Reinforcement Learning China Summer School

RLChina 2020

Multi-agent Systems

Bo AN (安波)
Nanyang Technological University

August 5, 2020

Outline

- Part 1-History and current status (10min)
- Part 2-Key research areas in MAS (30min)
- Part 3-Recent advances (40min)
  - *Computer poker*
  - *Game theory for security*
  - *Multi-agent RL*

专题 中国计算机学会通讯 第10卷 第9期 2014年9月

多智能体系统研究的历史、现状及挑战

安 波[1,2] 史忠植[2]
[1]新加坡南洋理工大学
[2]中国科学院计算技术研究所

关键词：智能体 多智能体系统 人工智能

多智能体系统 (multi-agent systems) 由分布式人工智能演变而来，其研究目的是解决大规模、复杂、实时和有不确定信息的现实问题，而这类问题是单个智能体所不能解决的。多智能体系统通常具有自主性、分布性、协调性等特征，并具有自组织能力、学习能力和推理能力。对多智能体系统的研究既包括构建单个智能体的技术，如建模、推理、学习及规划等，也包括使多个智能体协调运行的技术，例如交互通信、协调、合作、协商、调度、冲突消解等。随着互联网技术的发展，多智能体系统中往往会出现一些“自私”的智能体（如电子商务市场的交易方），因此需要引入博弈论来分析智能体的交互策略，其研究内容包括电子商务、拍卖、机制设计、社会选择理论等。

经过近 30 年的发展，多智能体系统已经成为国际人工智能领域的前沿和研究热点。在近年来的 AAAI 人工智能会议 (AAAI Conference on Artificial Intelligence) 和国际人工智能联合会议 (International Joint Conference On Artificial Intelligence, IJCAI) 上，

图1 多智能体系统创始人维克托·莱瑟

录用的关于多智能体系统的文章数量一直名列前两位。多智能体系统领域的创始人维克托·莱瑟 (Victor Lesser) 教授于 2009 年获得了 IJCAI 杰出研究奖 (IJCAI Award for Research Excellence)。自 1995 年以来，IJCAI 计算机与思维奖 (IJCAI Computers and Thought Award) 的获奖者有一半以上来自多智能体系统领域。

多智能体系统研究的历史

自 1956 年约翰·麦卡锡 (John McCarthy) 在著名的达特茅斯研讨会上提出“人工智能”这一概念后，“智能体”的概念便开始兴起。例如，阿兰·图灵 (Alan Turing) 提出了用来判断一台机器是否具备人类智能的“图灵测试”。在此测试中，测试者通过监视设备向被测试的实体（也就是我们现在所说的智能体）提问。根据图灵的观点，如果测试者无法区分被测试对象是计算机还是人，那么被测试对象就是智能的。人工智能泰斗马文·明斯基在他的《心智社会》这本书中将智能体描述为实现人类智能的

8

Points I want to make:

- ✧Mainstream AI research
- ✧Still at the very early age: Lots of work to be done and we can make a difference in both theory and applications

MAS Research: History

40-50s
Agent concept

1978
CMU workshop on distributed sensor networks

70s
Components for building agents (e.g., STRIPS)

1980
MIT workshop on distributed AI

80s
MAS research: coordination, cooperation, negotiation, organizations

1995
1st ICMAS, IFAAMAS

90s
BDI, KQML, AOP, DCOP, RoboCup

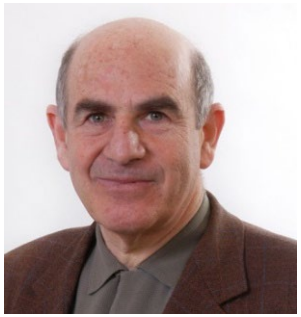
2002
AAMAS

00s
Game theory, social choice, TAC, DEC-MDP

10s
MAS+DRL

MAS: One of The Most Important Field in AI

- Over 30 years' history: Initially called **distributed AI**
- Now one of the most important/active fields of AI
  - *IJCAI Award for Research Excellence: Victor Lesser (09), Barbara Grosz (15)*
  - *IJCAI Computers and Thought Award:*
    - Sarit Kraus (95), Nicholas Jennings (99), Tuomas Sandholm (03), Peter Stone (07), Vincent Conitzer (11), Ariel Procaccia (15)





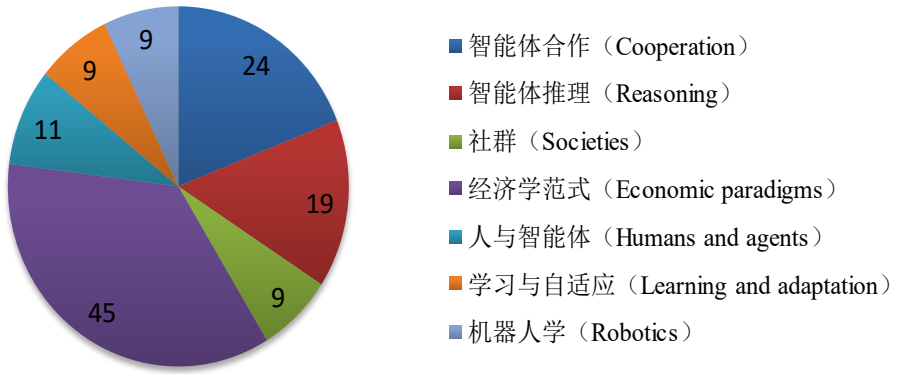



  - *Large number of accepted papers at AAAI/IJCAI*
  - *Researchers from top universities*
  - *Quality of AAMAS, JAAMAS*

MAS Research Areas

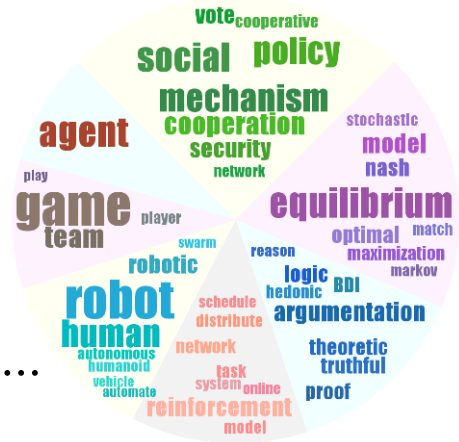
Theoretic foundations

- *Coordination, DEC-(PO)MDP*
- *DCOP*
- *Game theory*
  - mechanism design, coalition formation, social choice, negotiation
- *Multi-agent Learning*
- *Organizational design*

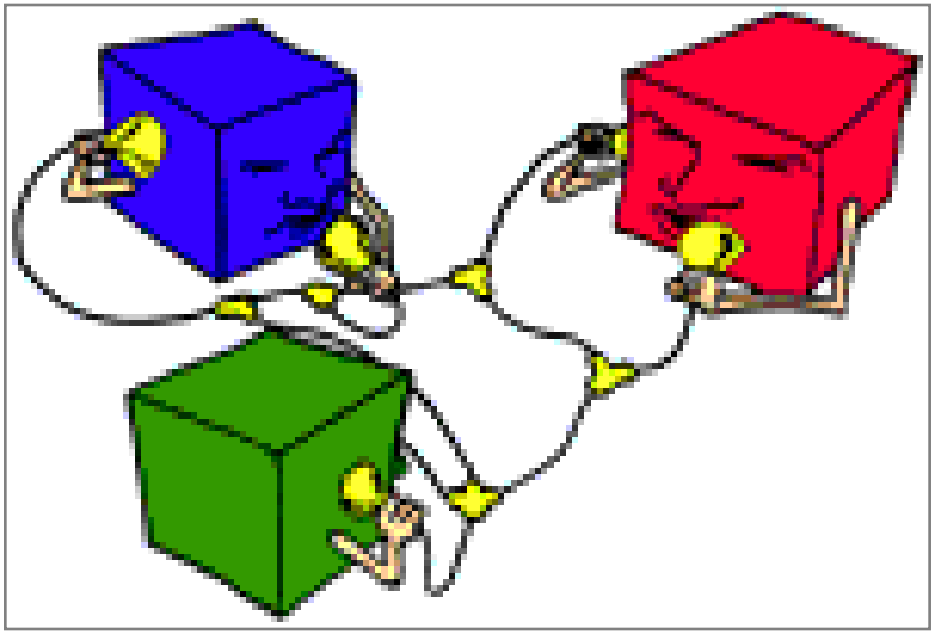


Applications

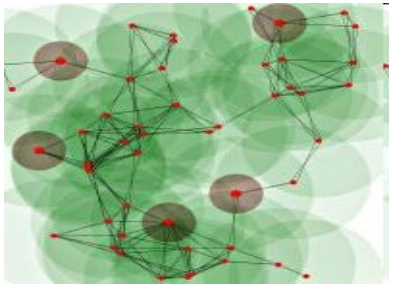
- *Robotics, security, sustainability, rescue, sensor networks, games…*



Cooperative Multi-Agent Systems:
Groups of Sophisticated AI systems that Work Together



- Open, dynamic, persistent systems
- Decentralized control
- Asynchronous
- Large scale (10s to 1000s)
- Partial observability
- No real-time global reward signal
- Communication delay



Partial Global Planning (PGP) [Lesser et al]

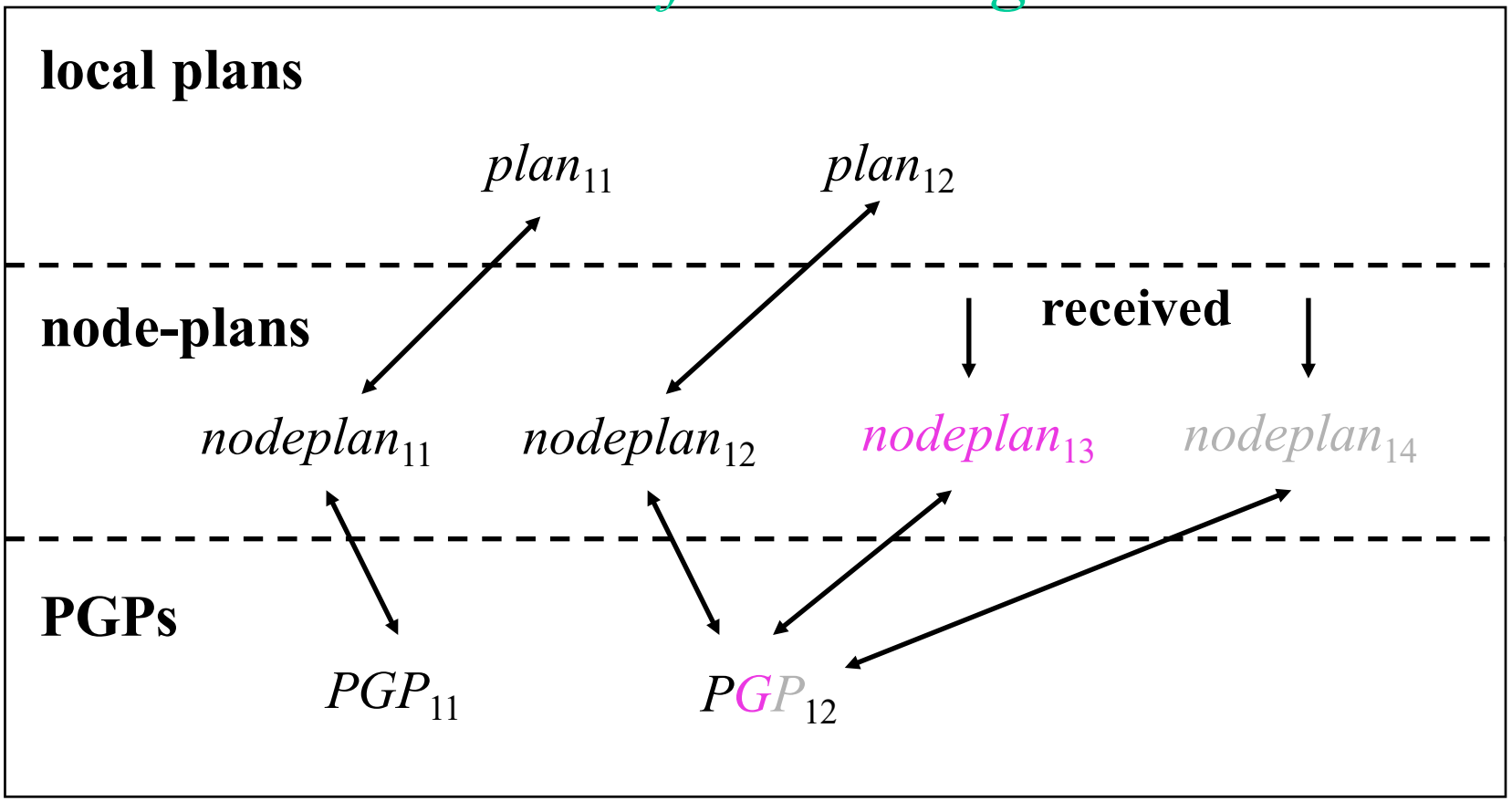
- Form a partial-global-plan (PGP) to achieve partial-global-goal
- PGP points to local models of participating plans

Process for Each Agent

Generation of local plans for an agent

Recognition of relationships among other agent plans

Modify plans to be coherent with other agents

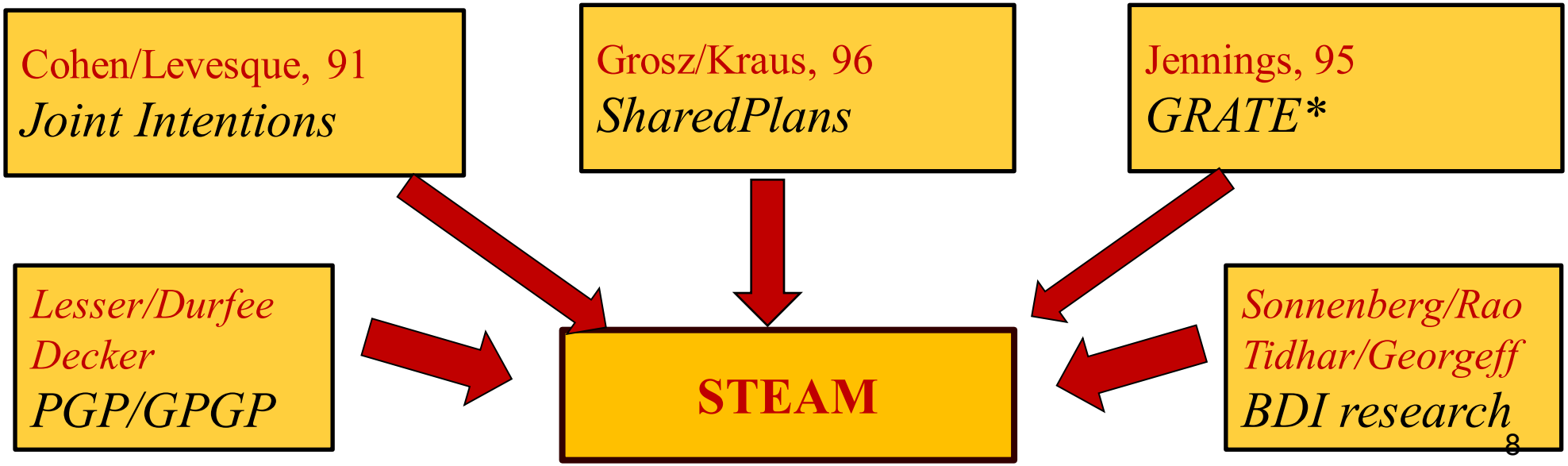


TÆMS, GPGP, distributed search

Further reading: Durfee, Lesser: Using Partial Global Plans to Coordinate Distributed Problem Solvers. IJCAI'87.

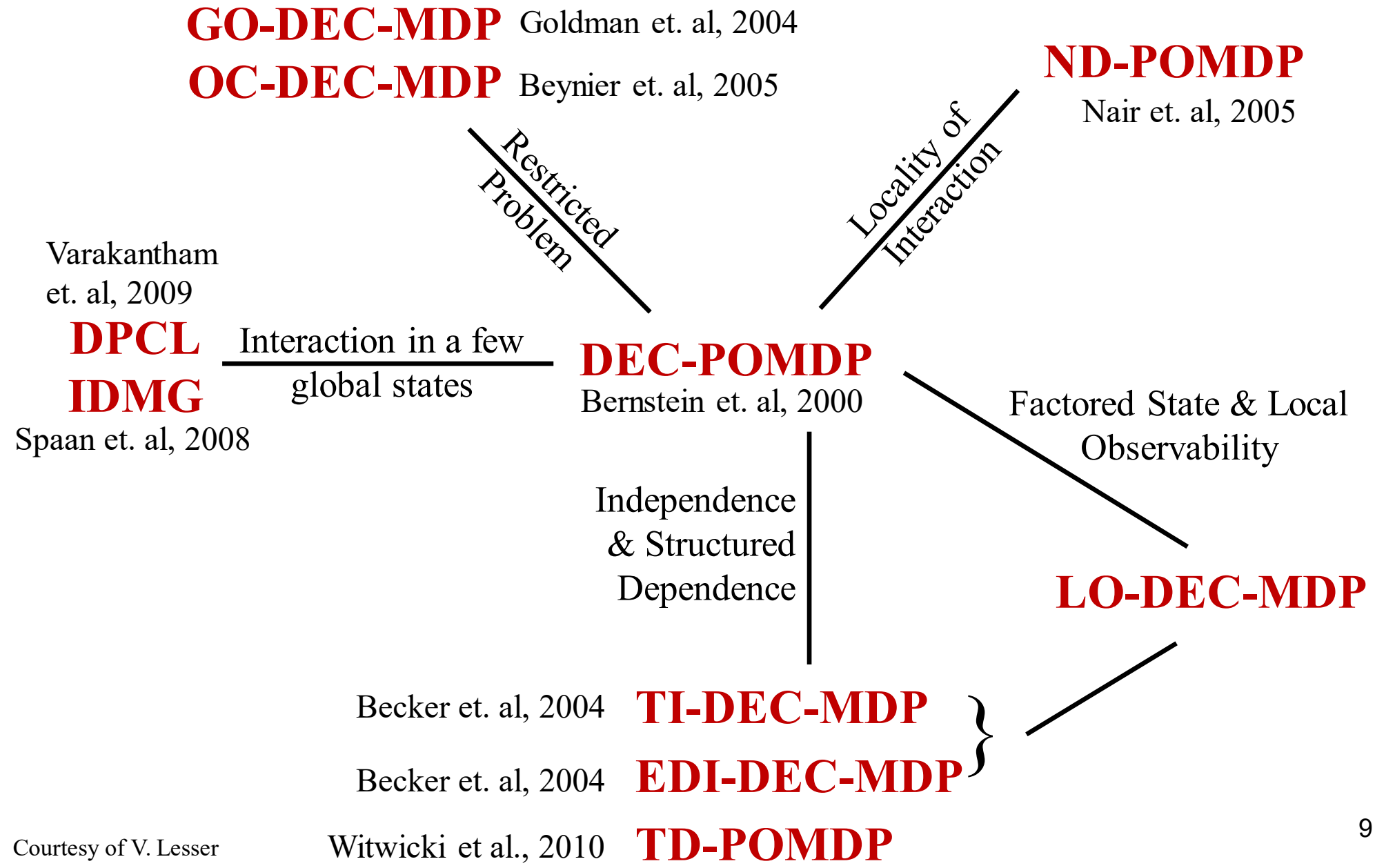
Towards Flexible Teamwork [Tambe]

- *Scout crashed, company waited forever*
- *Commander returned to home base alone…*
- Ad-hoc: No Framework to anticipate failures



Further reading: Tambe: Towards Flexible Teamwork. JAIR, 1997.

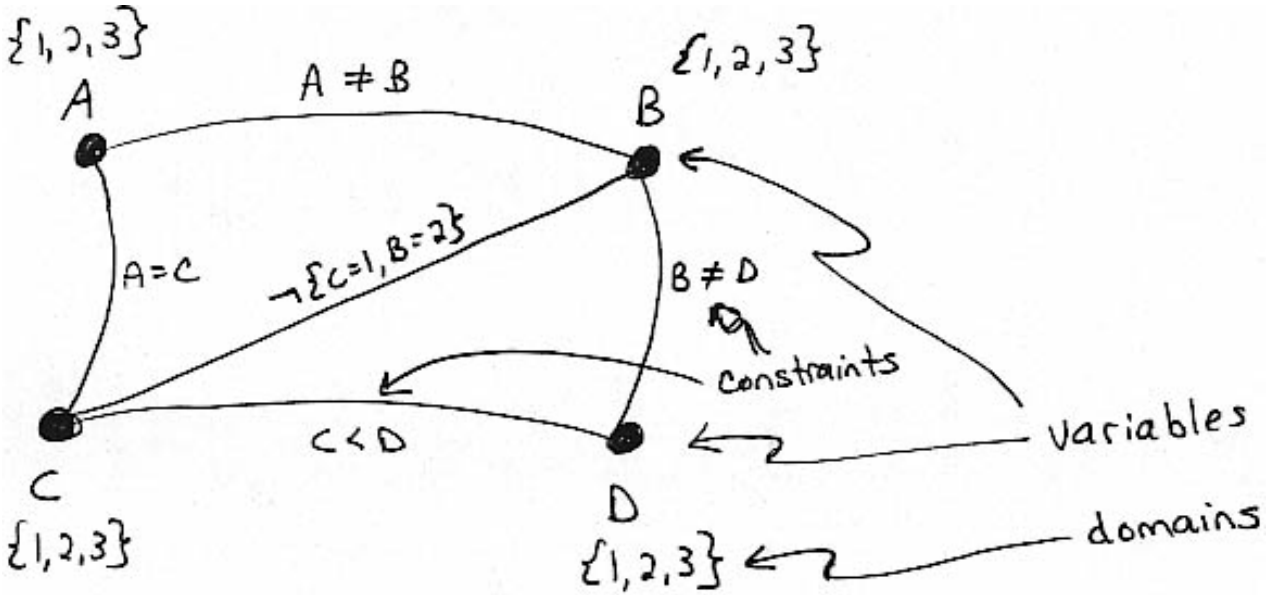
Dec-(PO)MDP Models



Courtesy of V. Lesser

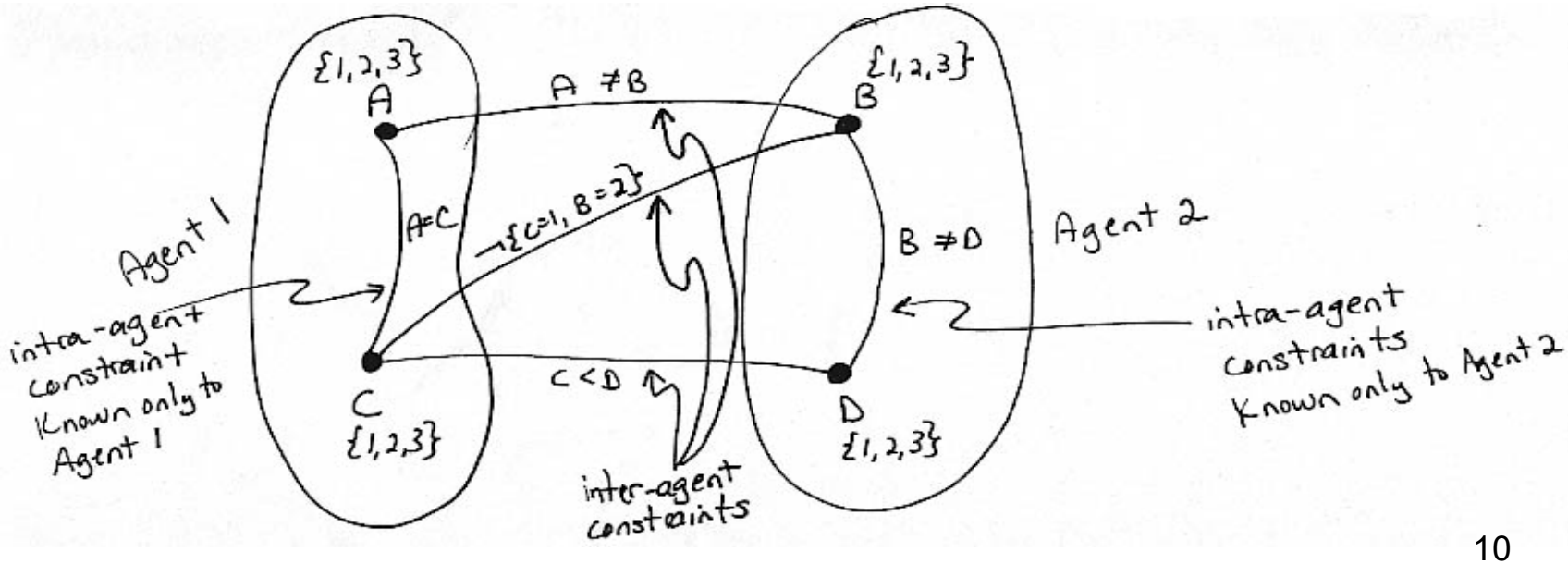
Further reading: Bernstein, Zilberstein, Immerman: The Complexity of Decentralized Control of Markov Decision Processes. UAI'00.

Distributed Constraint Satisfaction & Optimization



Applications:

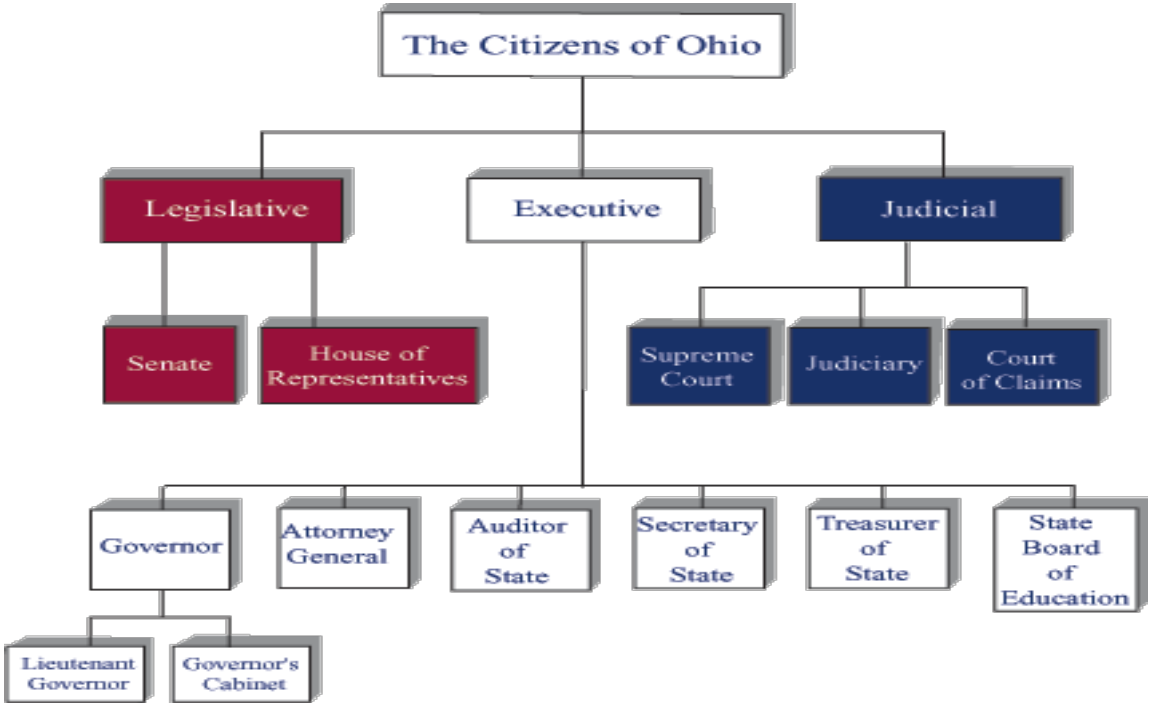
- Distributed sensor networks
- Resource allocation
- Meeting scheduling
- Engineering design



Further reading: Yeoh, Yokoo: Distributed Problem Solving. AI Magazine, 2012.

Organizational Design

- Human organizations
  - *Businesses*
  - *Governments*
  - *Universities*
- Long-term, repetitive, multi-participant



- Organizations give participants a set of responsibilities, goals, incentives and guidelines
- MAS organizations are the same
  - *Assignment of goals,incentives, roles, rights, authority, expectations, partners, rules, conventions…*

Further reading: Horling, Lesser: Using quantitative models to search for appropriate organizational designs. JAAMAS, 2008.

Game Theory for AI

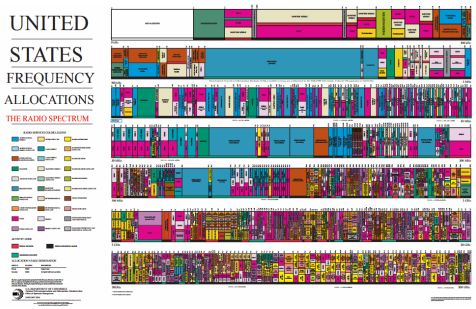
- GT analyzes *multi-agent* interaction (1940s-)



- AI: study and construction of *rational* agents [Russell & Norvig, 2003]

building a single agent (1950s-70s)	*Multi-agent systems (cooperative)* (1980s-)	*Multi-agent systems (competitive)* (1995-)

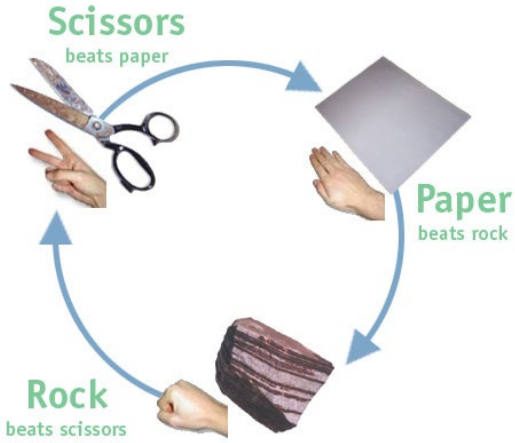
- GT for AI: success in computer poker, security, auction, GAN, …

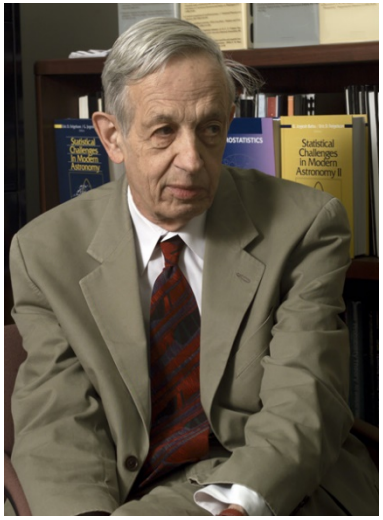

Google™ bing


Normal-form Game and Nash Equilibrium

- Players, strategies, payoffs

Player A \ Player B			
	0, 0	1, -1	-1, 1
	-1, 1	0, 0	1, -1
	1, -1	-1, 1	0, 0





- Nash Equilibrium: no agent has incentive to unilaterally deviate
  - *In any (finite) game, at least one Nash equilibrium (possibly mixed) exists* [Nash, 50]
  - *In 2-player zero-sum games, a profile is an NE iff both players play minimax strategies*
  - *Computing one (any) Nash equilibrium is PPAD-complete (even in 2-player games)* [Daskalakis, Goldberg, Papadimitriou 2006; Chen, Deng 2006]
  - *All known algorithms require exponential time (in the worst case)*
    - *Lemke-Howson, support enumeration*
- Mechanism design

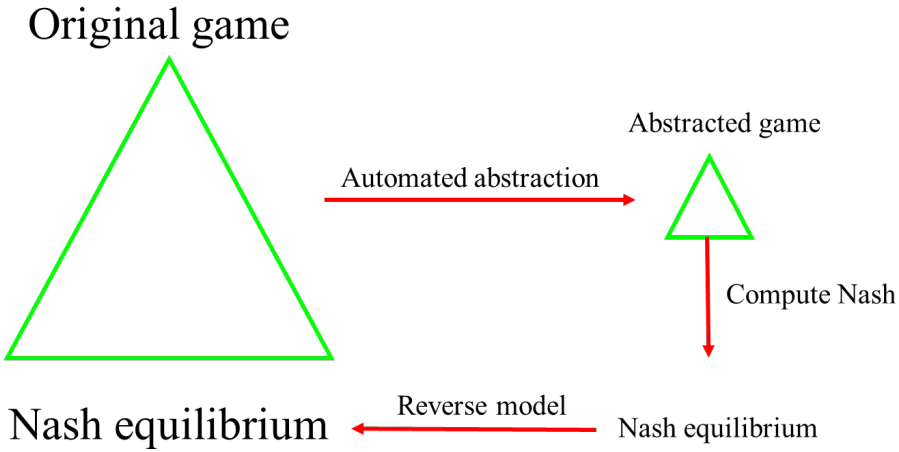
Further reading: Nisan, Roughgarden, Tardos, Vazirani: Algorithmic Game Theory. Cambridge University Press 2007.

Recent advances: Libratus and Pluribus



Abstraction (offline)

- action abstraction
- card abstraction
- took the game size from 10^{161} to 10^{12}



Equilibrium Finding (offline)

- CFR
- CFR+
- Monte Carlo CFR

Decomposition and Subgame Refinement (online)

- endgame solving
- subgame re-solving
- max-margin subgame refinement

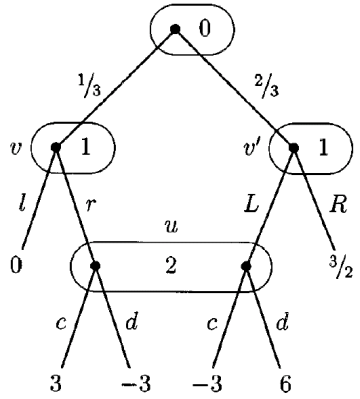
➢ Deep learning: Alberta's DeepStack, DeepMind

Regret and Regret Matching

- For each player i, action a_i and time t, define the regret $r_i(a_i, t)$ as

$$\Sigma_{1 \le t' \le t-1}(u_i(a_i, a_{-i,t'}) - u_i(a_{i,t'}, a_{-i,t'}))/(t-1)$$

- Regret matching: at time t, play an action that has positive regret $r_i(a_i, t)$ with probability proportional to $r_i(a_i, t)$
  - If none of the actions have positive regret, play uniformly at random
- **Theorem** [Hart & Mas-Colell 2000]: Informally, if we pick actions according to RM in a zero-sum game, the regret for any a_i is approaching 0 as t goes to infinity
- **Folk Theorem**: In a zero-sum game at time T, if both players' average overall regret is less than ε, then the average strategy is a 2ε equilibrium



	c	d
(l, L)	-2	4
(l, R)	1	1
(r, L)	-1	3
(r, R)	2	0

CFR: Regret Computing

- Key idea: minimize regret independently in each information set of an extensive-form game by maintaining the average immediate regret:

$$R_i^T(I,a) = \frac{1}{T}\sum_{t=1}^{T} \pi_{-i}^{\sigma^t}(I) \cdot \left(u_i(\sigma^t|_{I\to a}, I) - u_i(\sigma^t, I)\right)$$

  - π is the probability of reaching information set I given the strategies at time t and the player i is trying to reach that information;
  - μ is the utility of a strategy, including that we switch our action only at that information set

- Define the strategy at information set I and time $t+1$ as

$$\sigma_i^{t+1}(I,a) = \begin{cases} \dfrac{R_i^{t,+}(I,a)}{\sum_{a\in A(I)} R_i^{t,+}(I,a)} & \text{if } \sum_{a\in A(I)} R_i^{t,+}(I,a) > 0 \\ \dfrac{1}{|A(I)|} & otherwise. \end{cases}$$

Further reading: Zinkevich, Johanson, Bowling, Piccione: Regret Minimization in Games with Incomplete Information. NIPS'07.

Game Theory for Security [50+@ AAAI, AAMAS, IJCAI, ICML, NeurIPS]

- Global challenges for security

Boston Marathon bombings

French oil tanker hit by a boat

Cyber physical attacks

- Security resource allocation
  - *Limited security resources*
  - *Adversary monitors defenses, exploits patterns*
- We pioneered the first set of applications of game theory for security resource scheduling(2007-)

- 40+ papers at premier conferences/journals, 2 best paper awards
- INFORMS Daniel H. Wagner Prize for Excellence in Operations Research Practice (2012), etc
- Operational Excellence Award from US Coast Guard (2012), etc
- Media reports: FOX News, CNN News, Federal News Radio, Defense News, The Economics Times News, Los Angeles Times, etc
- United States congressional hearing (4 times)

Security Games

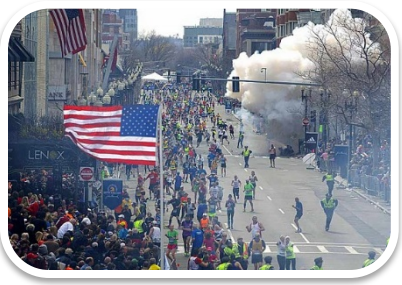
- Security allocation: (i) Target weights; (ii) Opponent reaction
- Stackelberg game: Security forces commit first
- Strong Stackelberg Equilibrium
- Our contributions:
  - *Algorithms for solving large scale games*
  - *Learning adversary behavior*
  - *Applications in the real world*

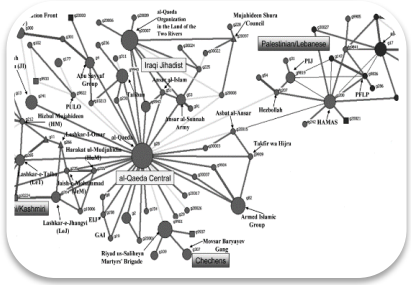
Attacker

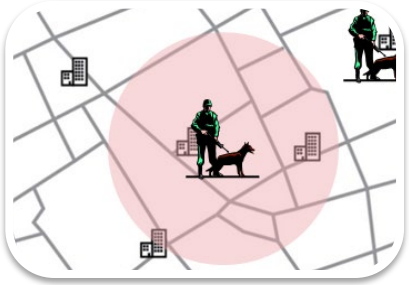
Defender

	Attacker Action #1	Attacker Action #2	……	……
Defender Action #1	7, -3	-6, 2	……	……
Defender Action #2	-3, 4	4, -6	……	……
……	……	……	……	……
……	……	……	……	……

Analyzing Complex Security Games [2016-]


Dynamic Payoff

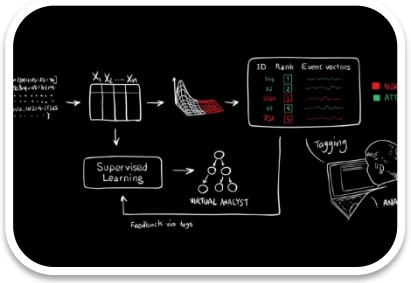

Network Games


Protection Externality

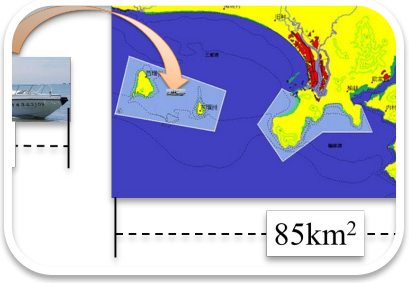
Uncertainty

Strategic Deception

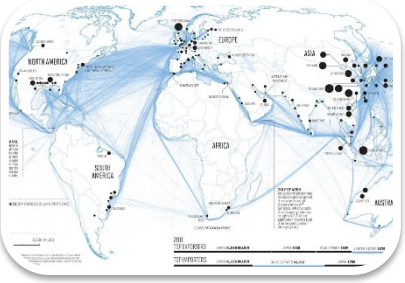

Cyber Security



Adversarial Machine Learning



Coral Reef


Nuclear Smuggling

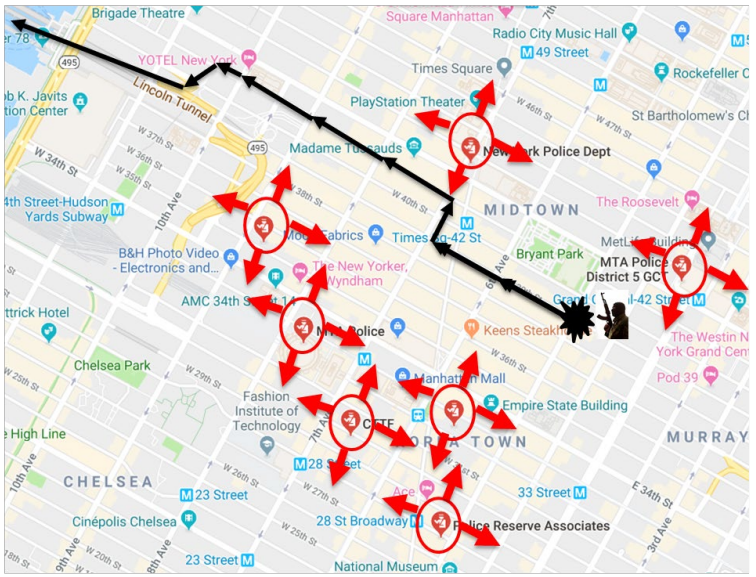
Elections

- Combining techniques from AI, Game Theory, Operations Research …
- Marry theory with practice
- Approaches can be applied to other domains *With proper tuning & extension*
  - *Incremental strategy generation*
  - *Construct (multiple) equivalent games*
  - *Exploit compact representation*
  - *Abstraction*
  - *Tradeoff between optimality and efficiency*
  - *Approximation*

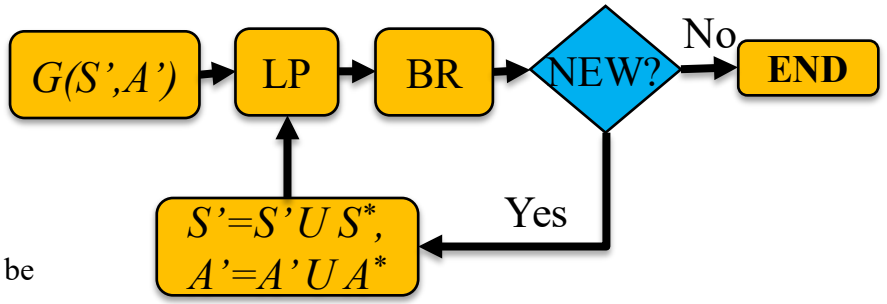
Further reading: An: Game theoretic analysis of security and sustainability. IJCAI'17.

Converging to Team-Maxmin Equilibria in Multiplayer Norm-Form Games [ICML'20]

- Equilibria in Multiplayer Games
  - *Hard to compute: PPAD-Complete*
  - *Hard to select: NEs are not unique*
  - *Few results:*
    - Special structure: congestion games
    - No theoretical guarantee: Pluribus (Brown and Sandholm 2019)
- Team-Maxmin Equilibria (von Stengel and Koller 1997)
  - *A team of players independently plays against an adversary*
  - *Unique in general*
  - *FNP-hard to compute a team-maxmin equilibrium*
    - Formulated as a non-convex program
    - Solved by a global optimization solver
- Converging to Team-Maxmin Equilibria
  - *Existing ISG for multiplayer games*
    - Converge to an NE but many not to a TME
    - Difficult to extend the current ISG to converge to a TME
  - *ISGT: the first ISG guaranteeing to converging to a TME*
    - Conditions in ISGT cannot be further relaxed
  - *CISGT: further improve the scalability*
    - Initialize the strategy space by computing an equilibrium that is easier to be computed



Incremental Strategy Generation (ISG)



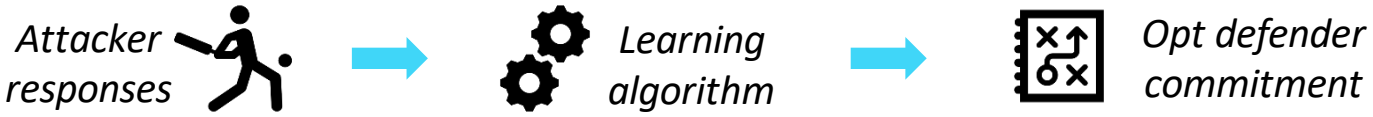
L×W	5×5	5×5	5×5	5×5	5×5	4×4	6×6	8×8	10×10
(p,q)	(0.8,0.6)	(0.7,0.5)	(0.6,0.4)	(0.5,0.3)	(0.4,0.2)	(0.4,0.2)	(0.4,0.2)	(0.4,0.2)	(0.4,0.2)
FullTME		∞	448s	50.4s	17.8s	0.3s	∞		
ISGT					>1000s	4s	>1000s		
CISGT	9.8s	5.9s	4.7s	3.7s	2.3s	2.2s	8.3s	24s	57s

Table 1. Computing TMEs: ∞ represents out of memory.

Manipulating a Learning Defender and Ways to Counteract [NeurIPS'19]

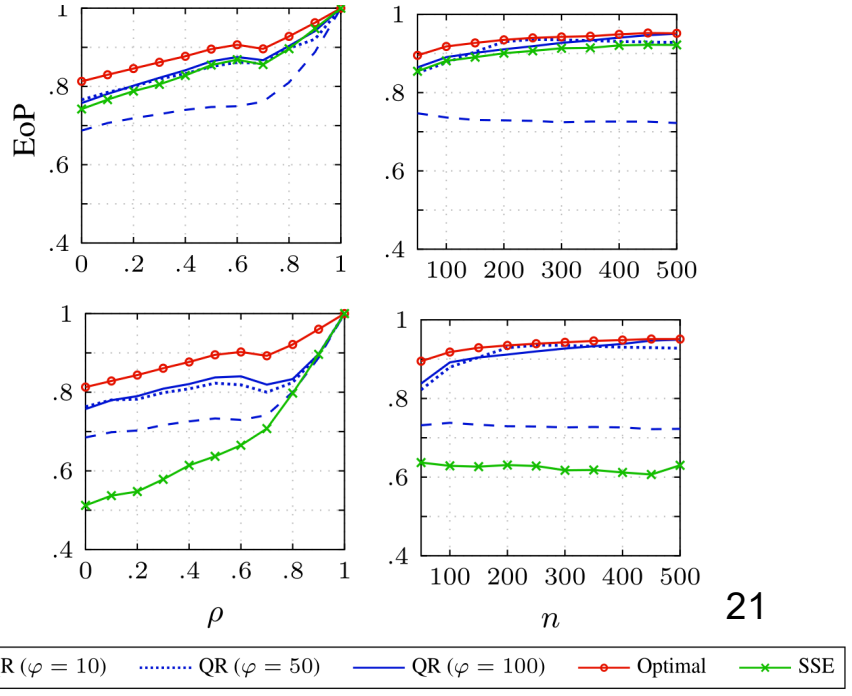
- Learn to play optimally in a Stackelberg Security Game (SSG)

[Letchford et al., 2009; Blum et al., 2014; Haghtalab et al., 2016; Roth et al., 2016; Peng et al., 2019]



- Unrealistic assumption about ***truthful*** attacker responses → *What if **untruthful**?*

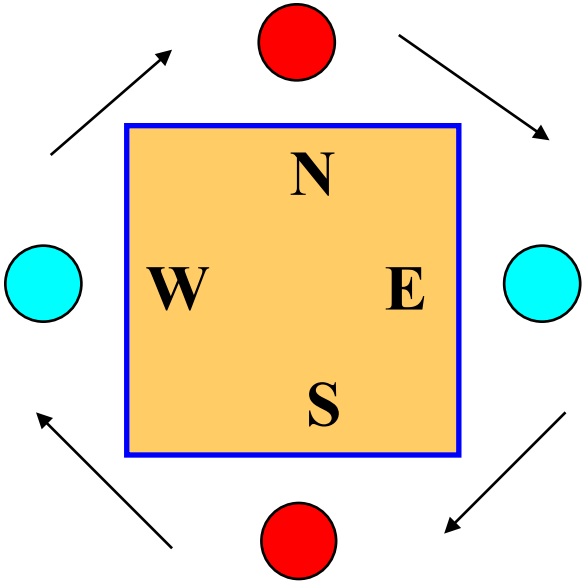
- In our work:
  - *Learning algorithms can be easily manipulated by untruthful attacker*
  - *Often optimal for attacker to deceive defender into playing a zero-sum game*
  - *A policy-based framework to play against attacker deception*
  - *A poly-time algorithm to compute optimal policy and a heuristic approach for infinite attacker types*

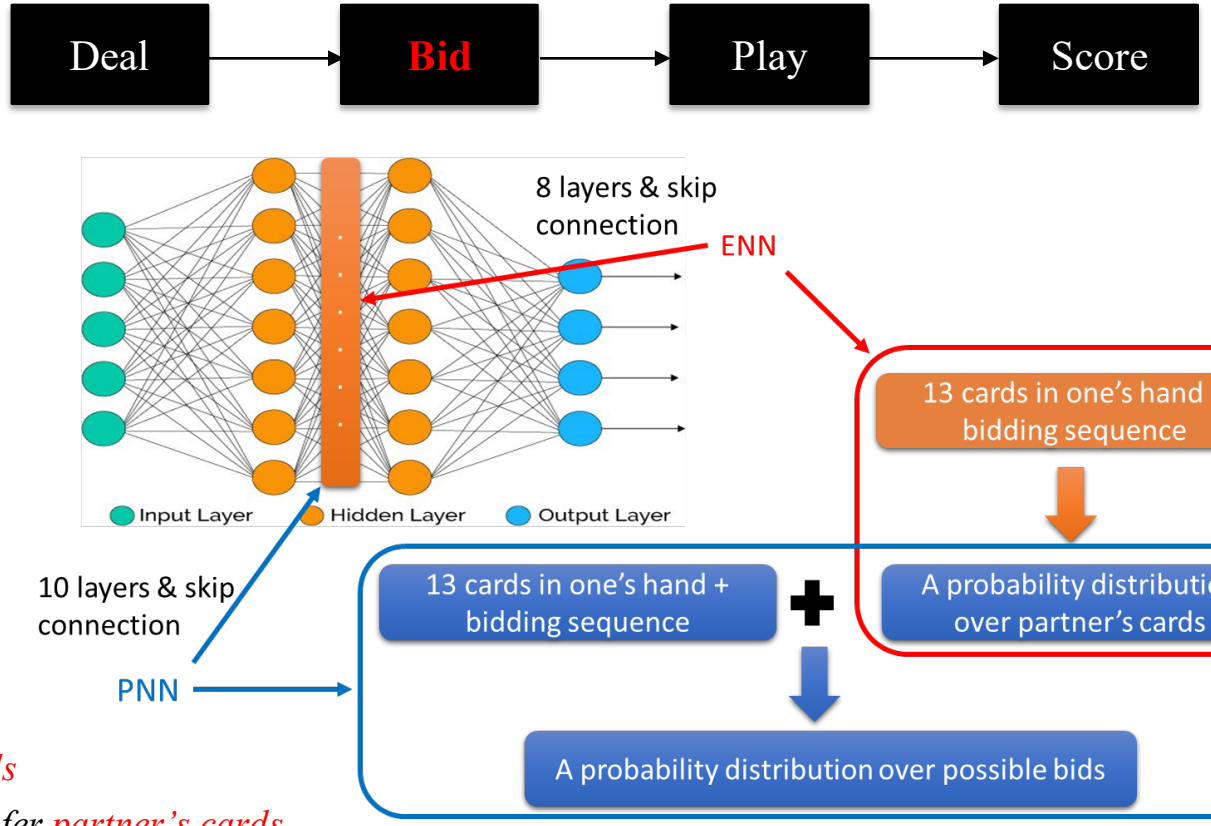


When do We Need RL?

- RL might be more appropriate when
  - *Problem cannot be well modelled*
  - *Large scale*
  - *Non-convex and cannot be approximated*
  - *No domain structures can be exploited*
- Multi-agent RL is receiving increasing attention
  - *Agents collaborate/compete with each other without the coordinator*
  - *Often centralized training and decentralized execution (CTDE)*
  - *MADDPG, COMA, VDN, QMIX …*
  - *More resources:*
    - DeepMind tutorial: https://www.karltuyls.net/wp-content/uploads/2020/06/MA-DM-ICML-ACAI.pdf
    - Thore Graepel (DeepMind) AAMAS'20 Keynote: https://underline.io/lecture/63-automatic-curricula-in-deep-multi-agent-reinforcement-learning
- Does not mean multi-agent RL can always work!
- Rest of the lecture: quick overview of some of our recent works on RL
  - *Game, security, e-commerce, urban planning: competitiveness - data rich/poor*

Competitive Bridge Bidding with Deep Neural Networks [AAMAS'19]





- Imperfect information
  - *Not necessary to infer opponents' cards*
  - *Estimation neural network (ENN) to infer partner's cards*
- Cooperation & competition
  - *Train the policy neural network (PNN) & ENN simultaneously*
  - *Maintain an opponent pool for competitors*
  - *Self-play with aapproximate reward from DDA*
- Large state space
  - *Efficient bidding sequence representation*

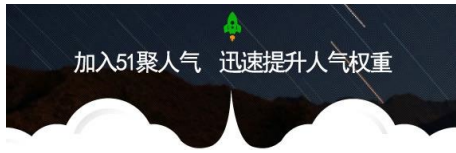
	SL-PNN	RL-PNN	SL-PNN+ENN	RL-PNN+ENN	Wbridge5
SL-PNN	N/A	-8.7793	-5.653	-9.2957	–
RL-PNN	8.7793	N/A	2.1006	-1.0856	–
SL-PNN+ENN	5.653	-2.1006	N/A	-2.2854	–
RL-PNN+ENN	9.2957	1.0856	2.2854	N/A	6.25

test on 10 thousand boards

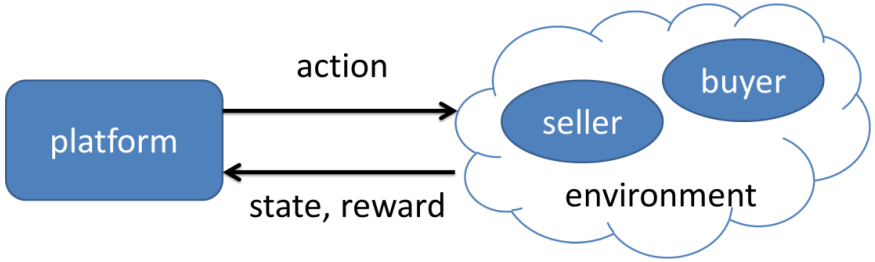
test on 64 boards manually

Impression Allocation for Combating Fraud in E-Commerce via Deep RL [IJCAI'18]

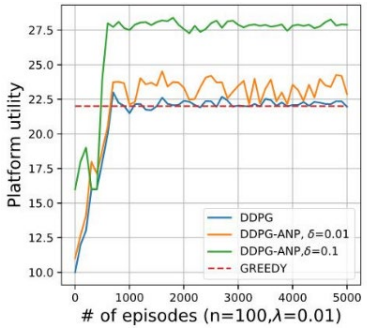
- **Fraud transactions in e-commerce:**
  - *Sellers buy their own products to fake popularity*
  - *Deliberate, large scale fraud: Grey industry*
  - *Existing approach: Machine learning for fraud detection*



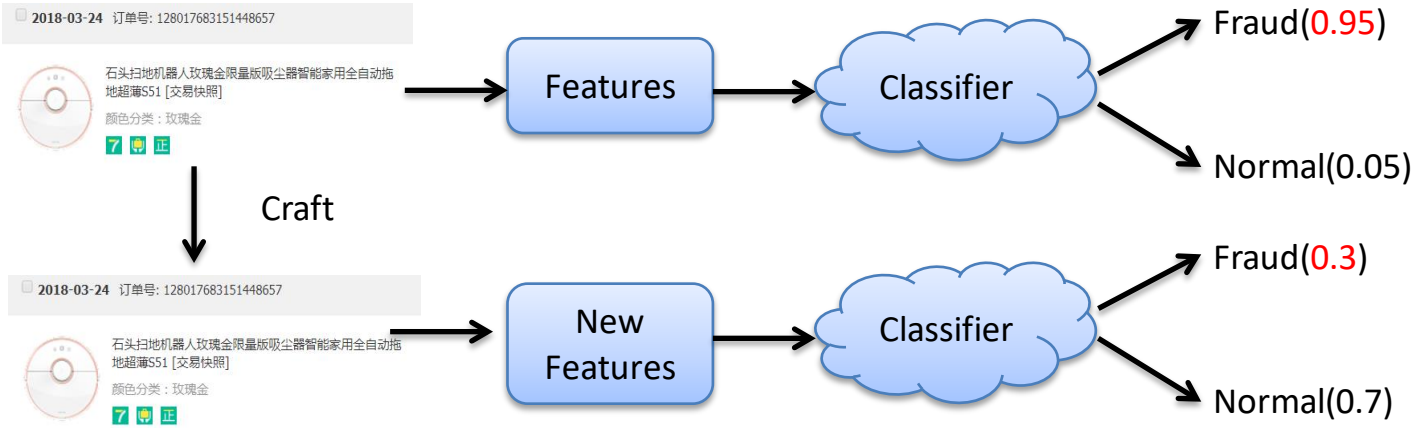
- **Reducing fraud by optimal impression allocation**
  - *Intuition: reward honest sellers and penalize cheating sellers*
  - *MDP model:*

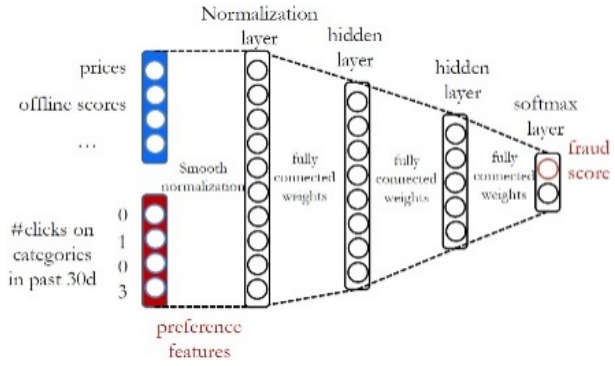


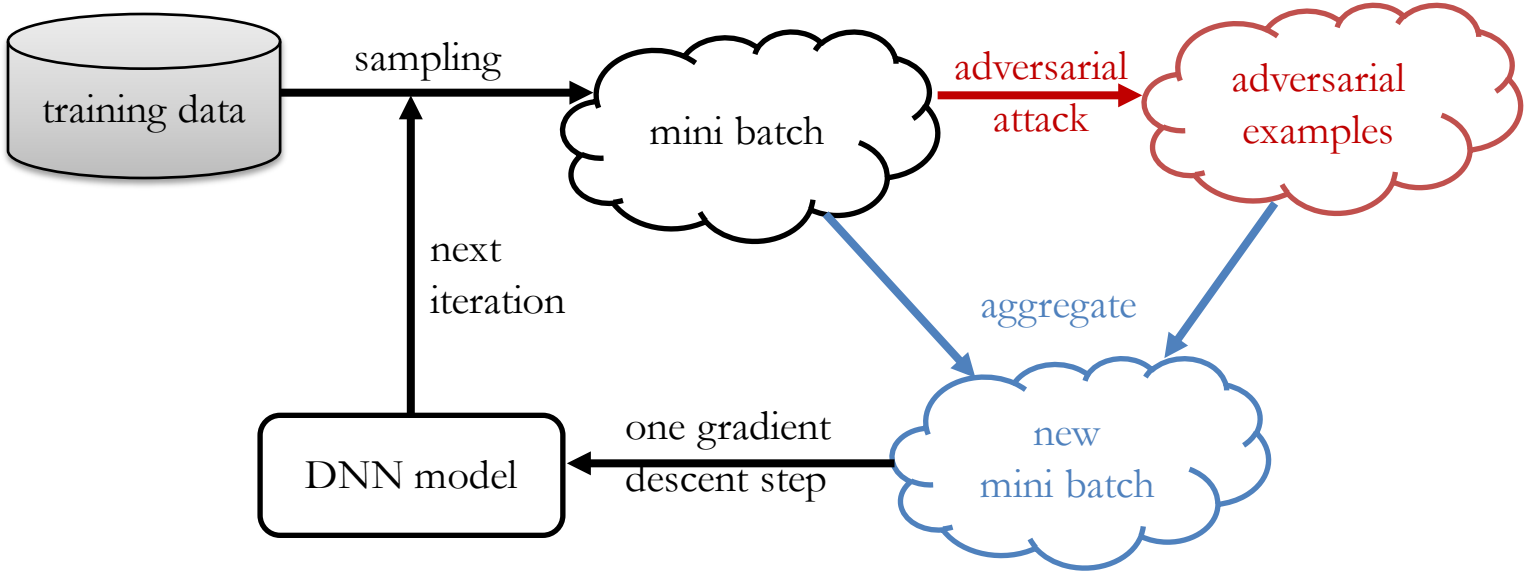
  - *Solving the MDP: Deep Deterministic Policy Gradient (DDPG) + reward shaping*
    - *Shaped reward*: $R(M^t, \boldsymbol{w}^t) = \frac{1}{n}\sum_{i=1}^{n}\left(r_i^{t+1} - \lambda f_i^{t+1}\right) * price_i^{t+1} - \delta||\boldsymbol{w}^t||_2$
    - Avoid too large action ($\boldsymbol{w}^t$)
    - Make the reward signal continuous
    - Outperform all existing approaches using Alibaba's real data

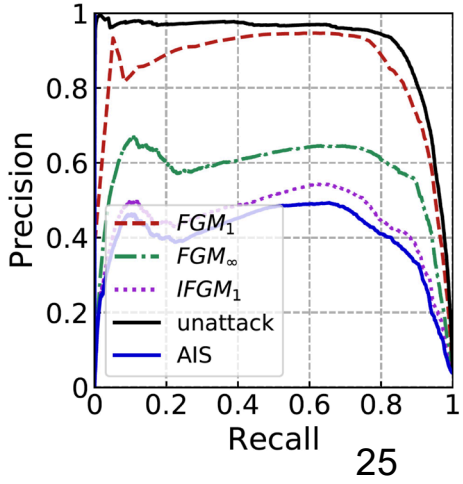


Improving Robustness of Fraud Detection Using Adversarial Examples [WWW'19]









Dynamic Electronic Road Pricing (ERP) [AAAI'18]

Woodsville Tunnel (71)

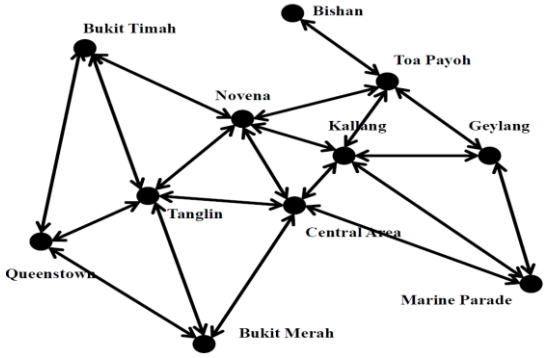
Cars/Light Goods/Taxis (Weekdays)

07:00 - 07:30	$0.00
07:30 - 08:30	$0.50
08:30 - 08:35	$1.00
08:35 - 08:55	$1.50
08:55 - 09:00	$1.00
09:00 - 09:30	$0.50
09:30 - 22:30	$0.00

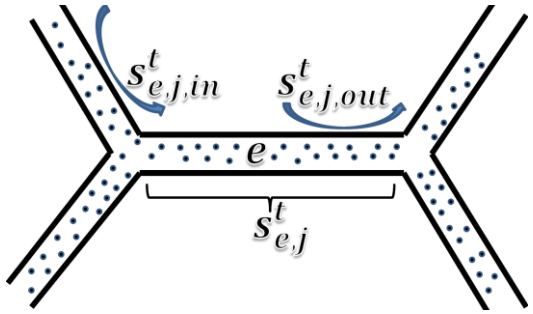
Static!!

Dynamic ERP

Abstracted road network



An MDP formulation (state transition)



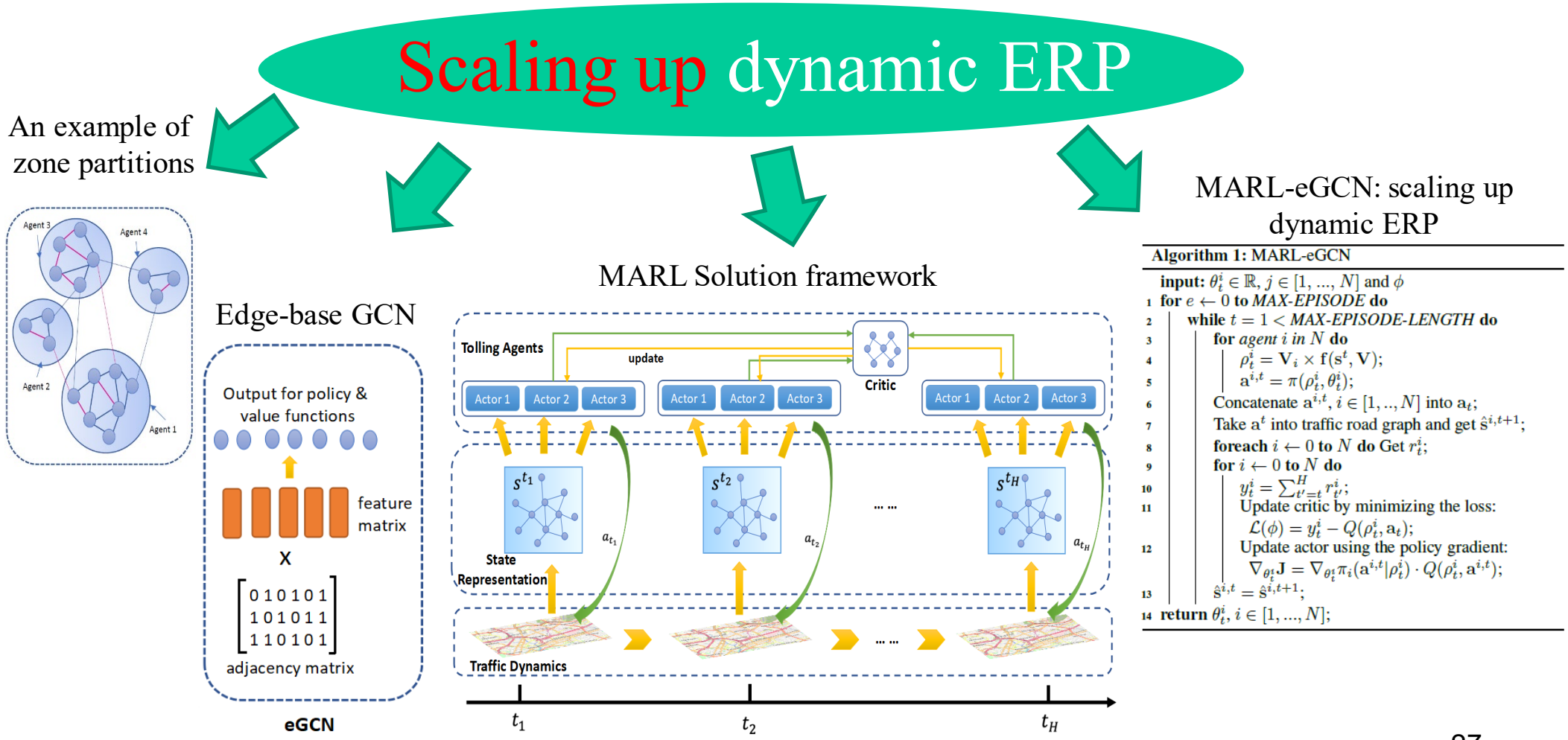
PG-β: scalable RL based on function approximation

Algorithm 1: PG-β

Initialize $\vartheta^t \leftarrow \vartheta_0, \boldsymbol{\theta}^t \leftarrow \boldsymbol{\theta}_0, \forall t = 0, 1, \ldots, H-1$;
repeat
  Generate an episode $\mathbf{s}^0, \mathbf{a}^0, R^1, \ldots, \mathbf{s}^{H-1}, \mathbf{a}^{H-1}, R^H$;
  for $t = 0, \ldots, H-1$ **do**
    $Q^t \leftarrow \sum_{t'=t}^{H} R^\tau$;
    $\delta \leftarrow Q^t - \hat{v}(\mathbf{s}^t, \mathbf{a}^t, \vartheta^t)$
    $\vartheta^t \leftarrow \vartheta^t + \beta \delta \nabla_{\vartheta^t} \hat{v}(\mathbf{s}^t, \vartheta^t)$;
    $\boldsymbol{\theta}^t \leftarrow \boldsymbol{\theta}^t + \beta' \gamma^t \delta \nabla_{\boldsymbol{\theta}^t} \log \pi^t(\mathbf{a}^t | \mathbf{s}^t, \boldsymbol{\theta}^t)$
until *converge*
return $\boldsymbol{\theta}^t, \forall t = 0, 1, \ldots, H-1$;

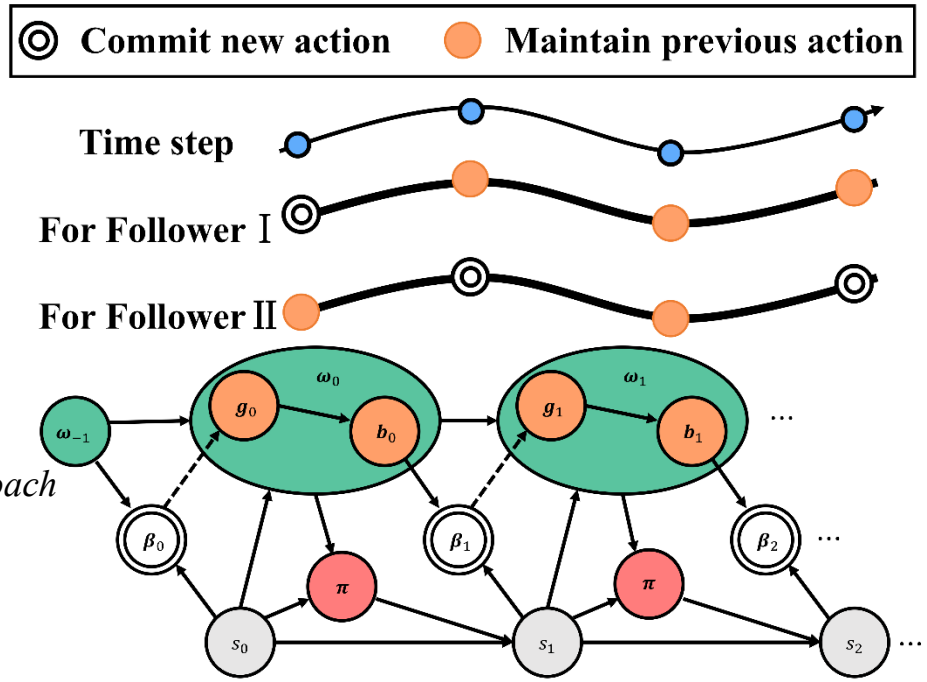
Scaling up Dynamic Electronic Road Pricing [IJCAI'19]

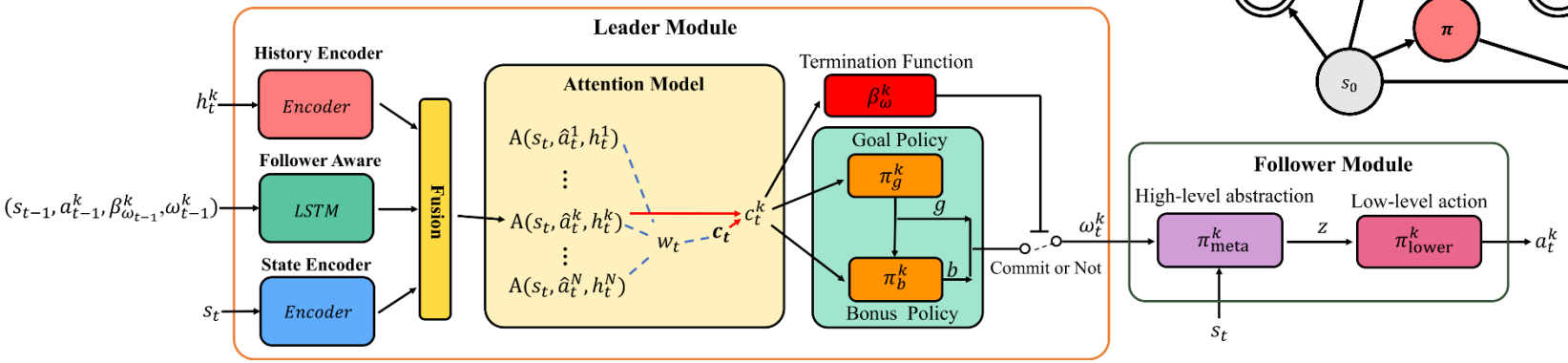
- Multi-agent reinforcement learning (MARL) for scaling up
- Edge-based graph convolutional network (GCN) for domain feature learning

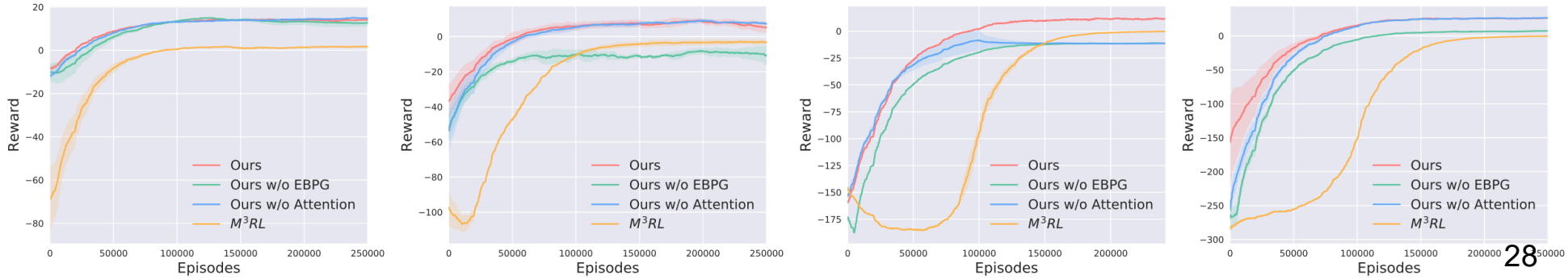


Learning Expensive Coordination: An Event-Based Deep RL Approach [ICLR'20]

- Large state space based expensive coordination
  - *Considering the followers are self-interested*
  - *The leader should coordinate them by assigning incentives*
- Event-based policy gradient
  - *Modeling the leader's strategy as events*
  - *A novel event-based policy is induced based on the events*
- Action abstraction for followers
  - *Accelerating the training process through action abstraction approach*



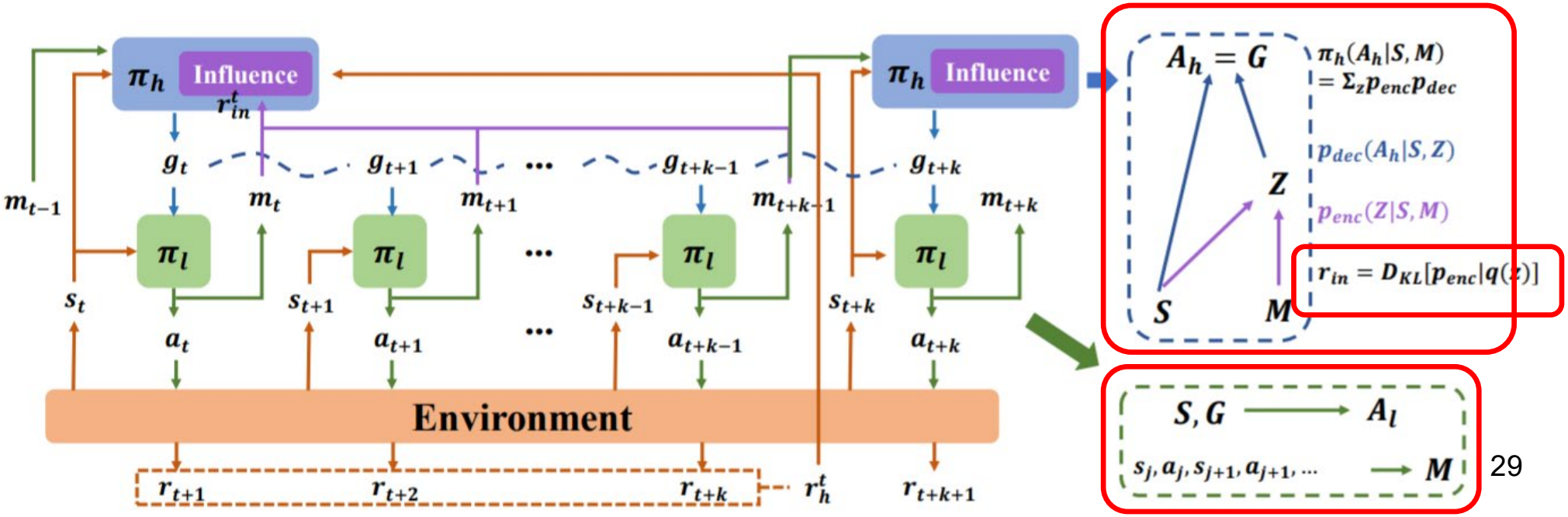




(a) Resource Collections. (b) Multi-Bonus Resource (c) Navigation. (d) Predator-Prey.

Interactive Influence-based HRL [IJCAI'20]

- Challenge: high-level policy suffers the non-stationarity problem
  - *the constantly changing low-level policy leads to different state transitions and rewards of high-level policy*
- Key idea: the high-level policy makes decisions conditioned on the low-level policy representation
- Solutions:
  - *Low-level policy modeling*
  - *High-level policy stabilization via information-theoretic regularization*
  - *Influence-based adaptive exploration with auxiliary rewards*

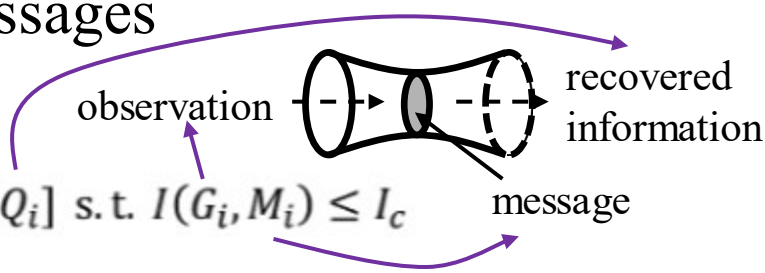


Efficient Multi-agent Communication [ICML'20]

- Limited bandwidth in multi-agent communication:




- Theorem: limited bandwidth constrains the message's entropy
- Key idea: compressing the communication messages
- Information bottleneck principle

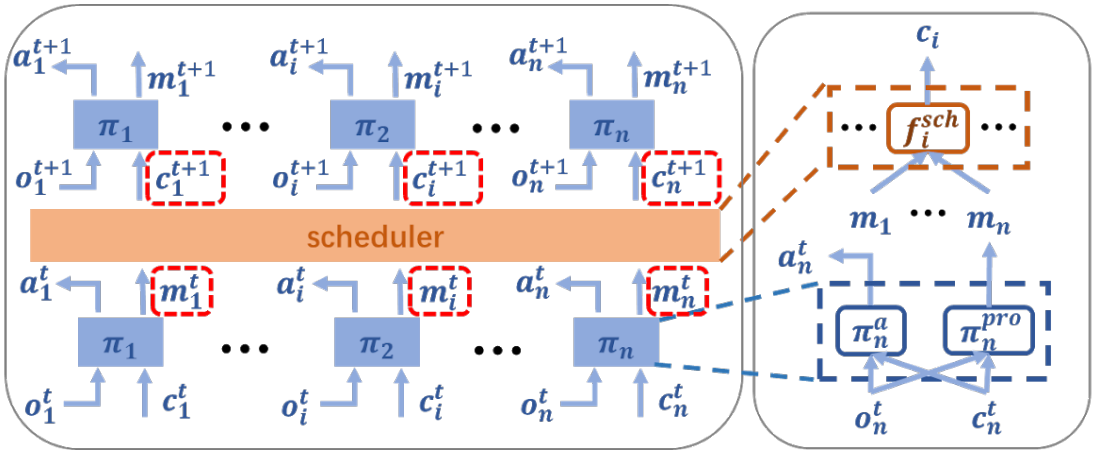


$$\max_{T} I(T,Y) \text{ s.t. } I(X,T) \le I_c \longrightarrow \max_{M_i} J(\theta_i) := E_{\pi_i}[Q_i] \text{ s.t. } I(G_i,M_i) \le I_c$$

  - *lower bound for control entropy*

$$\tilde{J}(\theta_i) = E_{\pi_i}[Q_i] - \beta E\left[D_{KL}\left[\pi_i^{pro}(m_i|g_i)||z(m_i)\right]\right]$$

- Architecture:



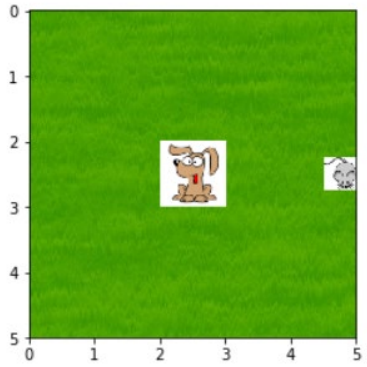
Learning Behaviors with Uncertain Human Feedback [UAI'20]

Uncertain Human Feedback

- Human may not give any feedback
- Positive feedback → sub-optimal action
- Negative feedback → optimal action

$$p(a, \lambda(s); \sigma, \mu, f) = \begin{cases} p^+(a, \lambda(s); \sigma, \mu^+) & f = f^+ \\ p^-(a, \lambda(s); \sigma, \mu^-) & f = f^- \\ 1 - p^+(a, \lambda(s); \sigma, \mu^+) & \\ \quad -p^-(a, \lambda(s); \sigma, \mu^-) & f = f^0 \end{cases}$$

p^+ and p^- are functions like Gauss function where λ is the optimal action, μ and σ are unknown parameters controlling mean and variance.



Rat catching

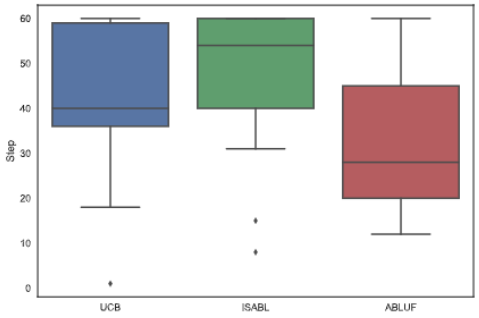
Light controlling

Inferring Optimal Behavior

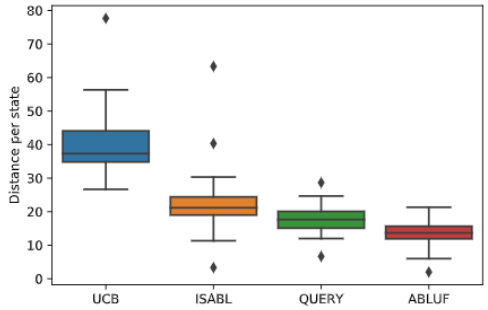
- Maximizing likelihood estimation of receiving different kinds of feedback:

$$\arg\max_\lambda P(h|\lambda; \mu, \sigma)$$

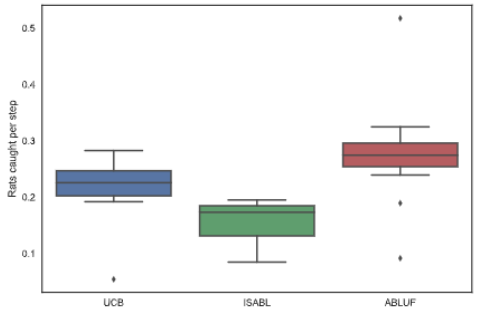
- EM + GD:
  - EM updates λ with σ fixed
  - μ is latent variable in integral
  - GD updates σ with λ fixed
  - μ is eliminated by a trick



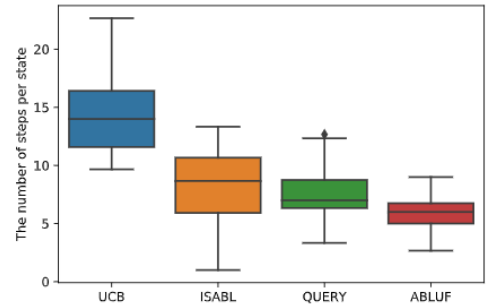
(a) Step



(a) Distance/state



(b) Rats caught per step

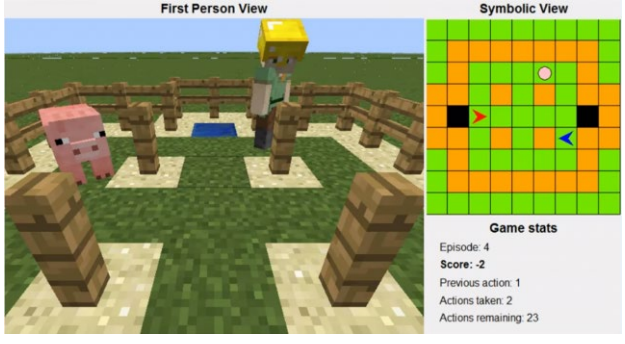


(b) #steps/state

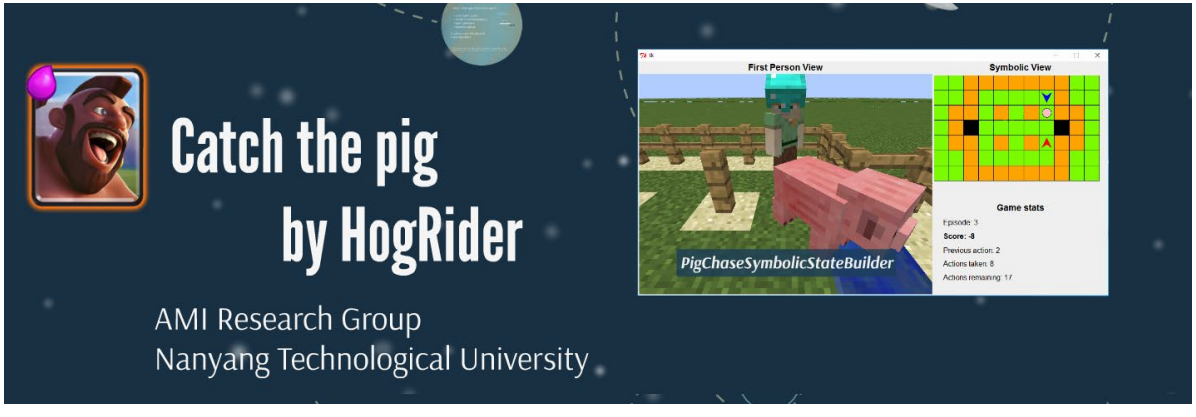
We Won Microsoft Collaborative AI Challenge

[AAAI'18]

- Collaborative AI
  - *How can AI agents learn to recognize someone's intent (that is, what they are trying to achieve)?*
  - *How can AI agents learn what behaviors are helpful when working toward a common goal?*
  - *How can they coordinate or communicate with another agent to agree on a shared strategy for problem-solving?*
- Microsoft Malmo Collaborative AI Challenge
  - *Collaborative mini-game, based on an extension "stag hunt"*
  - *Uncertainty of pig movement*
  - *Unknown type of the other agent*
  - *Detection noise (frequency 25%)*
  - *Efficient learning*



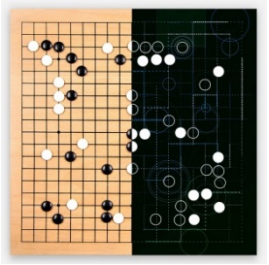
- Our team HogRider won the challenge (out of more than 80 teams from 26 countries)
  - *learning + game theoretic reasoning + sequential decision making + optimization*



AI for Complex Interaction

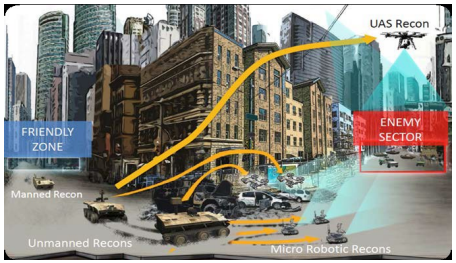
- Recent AI breakthrough

IMAGENET



- What's next: AI for ***complex*** interaction

- Stochastic, open environment
- Multiple players
- Sequential decision, online
- Strategic (selfish) behavior
- Distributed optimization



Google

bing

- MARL will play a very important role…