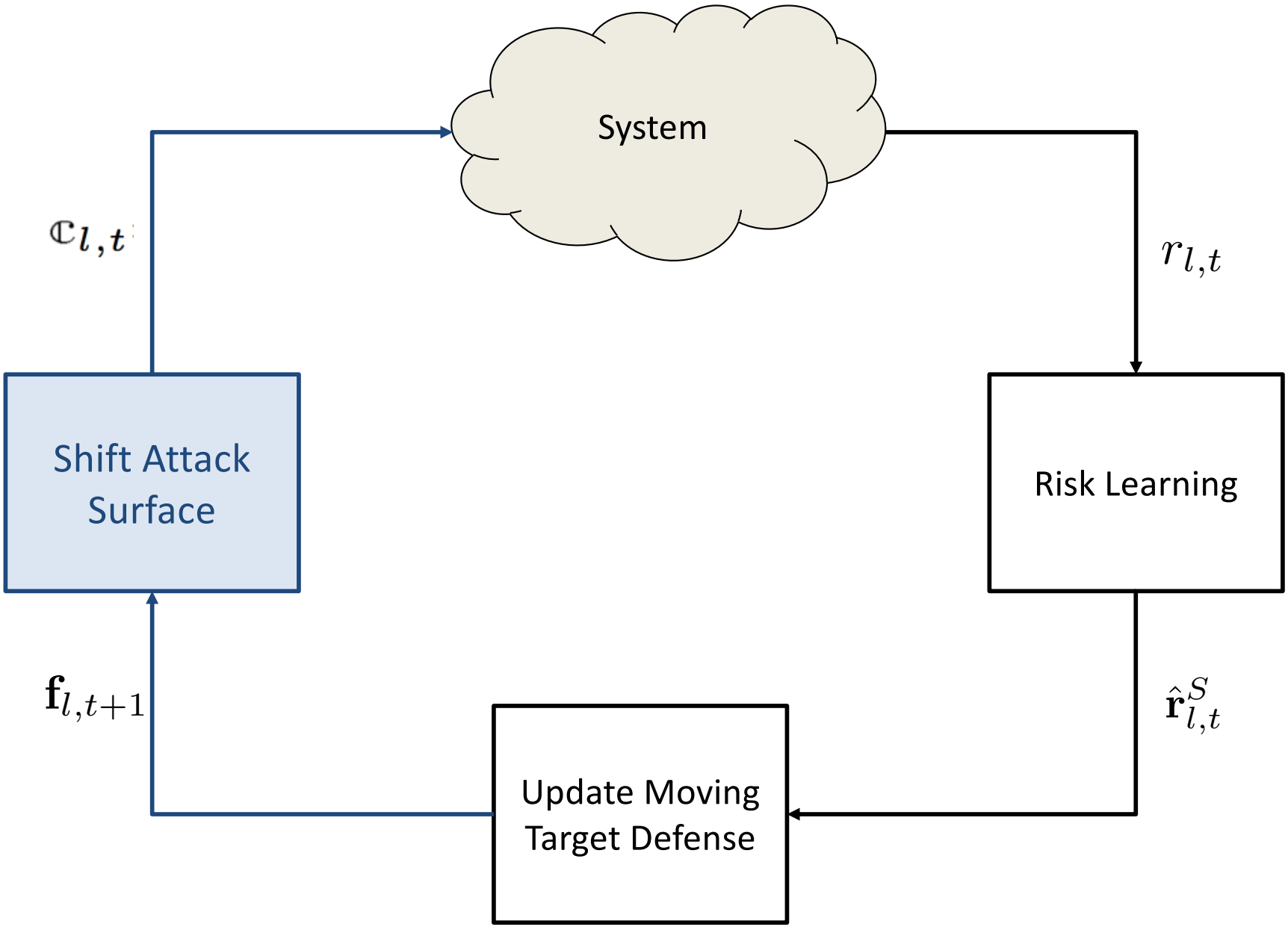
Relative entropy: Distance between two distributions

Cost on changing the strategy (usability)

$$f_{l,h,t+1} = \frac{f_{l,h,t} e^{-\frac{\hat{r}_{l,t}(c_{l,h})}{\epsilon^S_{l,t}}}}{\sum_{h'=1}^{m_l} f_{l,h',t} e^{-\frac{\hat{r}_{l,t}(c_{l,h'})}{\epsilon^S_{l,t}}}};$$

- High risk → low probability
- High ε → Costly to change → $f_{l,h,t}$ (Lower rationality)
- Low ε → Less costly to change (High Rationality)

A Feedback System Model



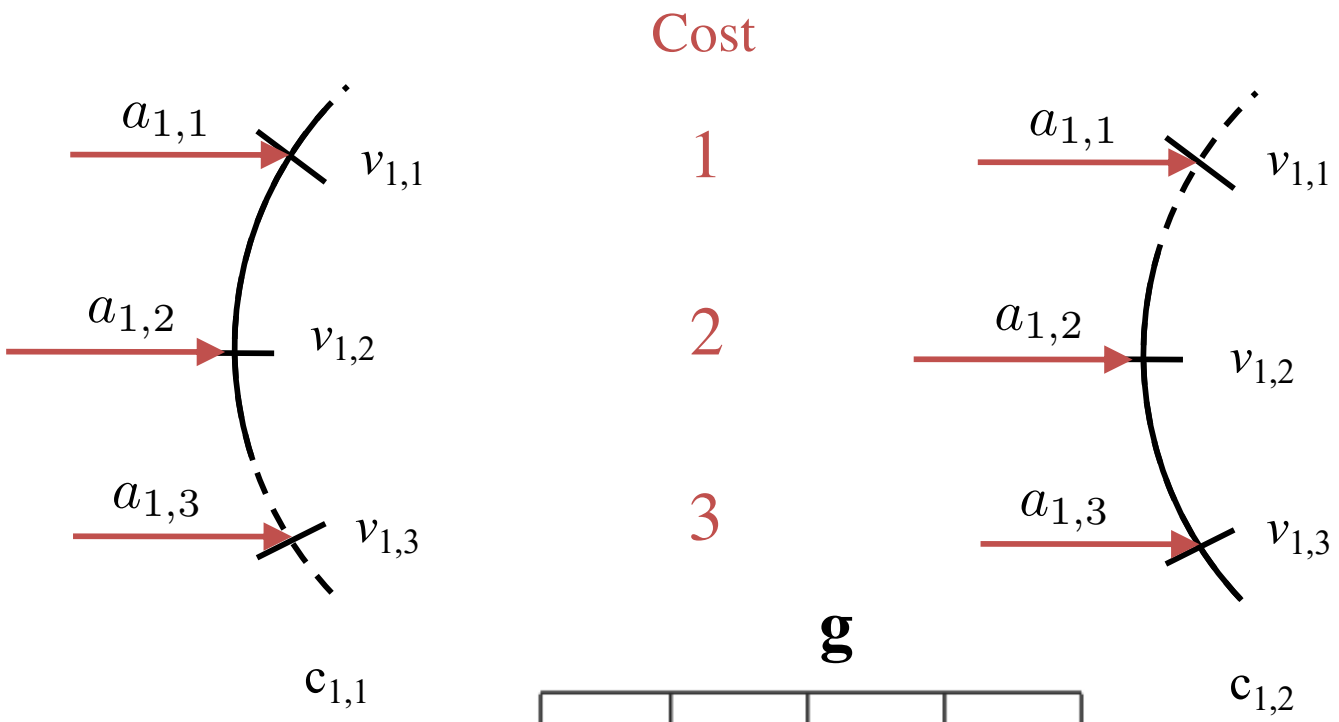
Mathematical Analysis of the Feedback System

$$\hat{r}^S_{l,t+1}(c_{l,h}) = \hat{r}^S_{l,t}(c_{l,h}) + \mu^S_t \mathbb{1}_{\{\mathbb{c}_{l,t}=c_{l,h}\}}(r_{l,t} - \hat{r}^S_{l,t}(c_{l,h}))$$

$$f_{l,h,t+1} = (1-\lambda^S_{l,t}) f_{l,h,t} + \lambda^S_{l,t} \left(\frac{f_{l,h,t} e^{-\frac{\hat{r}_{l,t}(c_{l,h})}{\epsilon^S_{l,t}}}}{\sum_{h'=1}^{m_l} f_{l,h',t} e^{-\frac{\hat{r}_{l,t}(c_{l,h'})}{\epsilon^S_{l,t}}}} \right)$$

- Use stochastic approximation to show the convergence to an ordinary differential equation (ODE).
- Use ODE to show the convergence of the coupled dynamics to the equilibrium

Game-Theoretic Model Revisited



Ξ_1 : **f**

g

	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$
$c_{1,1}$	1	2	0
$c_{1,2}$	0	2	3

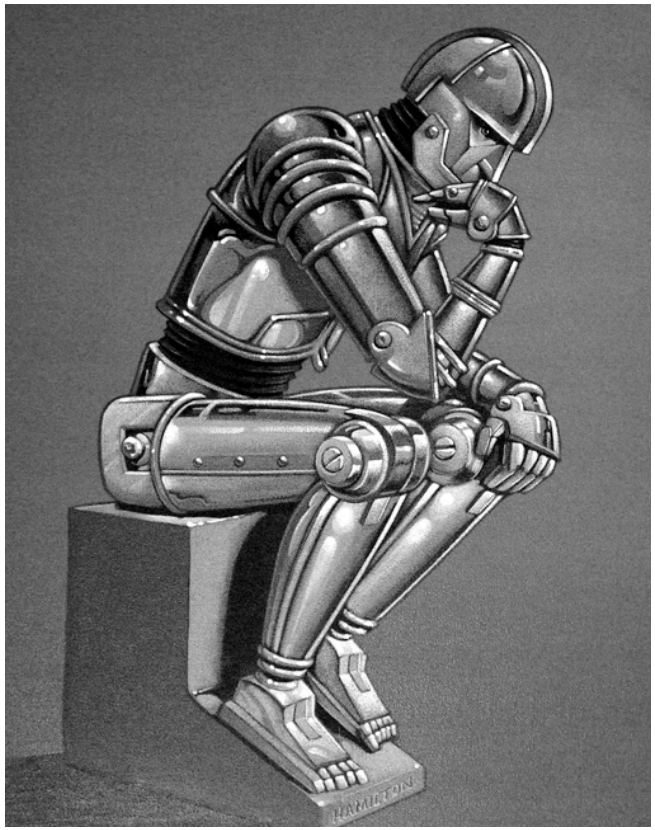
- **f** and **g** are mixed strategies. Mixed strategies are *randomizing* strategies.
- Nash equilibrium to the zero-sum game exists, and yields the worst-case defense strategy.

Boltzmann Learning

$$\begin{cases} \mathbf{f}_{t+1} & = & (1-\lambda_{1,t})\mathbf{f}_t + \lambda_{1,t}\tilde{\beta}_{1,\epsilon}(\hat{\mathbf{u}}_{1,t}) \\ \hat{\mathbf{u}}_{1,t+1} & = & \hat{\mathbf{u}}_{1,t} + \frac{\mu_{1,t}}{f_t(a_1)}\mathbb{1}_{\{a_{1,t}=a_1\}}(U_{1,t} - \hat{\mathbf{u}}_{1,t}) \\ \mathbf{g}_{t+1} & = & (1-\lambda_{2,t})\mathbf{g}_t + \lambda_{2,t}\tilde{\beta}_{2,\epsilon}(\hat{\mathbf{u}}_{2,t}) \\ \hat{\mathbf{u}}_{2,t+1} & = & \hat{\mathbf{u}}_{2,t} + \frac{\mu_{2,t}}{g_t(a_2)}\mathbb{1}_{\{a_{2,t}=a_2\}}(U_{2,t} - \hat{\mathbf{u}}_{2,t}) \end{cases}$$

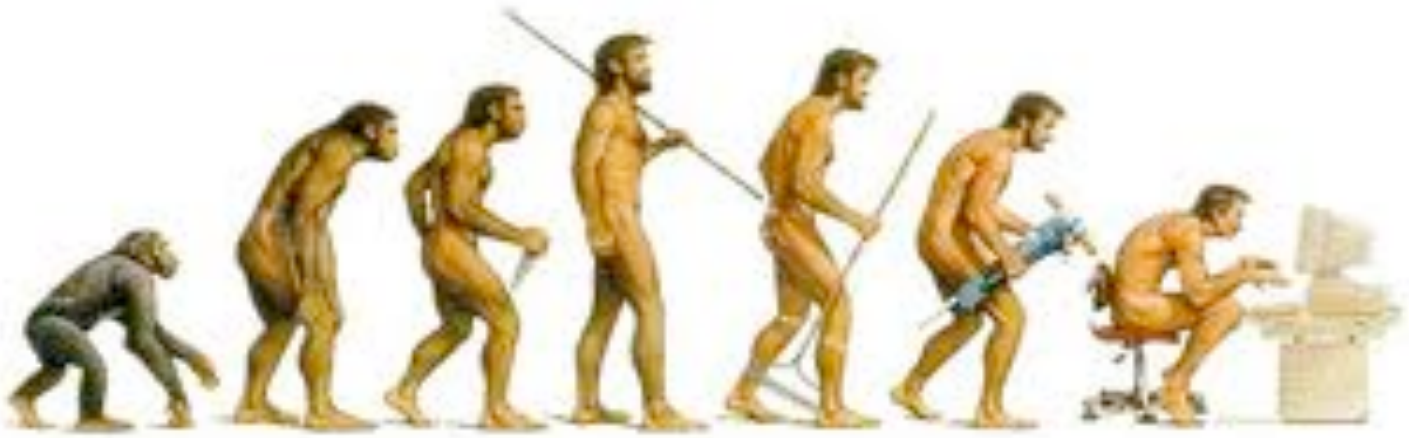
Soft-max function

$$\tilde{\beta}_{i,\epsilon}(\hat{\mathbf{u}}_{i,t})(a_i) = \frac{e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i)}}{\sum_{a'_i} e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a'_i)}}$$

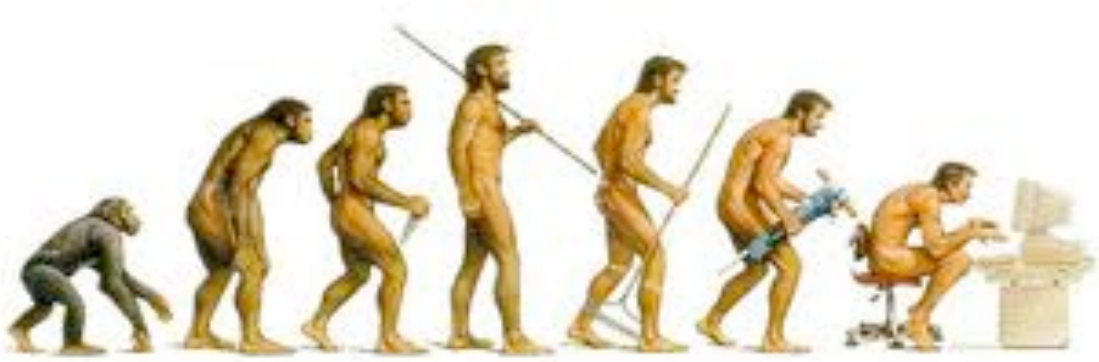


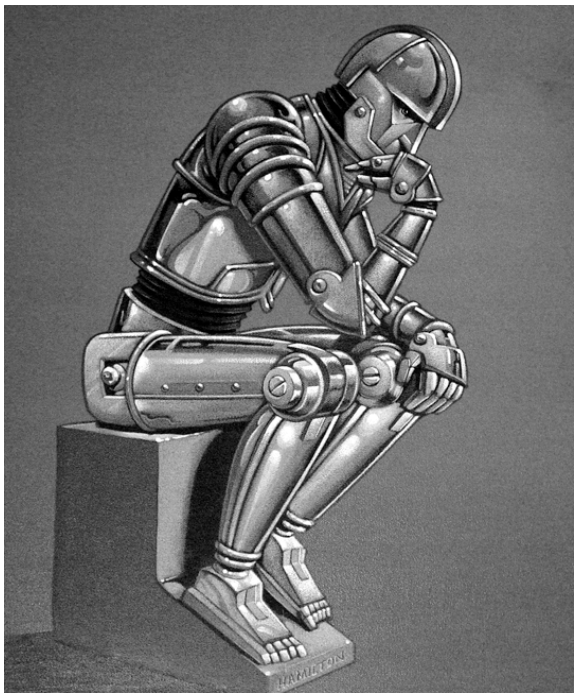
Replicator Dynamics

$$
\begin{cases}
\mathbf{f}_{t+1} & = & \mathbf{f}_t + \lambda_{1,t} U_{1,t} \cdot \left(\mathbb{1}_{\{a_{1,t}=a_1\}} - \mathbf{f}_t \right) \\
\hat{\mathbf{u}}_{1,t+1} & = & \hat{\mathbf{u}}_{1,t} + \mu_{1,t} \mathbb{1}_{\{a_{1,t}=a_1\}} \left(U_{1,t} - \hat{\mathbf{u}}_{1,t} \right) \\
\mathbf{g}_{t+1} & = & \mathbf{g}_t + \lambda_{2,t} U_{2,t} \cdot \left(\mathbb{1}_{\{a_{2,t}=a_2\}} - \mathbf{g}_t \right) \\
\hat{\mathbf{u}}_{2,t+1} & = & \hat{\mathbf{u}}_{2,t} + \mu_{2,t} \mathbb{1}_{\{a_{2,t}=a_2\}} \left(U_{2,t} - \hat{\mathbf{u}}_{2,t} \right)
\end{cases}
$$



Boltzmann Learning vs. Replicator Dynamics





Robust yet inefficient

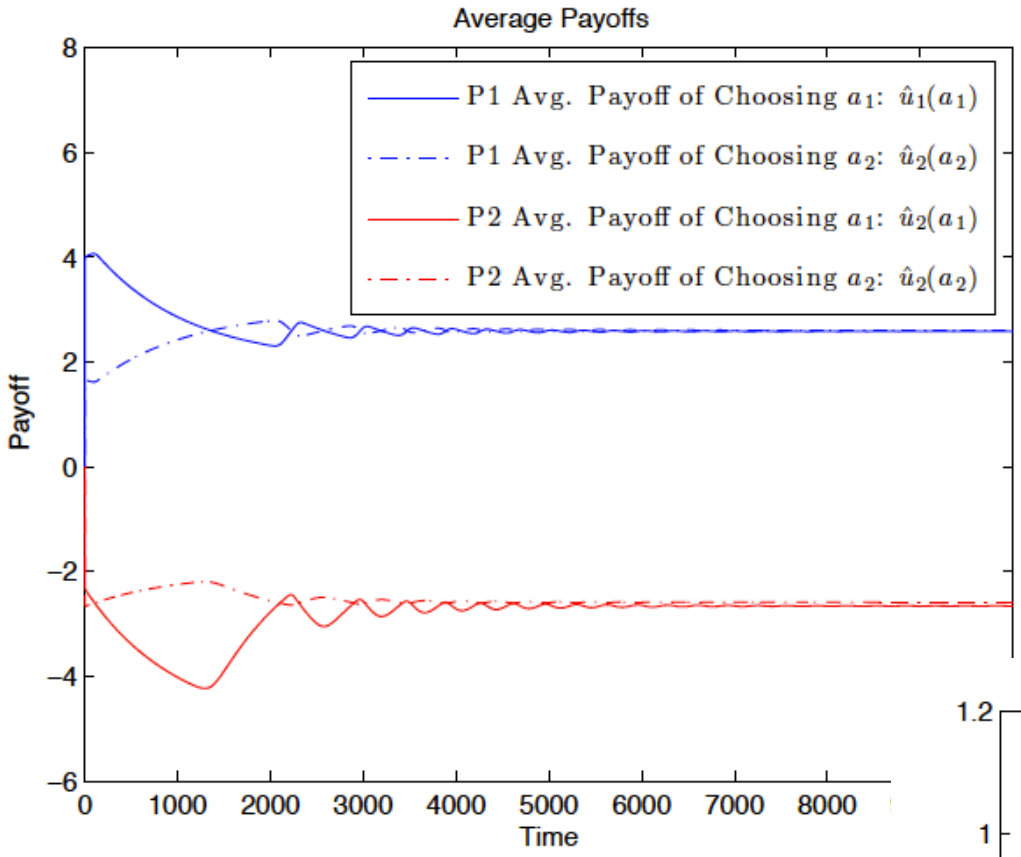
Fragile yet efficient

Heterogeneous and Hybrid Learning

- Heterogeneous learning: Different players adopt different learning schemes.

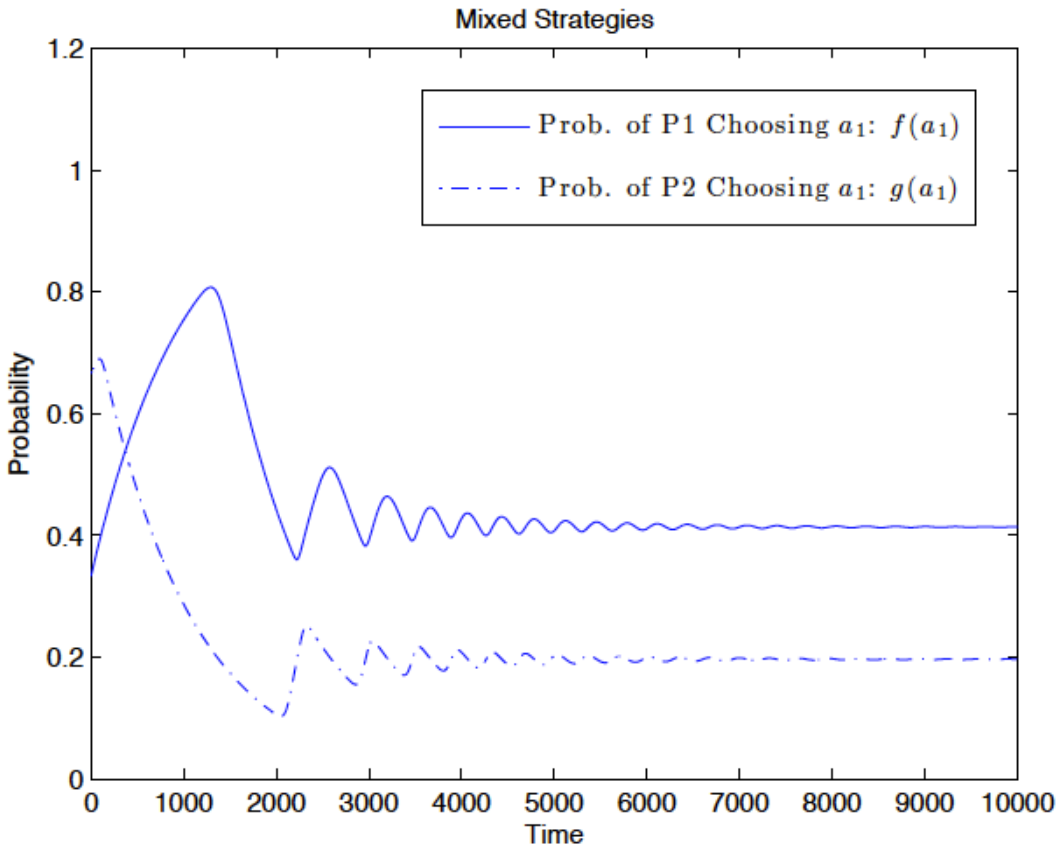
- Hybrid learning: Players adopt different learning schemes at different times.

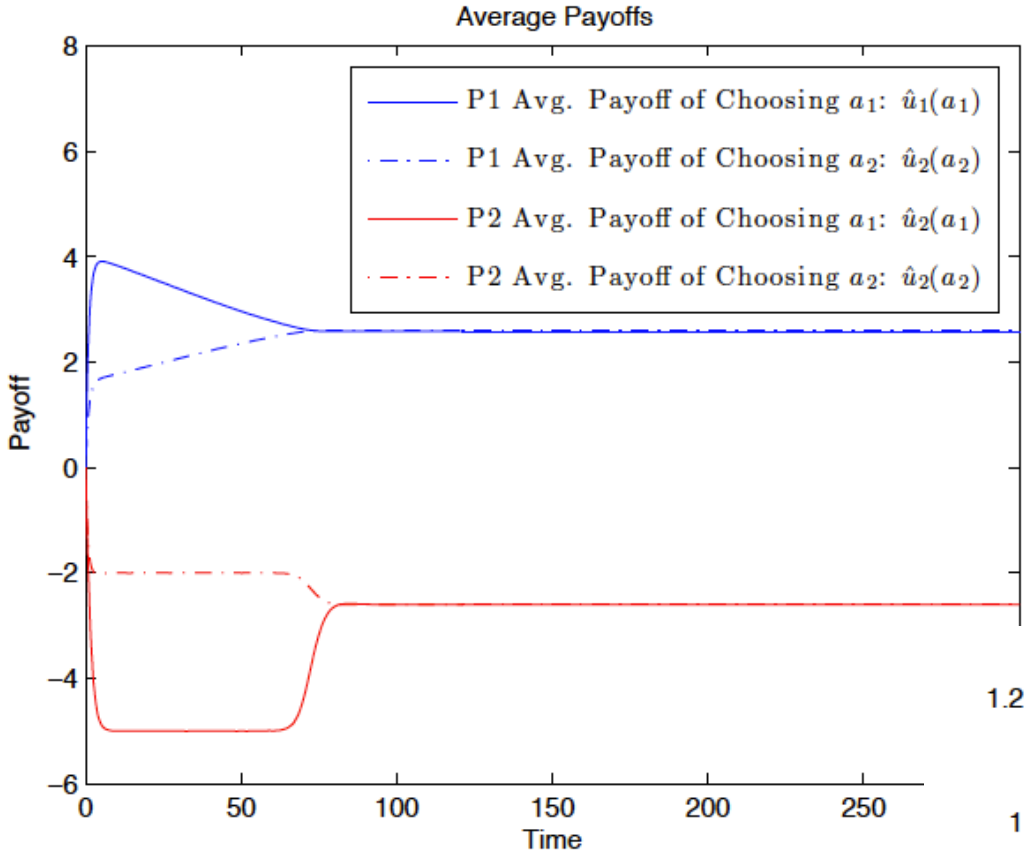
Q. Zhu, H. Tembine and T. Basar, “Hybrid learning in stochastic games and its application in network security,” In F. L. Lewis and D. Liu (Eds.), Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, IEEE Press Computational Intelligence Series, 2012.



Average Payoffs
Boltzmann Learning

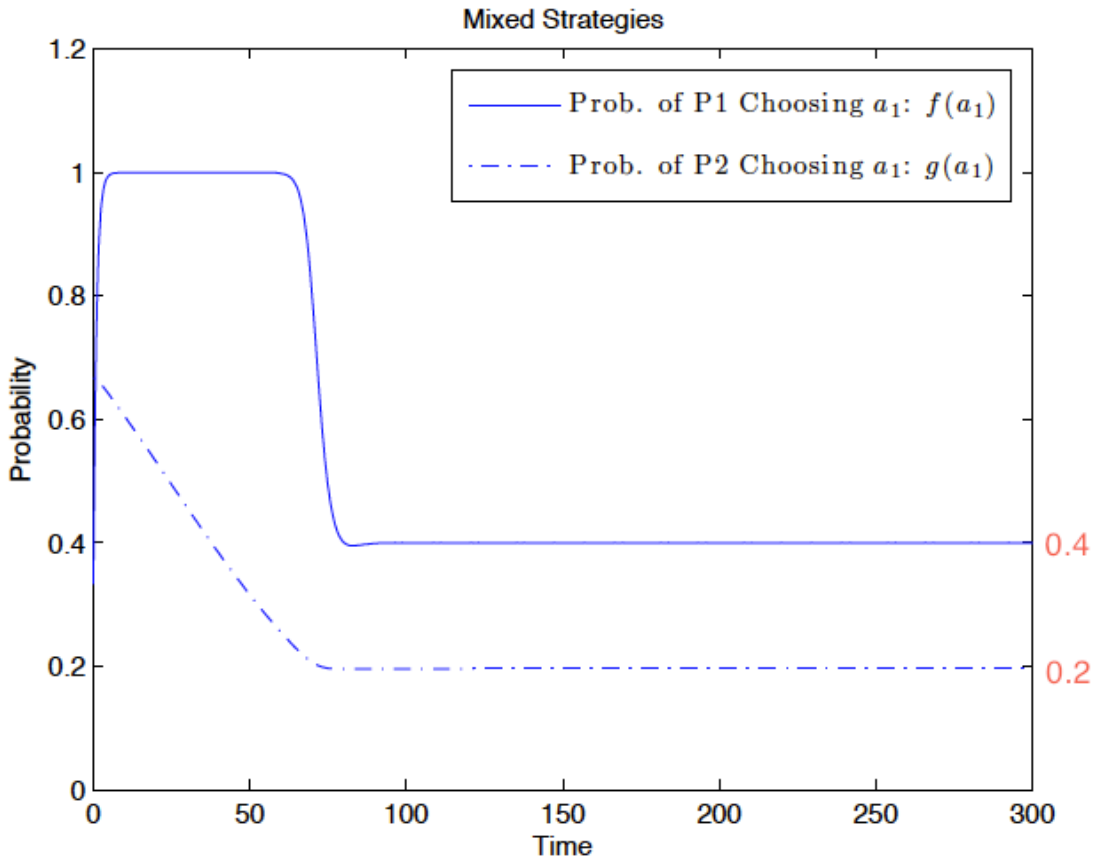
Mixed Strategies
Boltzmann Learning



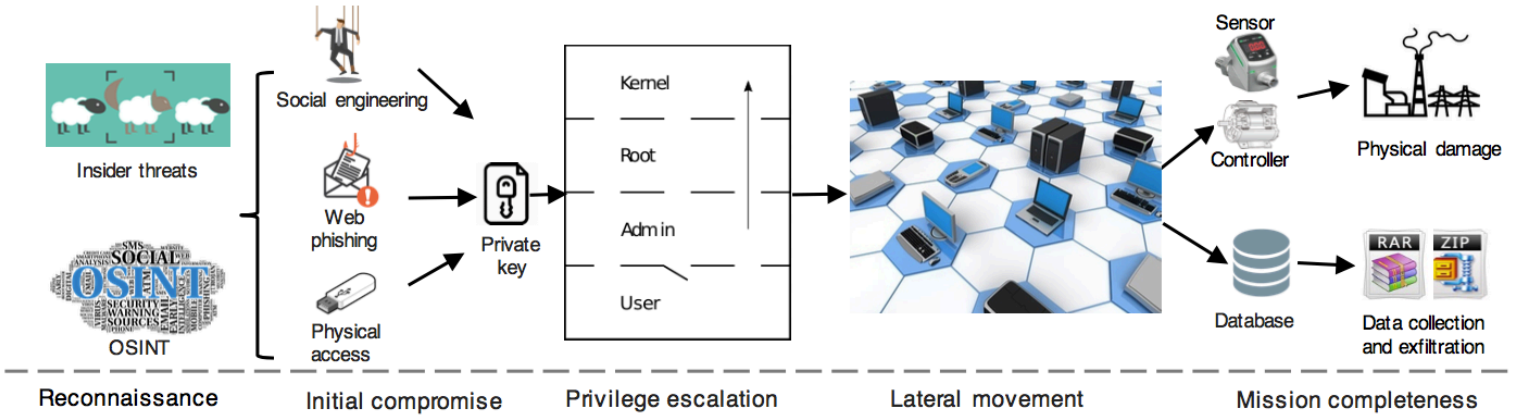


Average Payoffs
Boltzmann with Replicator

Mixed Strategies
Boltzmann with Replicator



Proactive Defense Against Advanced Persistent Threat



APT features

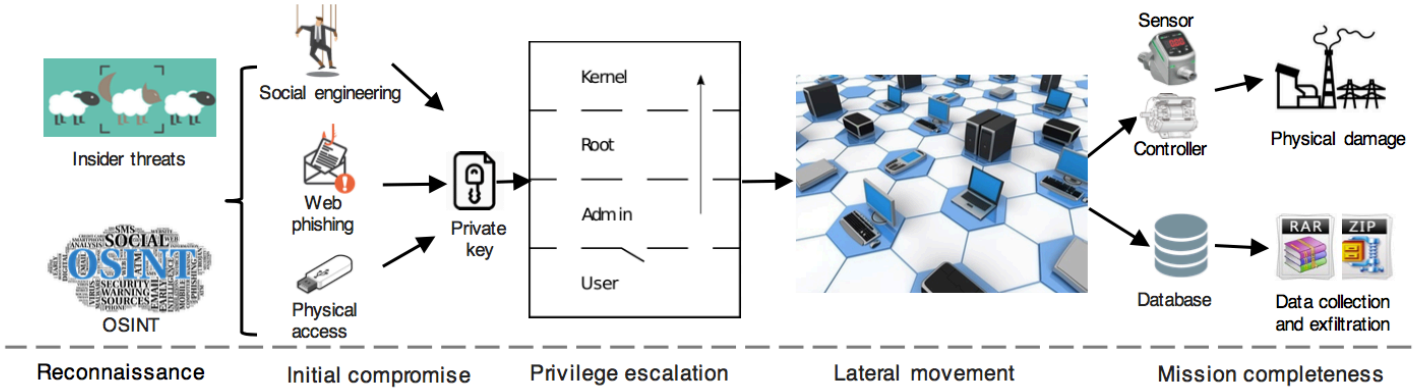
- Targeted attack: reconnaissance
- Persistent: firm and patient
- Advanced: technically and strategically

Cyber Deception

- Stealthy
- Deceptive

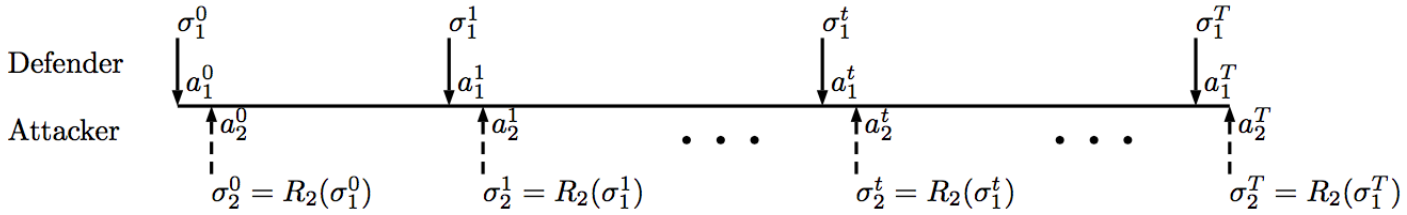
Multi-Stage Game Framework

Proactive defense: active response prior to the attack



Cross-layer Defense

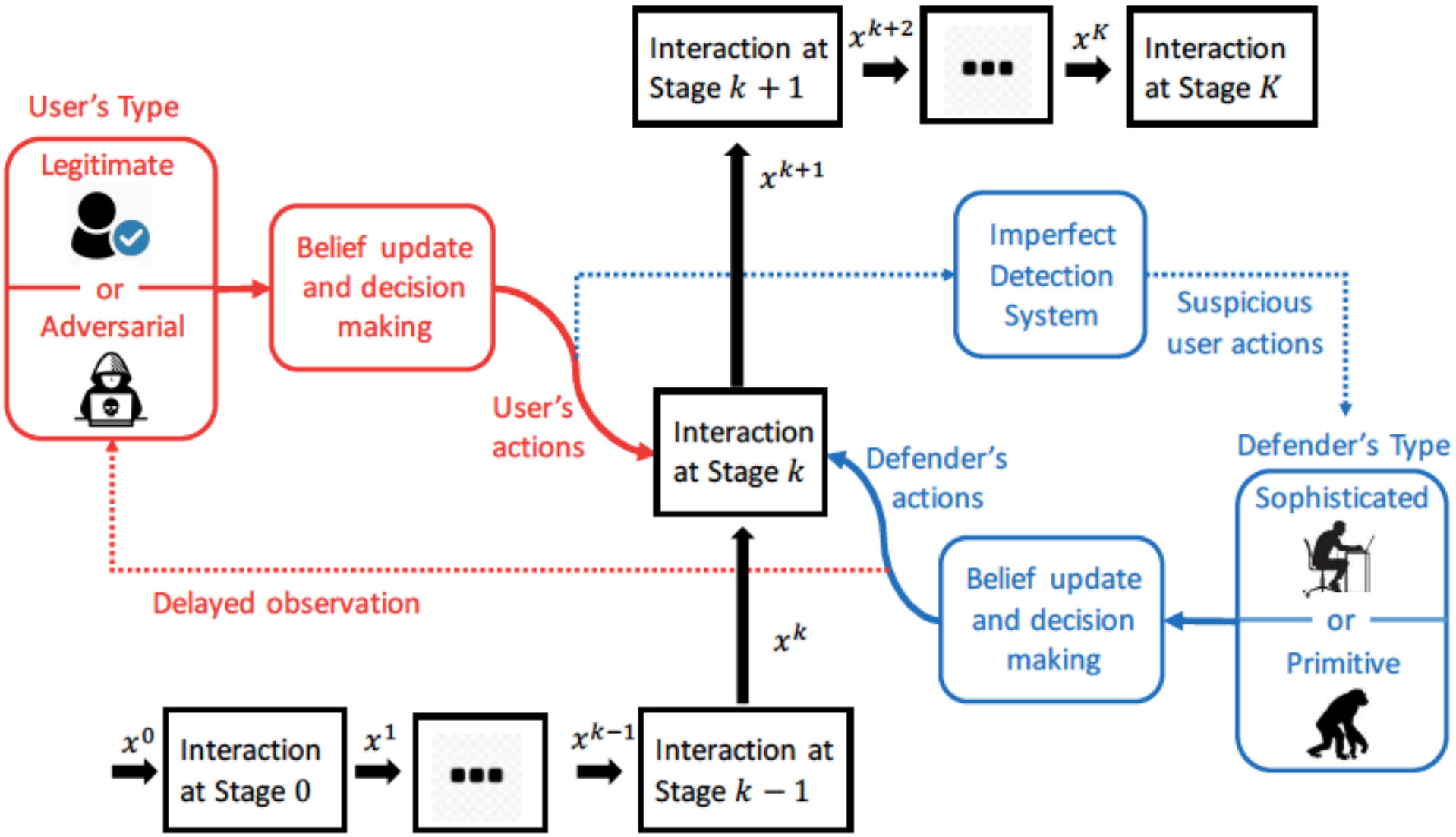
Online learning

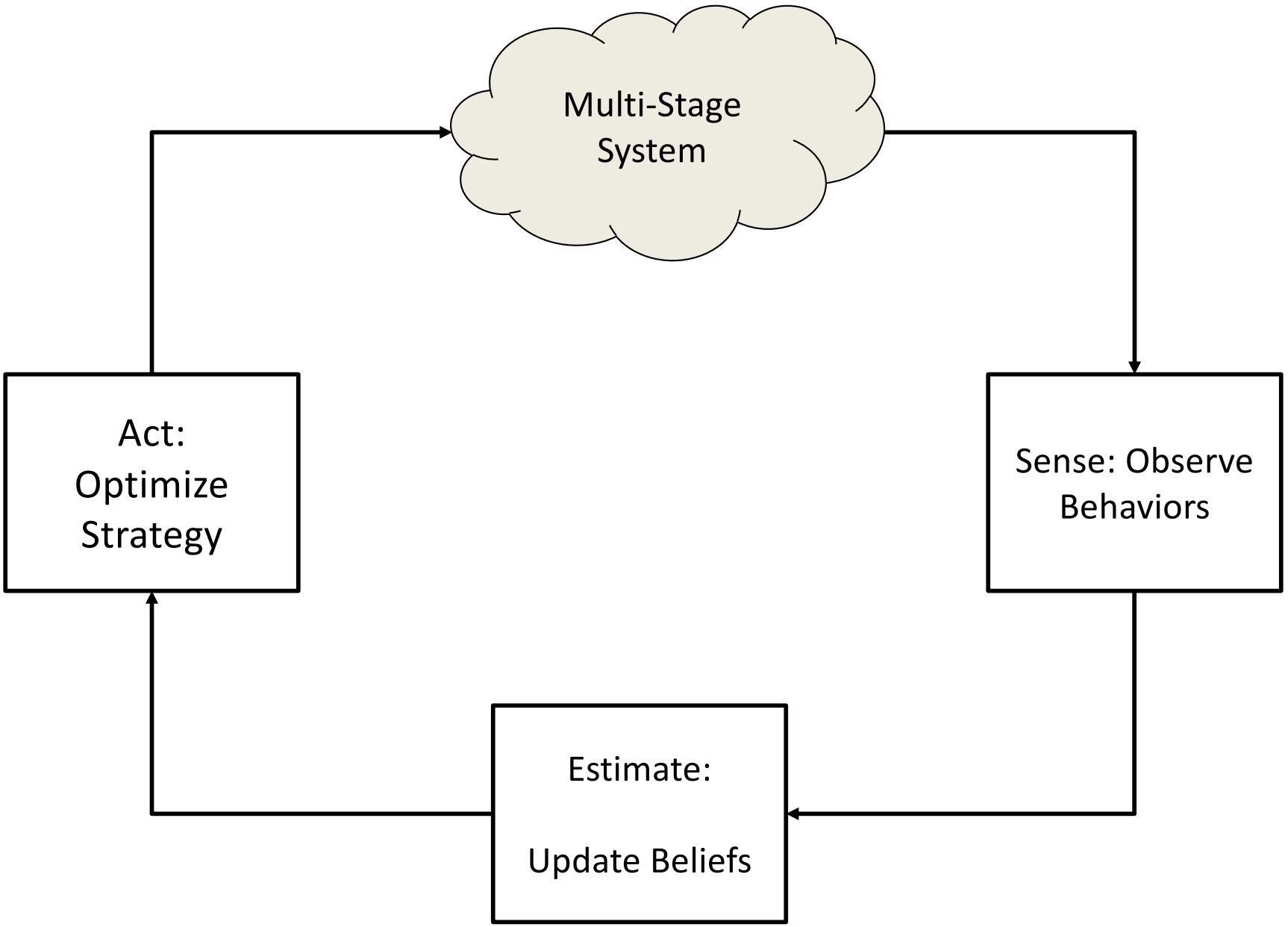


Observable History

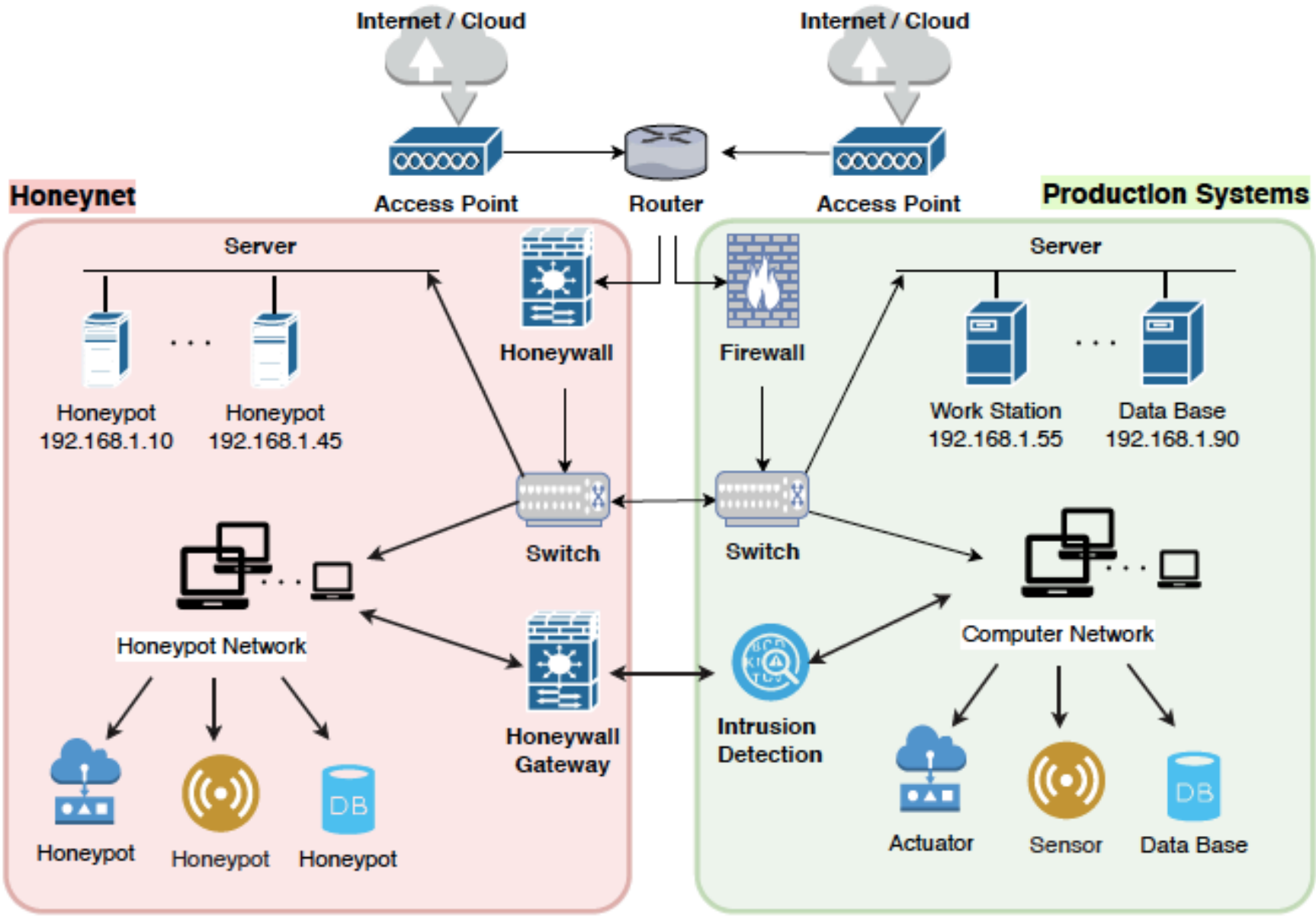
Stage 0	Stage 1	Stage t	Stage T
$\mathbf{h}^0 = \emptyset$	$\mathbf{h}^0 = \{a_1^0, a_2^0\}$	$\mathbf{h}^t = \{\mathbf{h}^{t-1}, a_1^{t-1}, a_2^{t-1}\}$	$\mathbf{h}^T = \{a_1^0, ..., a_1^{T-1}, a_2^0, ..., a_2^{T-1}\}$

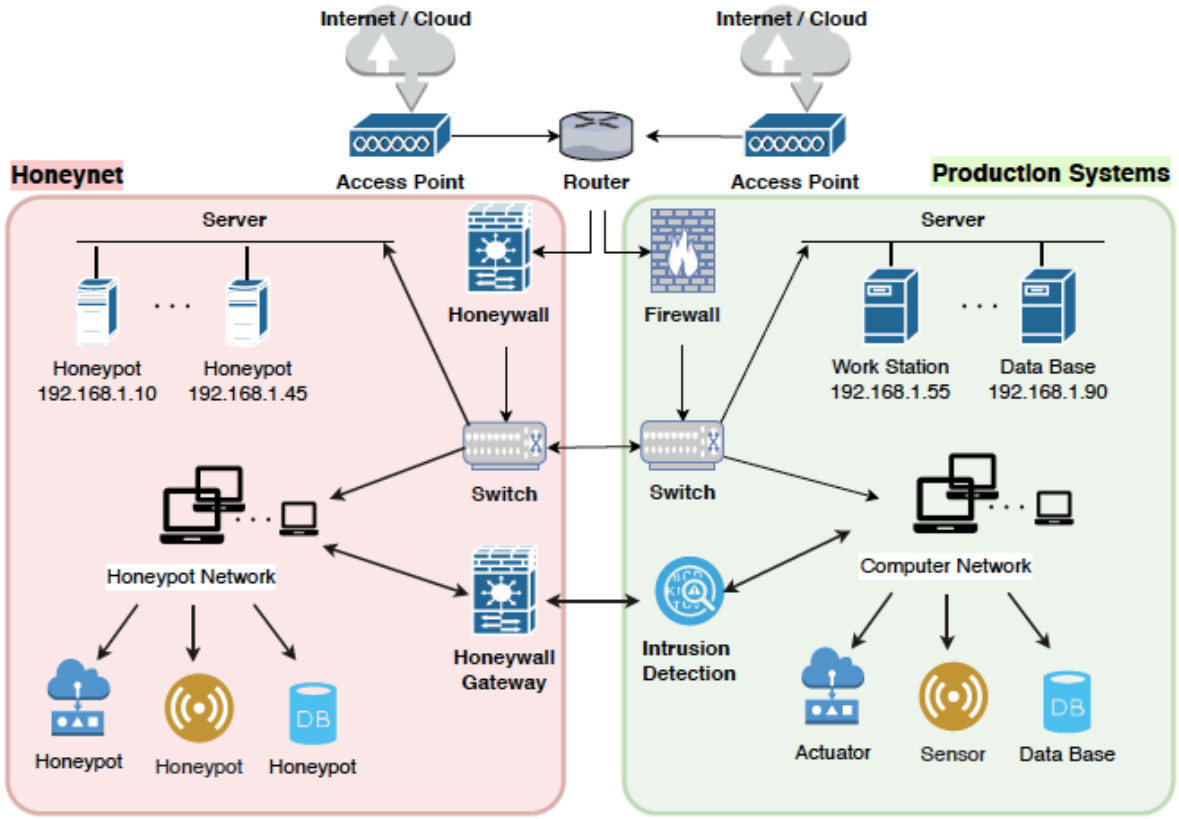
Multi-Stage Co-Learning




Multi-Stage System
Act: Optimize Strategy
Sense: Observe Behaviors
Estimate:
Update Beliefs

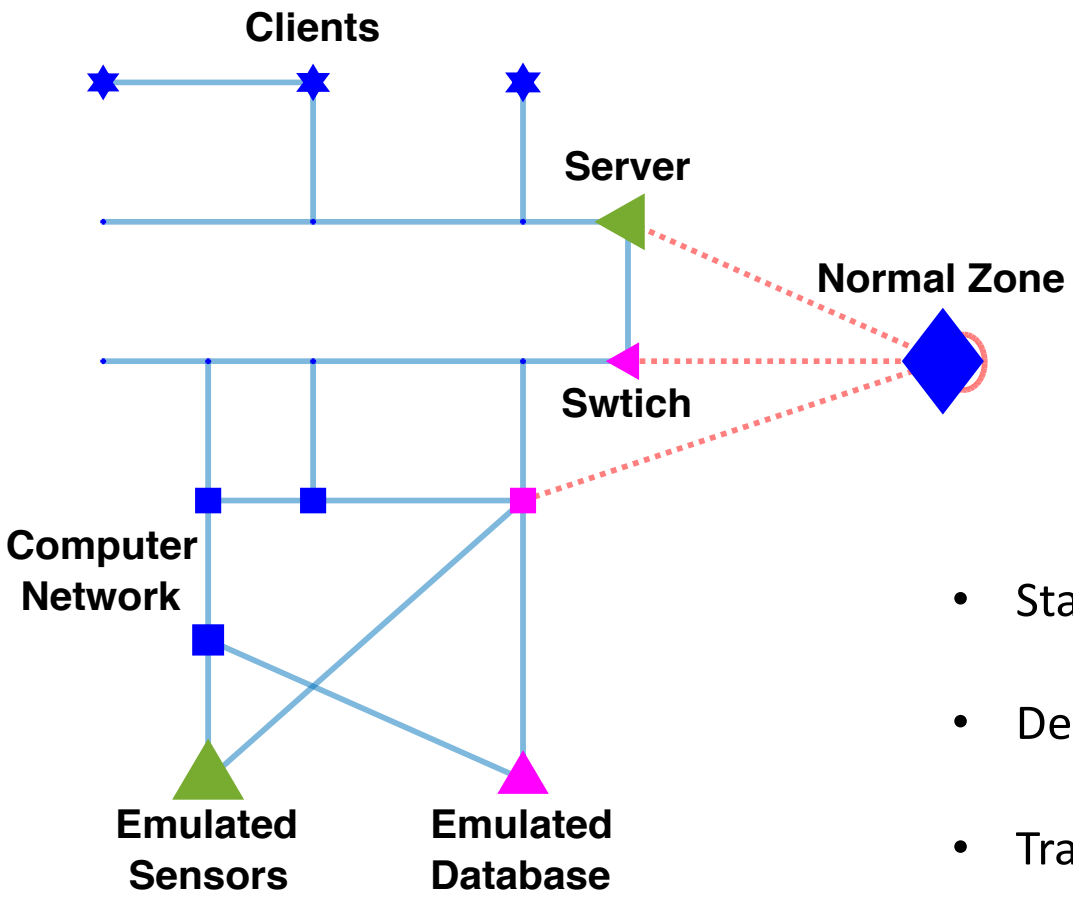
Attacker Engagement in Honeypot





- Use a honeynet to emulate a production system.
- Interact with rather than directly eject attackers.
- Quickly attract attackers to target honeypots and engage them for a desired time.
- Grant attackers proper degree of freedom to avoid the escape risk and the identification risk.

Abstraction: Semi-Markov Decision Process



Red: eject the attacker

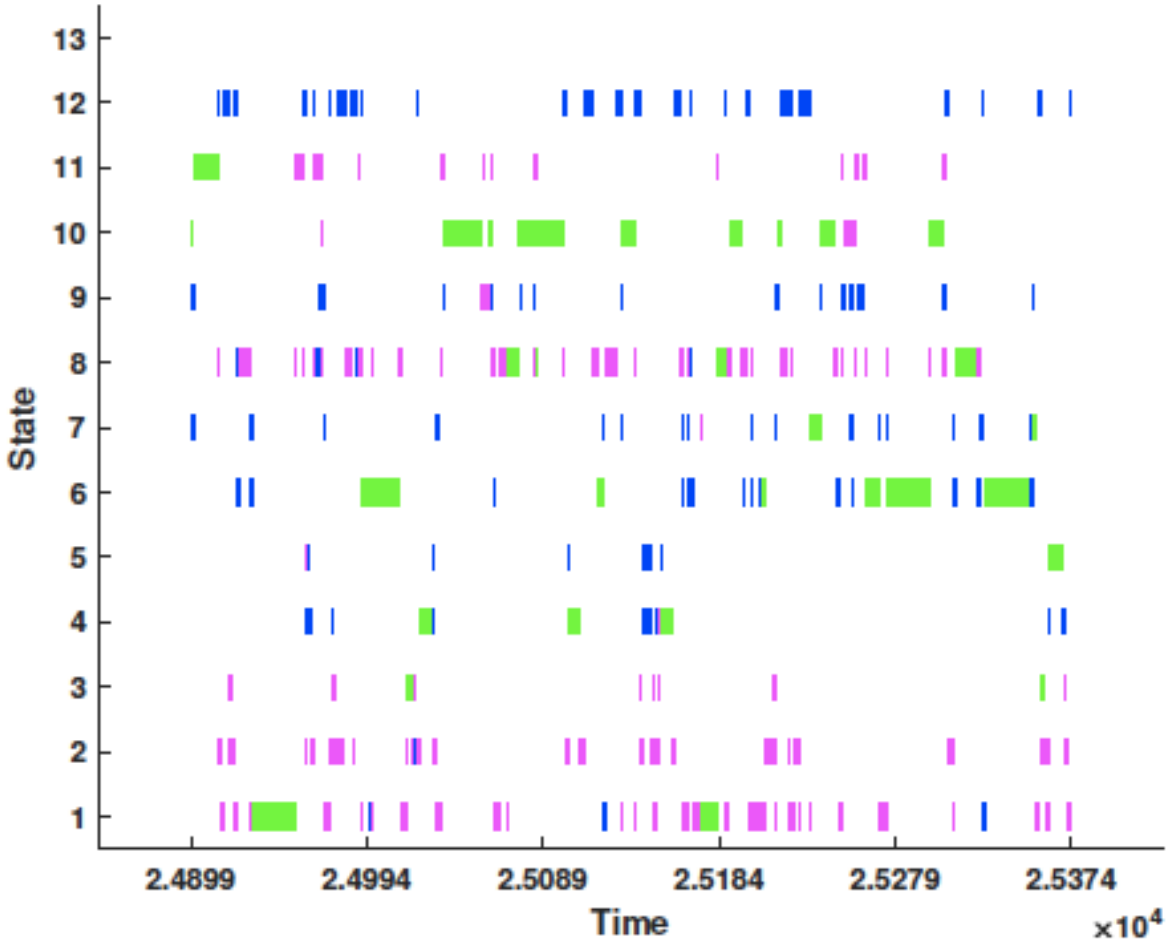
Blue: pure recording of attacker's activity

Purple: low-interaction with the attacker

Green: high-interaction with the attacker

- State at time t : $s^t \in S$
- Defender's action at state s: $a \in A(s)$
- Transition probability: $p(s^{t+1}|s^t, a)$
- Sojourn time distribution $q(t|s^t, s^{t+1}, a)$

Attacker's Footprint

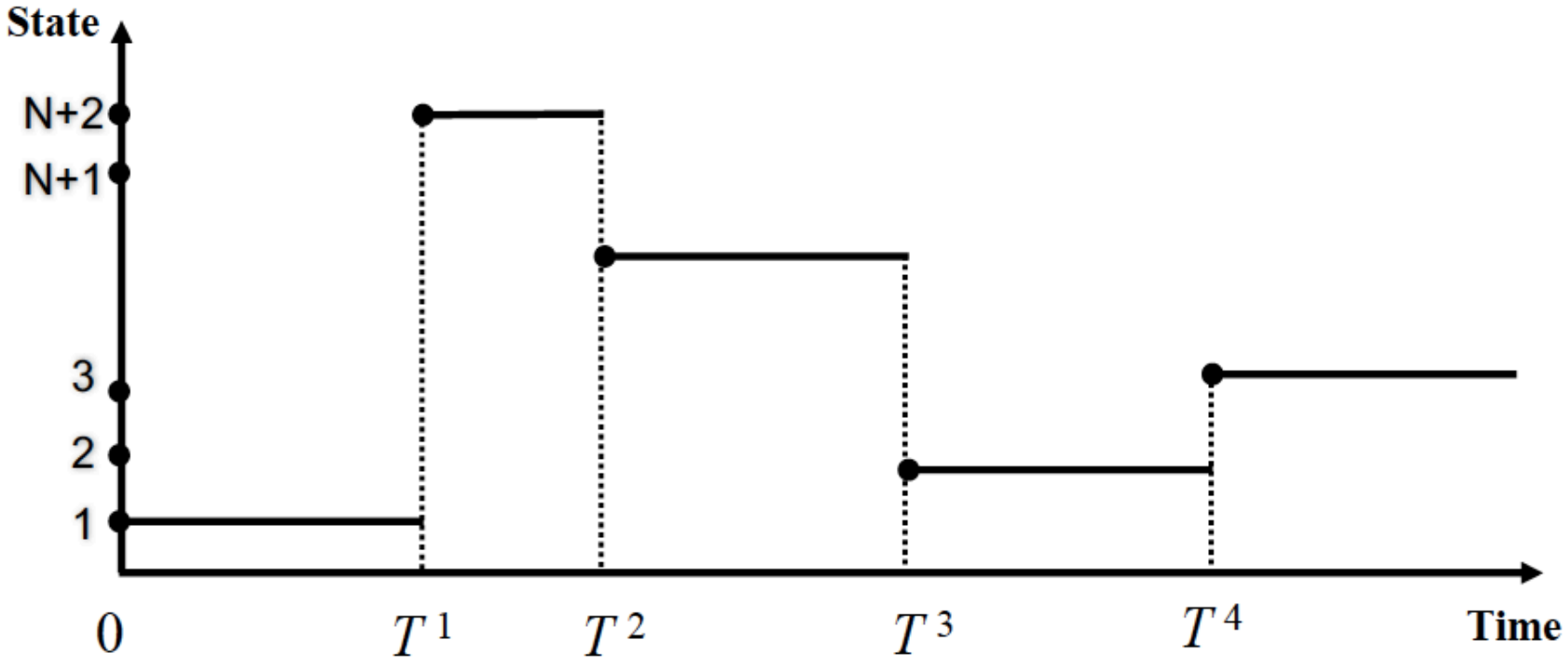


- Treat transition kernel and sojourn distribution as threat intelligence.
- Characterize the escape risk and the identification risk.

Semi-Markov Decision Process and Learning

- T^k is the time of the k^{th} transition, which is a random variable.

- Defender receives a reward r at time $\tau, T^k \leq \tau \leq T^{k+1}$ if the next state is *s'* and the duration time at current *s, a* is $T^{k+1} - T^k$.



Q-Learning for SMDP

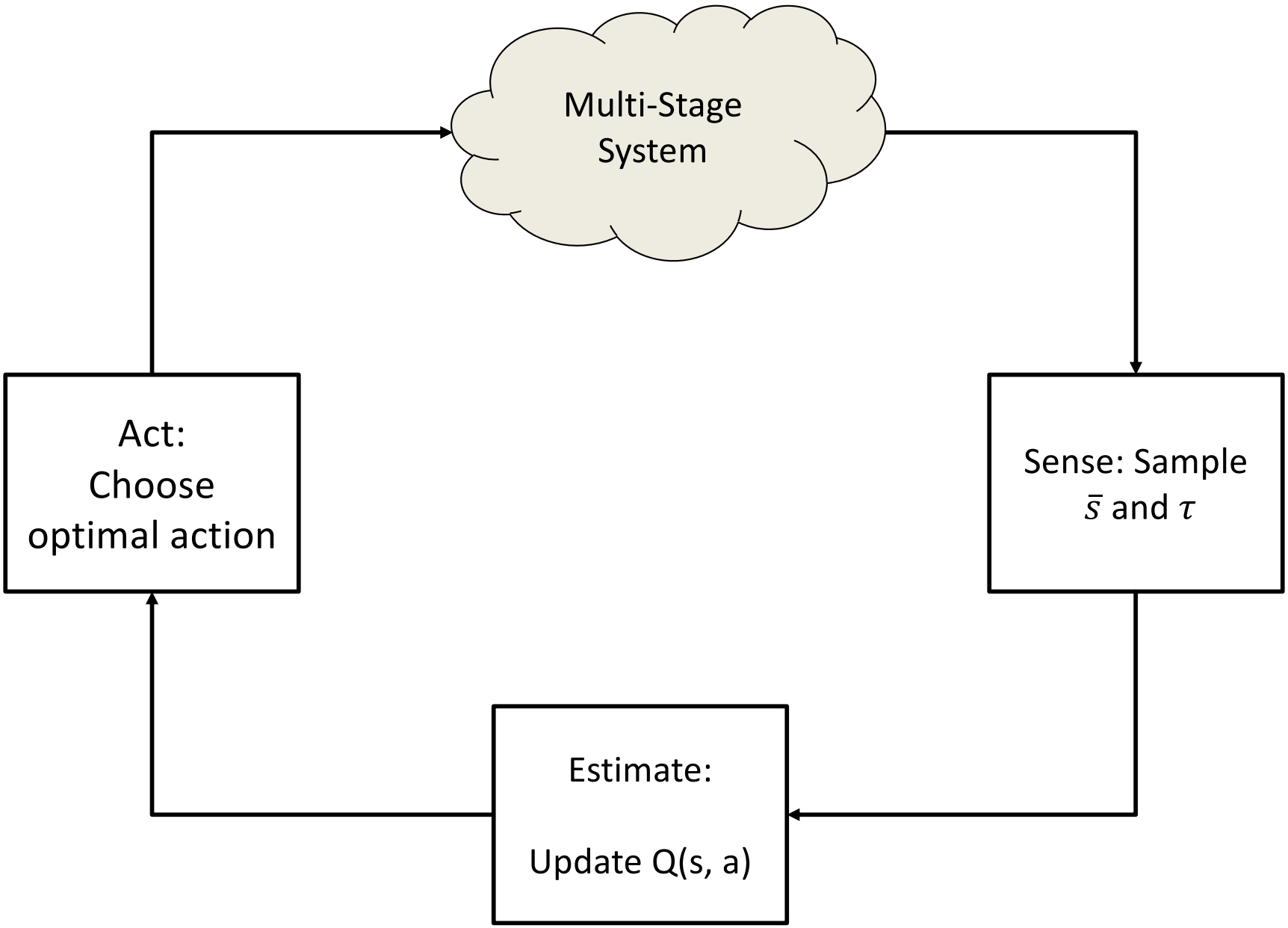
- Defender's stationary policy:π: $S \rightarrow A(s)$

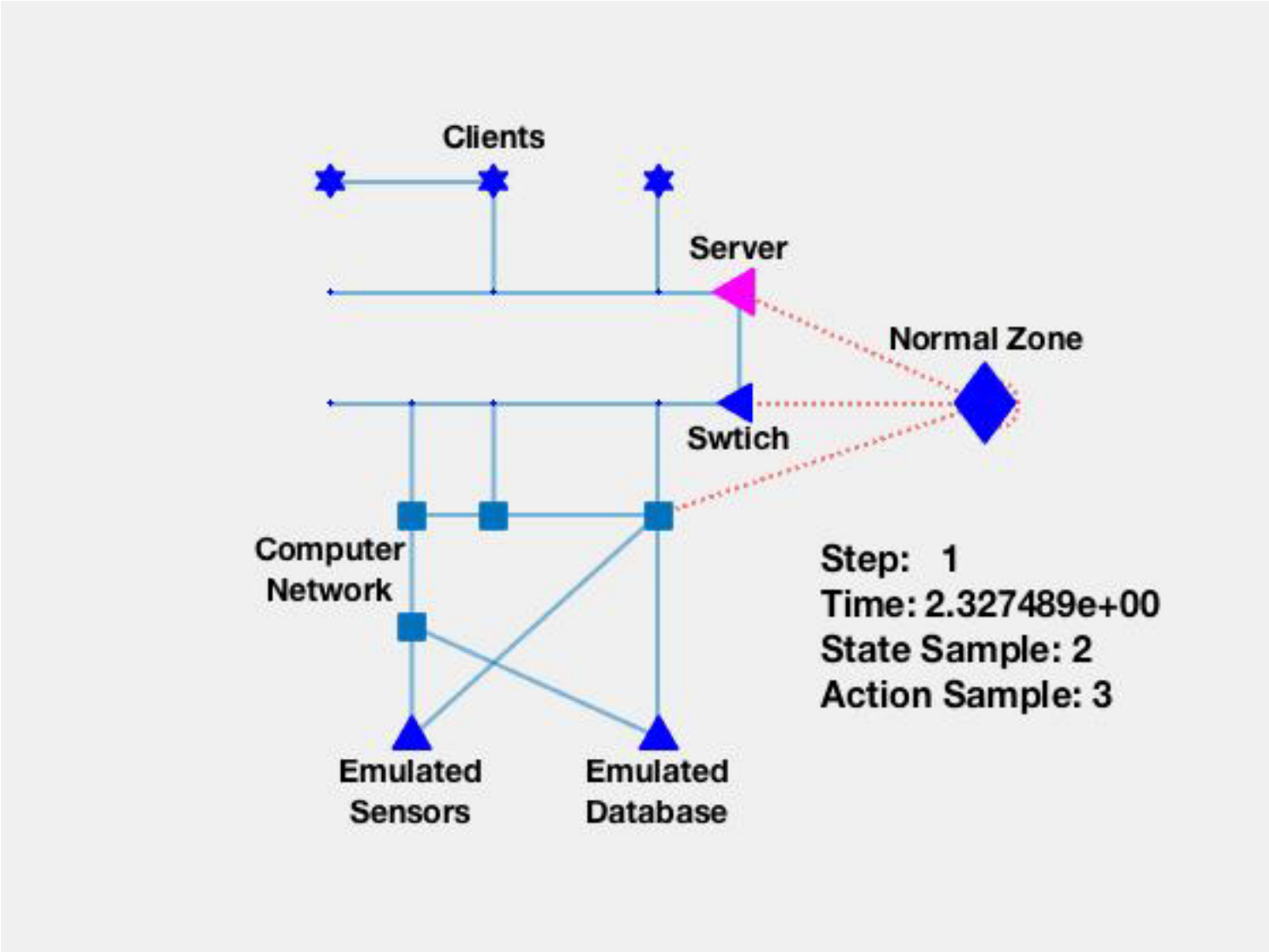
- DP representation

$$v(s) = \max_a \sum_{s'} p(s'|s,a)(r^{\gamma}(s,a,s') + \tilde{q}(s,a,s')v(s'))$$

- Q-Learning: sampled state $\bar{s}$ and the duration time τ

$$Q^{k+1}(s,a) = (1-\alpha^k)Q^k(s,a) + \alpha^k(r_1(s,a,\bar{s}) + \frac{(1-e^{-\gamma\tau})r_2(s,a,\bar{s})}{\gamma} + e^{-\gamma\tau}\max_{a'} Q^k(\bar{s},a'))$$


Multi-Stage System
Act: Choose optimal action
Sense: Sample $\bar{s}$ and τ
Estimate:
Update Q(s, a)

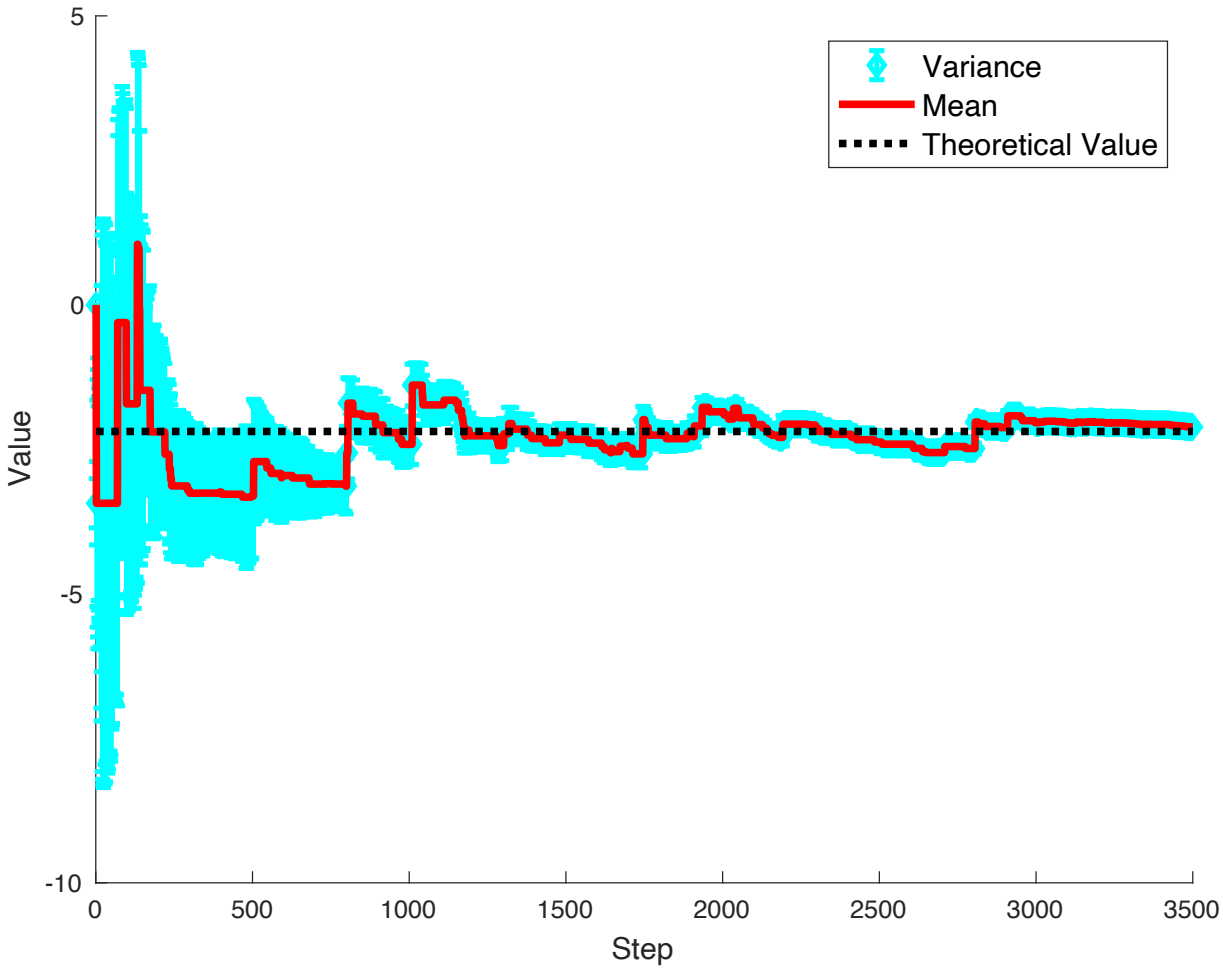


Red: eject the attacker

Purple: low-interaction with the attacker

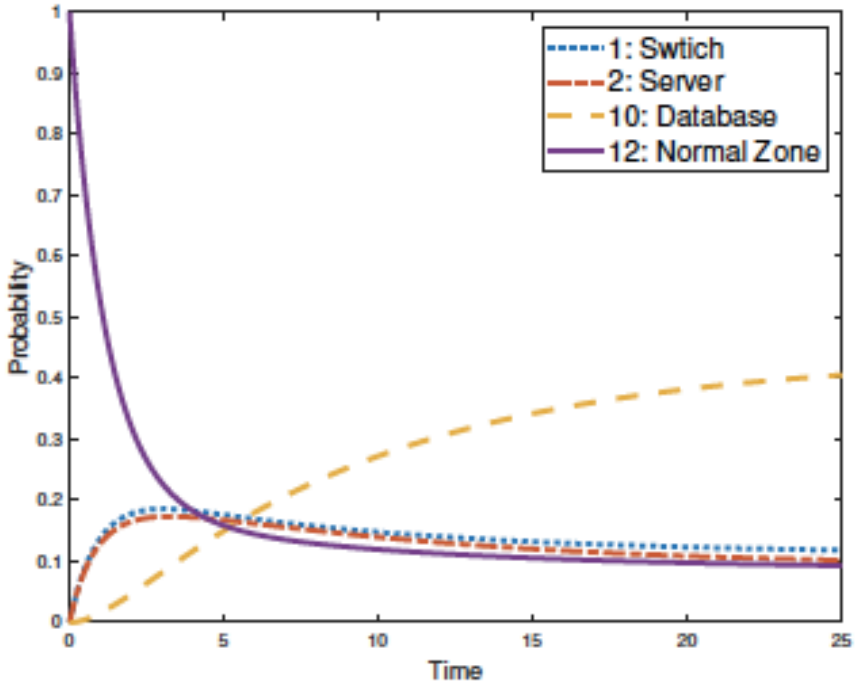
Blue: pure recording of attacker's activity

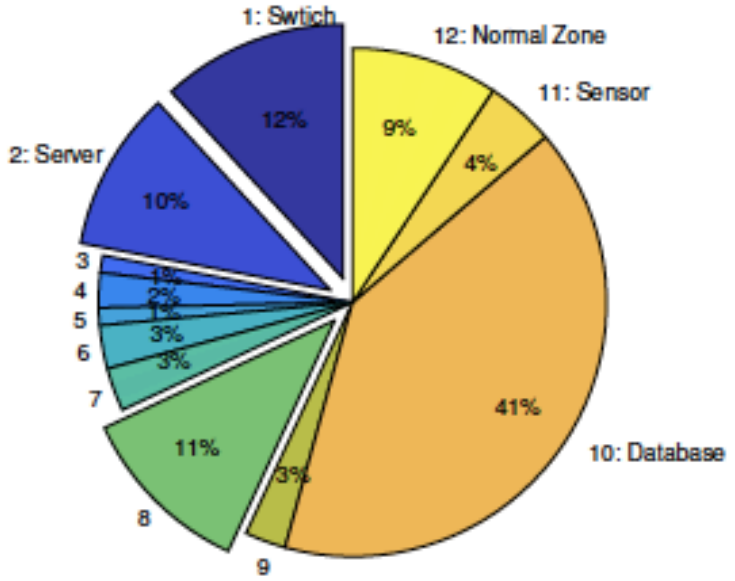
Green: high-interaction with the attacker



L. Huang and Q. Zhu, “Strategic Learning for Active, Adaptive, and Autonomous Cyber Defense, "Adaptive Autonomous Secure Cyber Systems, Jajodia, S., Cybenko, G., Subrahmanian, V.S., Swarup, V., Wang, C., Wellman, M. (Eds.), 2020.

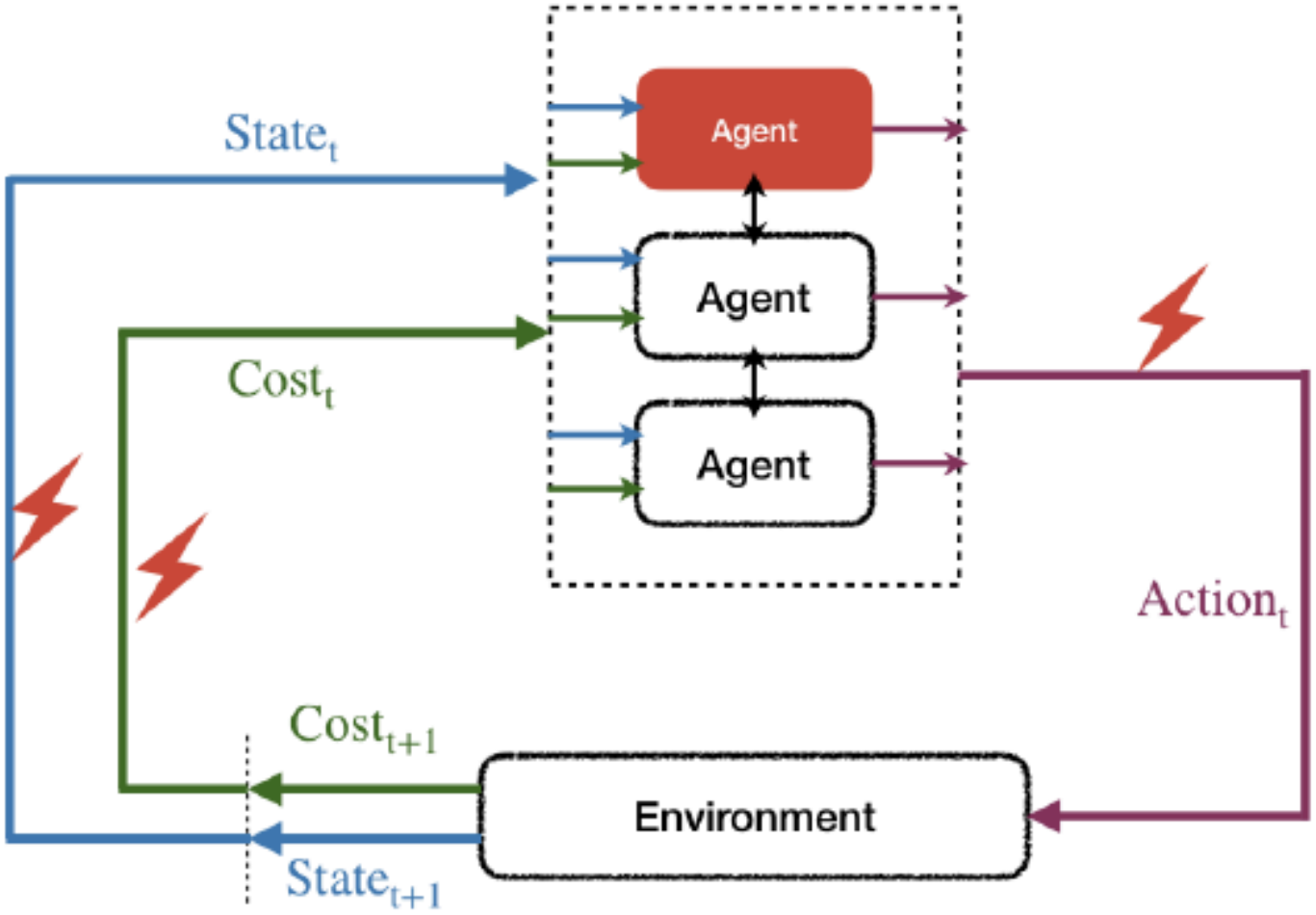
Security Metrics to Evaluate Engagement





L. Huang, Q. Zhu, “Adaptive Honeypot Engagement through Reinforcement Learning of Semi-Markov Decision Processes," Conference on Decision and Game Theory for Security (GameSec), Oct. 30 - Nov. 1, 2019, Stockholm, Sweden.

RL in Adversarial Environment



Y. Huang, Q. Zhu, “Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals," Conference on Decision and Game Theory for Security (GameSec), Oct. 30 - Nov. 1, 2019, Stockholm, Sweden.

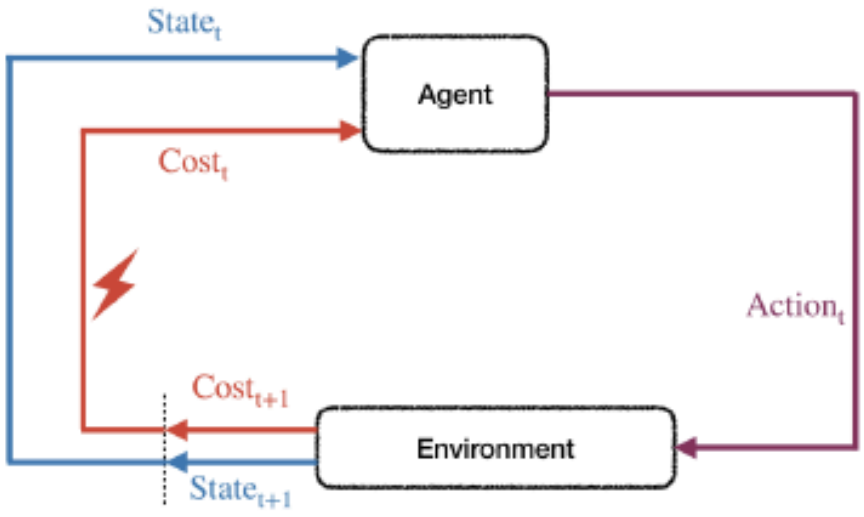
- Drones with RL techniques can be indirectly trained and weaponized by terrorists.

- Self-driving car can be misled to collisions by receiving false feedback.

RL with Manipulated Cost Signals

Consider a Markov Decision Process (MDP) with manipulated cost signals, denoted by $\langle \mathcal{S}, \mathcal{A}, c, \tilde{c}, \mathcal{P}, \beta \rangle$.

- State space $\mathcal{S} := \{1, 2, \cdots, S\}$.
- Action space $\mathcal{A} := \{a_1, \cdots, a_A\}$.
- Cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.
- Manipulated cost $\tilde{c}$.
- Transition probability kernel $\mathcal{P}$
- Discounted factor β



• The agent aims for **the optimal policy** $w : \mathcal{S} \to \mathcal{A}$ minimizing

$$J(i, \mathbf{Z}) := \mathbf{E}[\sum_{t=0}^{\infty} \beta^t c(\Phi(t), Z(t)) | \Phi(0) = i],$$

where $\Phi(\tau)$ *is the state at time* τ, $Z(\tau)$ *is the action chosen at time* τ *by* $Z(\tau) = w(\Phi(\tau))$.

• To find the optimal policy w^*, the agent implements **Q-learning algorithm**, i.e., for $i = \Phi(t), a = Z(t)$,

$$Q_{t+1}(i, a) = Q_t(i, a) + s(t) \times \Big[\beta \min_b Q_t(\Psi_{t+1}(i, a), b) + c(i, a) - Q_t(i, a)\Big],$$

where $\Psi_{n+1}(i, a)$ is a realized $\mathcal{S}$-valued random variable with law $p(i, \cdot, a)$.

• Under proper conditions [V.Borkar; SIAM JCO; 2000], $Q_t \to Q^*$ a.s., as $t \to \infty$. And Q^* satisfies

$$Q^*(i,a) = c(i,a) + \beta \sum_j p(i,j,a) \min_b Q^*(j,b), \quad i \in \mathcal{S}, a \in \mathcal{A}. \quad (1)$$

• We use $Q^* = F(Q^*)$ to capture relation (1), where $F : \mathbb{R}^{S \times A} \to \mathbb{R}^{S \times A}$.

• The agent can find an optimal policy w^* by

$$w^*(i) = \arg\min_{a \in \mathcal{A}} Q^*(i,a), i \in \mathcal{S}.$$

The Q-learning algorithm with manipulated cost signals:

$$Q_{t+1}(i,a) = Q_t(i,a) + s(t) \times \left[\beta \min_b Q_t(\Psi_{t+1}(i,a),b) + \tilde{c}_t - Q_t(i,a)\right].$$

Question: How should the attacker manipulates cost signals?

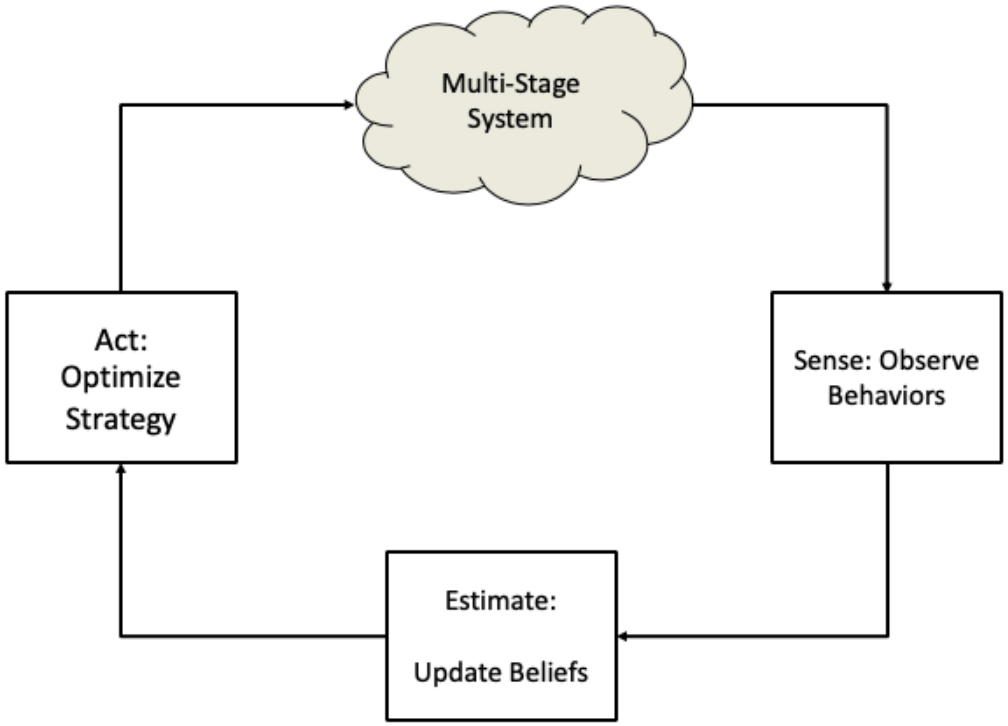
Theorem (Manipulation Rule)

Let $\tilde{Q}^* \in \mathbb{R}^{S\times A}$ be the Q-values learned under the falsified cost $\tilde{c} \in \mathbb{R}^{S\times A}$. Then $\tilde{Q}^* \in \mathcal{V}_{w^\dagger}$ if and only if the falsified cost signals $\tilde{c}$ designed by the adversary satisfy the following conditions

$$\tilde{c}(i,a) > (\mathbf{1}_i - \beta P_{ia})^T (I - \beta P_{w^\dagger})^{-1} \tilde{c}_{w^\dagger}. \tag{2}$$

for all $i \in \mathcal{S}$, $a \in \mathcal{A}\backslash\{w^\dagger(i)\}$.

Conclusion: Feedbacks and learning are fundamental for adaptive defense.



- *Moving target defense:* Baseline learning algorithms
- *Proactive defense against APT:* learning of dynamic games with incomplete information
- *Attacker engagement problem:* reinforcement learning
- *Security of reinforcement learning*: deceptive RL

References

- Q. Zhu, H. Tembine and T. Basar, “Hybrid learning in stochastic games and its application in network security,” In F. L. Lewis and D. Liu (Eds.), Reinforcement Learning and Approximate Dynamic Programming for Feedback Control, IEEE Press Computational Intelligence Series, 2012. DOI: 10.1002/9781118453988.ch14

- K Horak, Q. Zhu and B. Bosansky, “Manipulating Adversary's Belief: A Dynamic Game Approach to Deception by Design for Proactive Network Security, " 8th Conference on Decision and Game Theory for Security (GameSec), Oct. 23-25, 2017, Vienna, Austria.

- L. Huang and Q. Zhu, “Strategic Learning for Active, Adaptive, and Autonomous Cyber Defense, "Adaptive Autonomous Secure Cyber Systems, Jajodia, S., Cybenko, G., Subrahmanian, V.S., Swarup, V., Wang, C., Wellman, M. (Eds.), 2020.