Control of networked coordination games

Ming Cao

Faculty of Science and Engineering
& School of Digital Society, Technology and AI
University of Groningen
The Netherlands

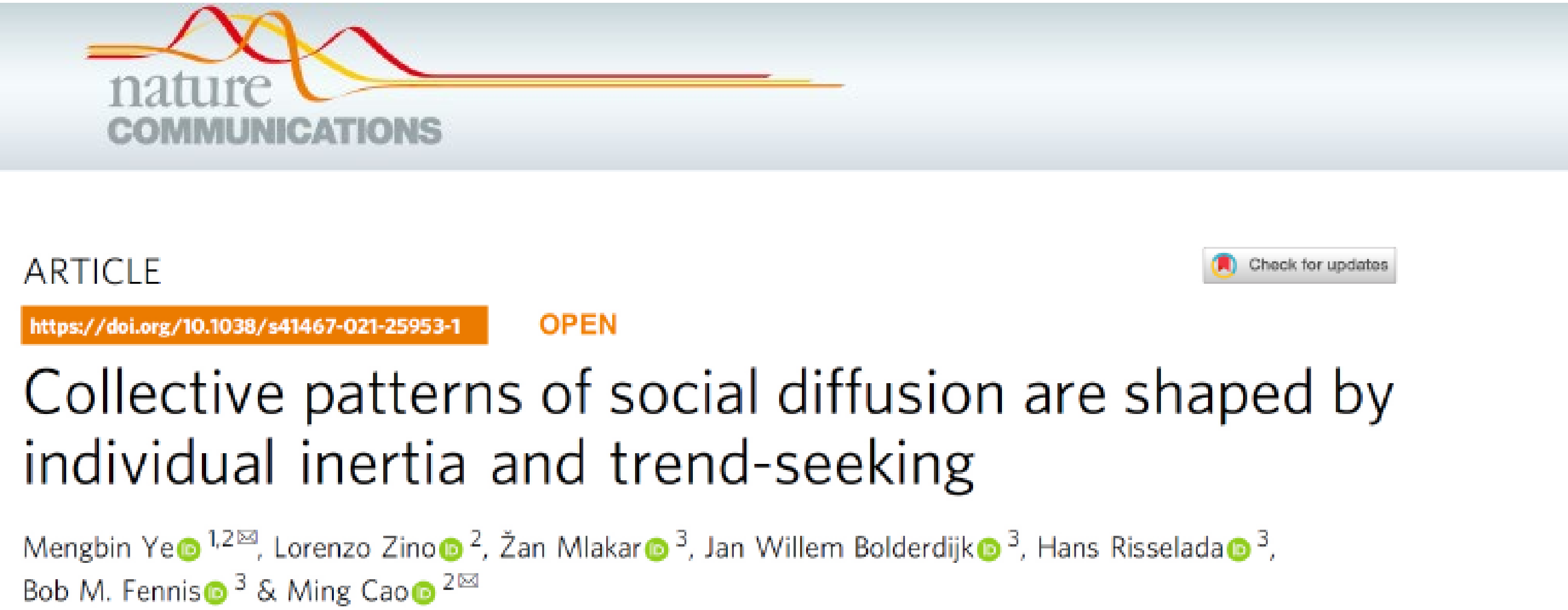
Coordination of self-interested agents

Sociology & Economics: social dilemma in modern society

Biology: understanding cooperating behavior in social animals

Key difference from controlling complex engineering systems

- incentives to change rather than commands
- system dynamics co-evolve with changing environment



nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-021-25953-1 OPEN

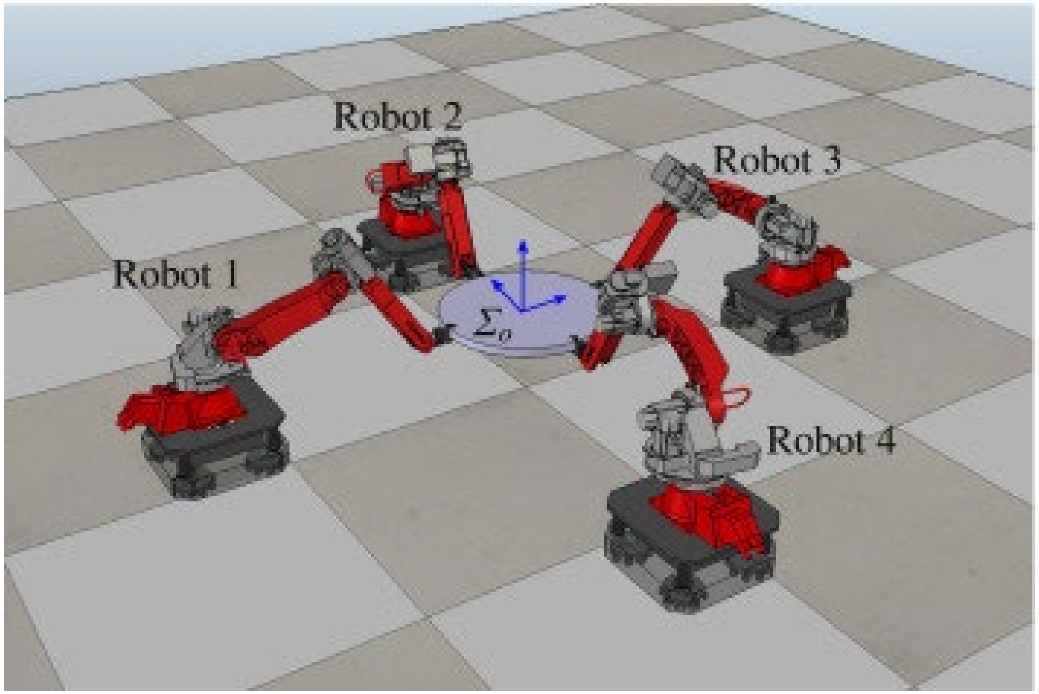
Collective patterns of social diffusion are shaped by individual inertia and trend-seeking

Mengbin Ye[1,2] ✉, Lorenzo Zino[2], Žan Mlakar[3], Jan Willem Bolderdijk[3], Hans Risselada[3], Bob M. Fennis[3] & Ming Cao[2] ✉

Experimental game setting to study *social diffusion*

- 12 participants: 8-10 students, 2-4 computer bots
- Each participant makes a binary choice {0,1} synchronized in each round, using the information of how many having chosen 0s and 1s in the previous round
- The process finishes when all participants reach the same decision or after 24 rounds
- Reward: the faster the game is finished, the more $; the more rounds to be on the winning side, the more $
- The bots: First choose 0, then choose 1

Robotic cooperative transportation task



Carry out the task repeatedly; adjust strategies each time

- each time the task is taken as a group game
- new insight into how cooperation emerge as an evolutionary outcome

Outline

- Convergence of game dynamics

- Controlling games through “incentives”
  - Uniform reward
  - Targeted reward
  - Budgeted targeted reward

- Controlling games through “targeted” agents
  - Formulation as a Markovian decision process
  - Q-learning
  - Ergodic condition

The game model

Consider an undirected network $\mathbb{G} = (\mathcal{V}, \mathcal{E})$ where the nodes $\mathcal{V} = \{1, \dots, n\}$ correspond to agents and each edge represents a 2-player game between neighboring agents.

Each agent $i \in \mathcal{V}$ chooses pure strategies from $\{A, B\}$ and receives a payoff according to the matrix:

$$\begin{array}{cc} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \end{array}, \qquad a_i, b_i, c_i, d_i \in \mathbb{R}.$$

The dynamics take place over a sequence of discrete time

Let $x_i(k)$ denote the strategy of agent i at time k, and denote the number of neighbors of agent i playing A and B at time k by $n_i^A(k)$ and $n_i^B(k)$, respectively.

The total payoffs to agent i are accumulated over all neighbors, and are equal to $a_i n_i^A(k) + b_i n_i^B(k)$ when $x_i(k) = A$

Each agent $i \in \mathcal{V}$ chooses pure strategies from $\{A, B\}$ and receives a payoff according to the matrix:

$$\begin{array}{cc} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix}, \end{array} \qquad a_i, b_i, c_i, d_i \in \mathbb{R}.$$

The dynamics take place over a sequence of discrete time

Let $x_i(k)$ denote the strategy of agent i at time k, and denote the number of neighbors of agent i playing A and B at time k by $n_i^A(k)$ and $n_i^B(k)$, respectively.

The total payoffs to agent i are accumulated over all neighbors, and are equal to $a_i n_i^A(k) + b_i n_i^B(k)$ when $x_i(k) = A$, or $c_i n_i^A(k) + d_i n_i^B(k)$ when $x_i(k) = B$.

Each agent $i \in \mathcal{V}$ chooses pure strategies from $\{A, B\}$ and receives a payoff according to the matrix:

$$\begin{array}{cc} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \end{array}, \qquad a_i, b_i, c_i, d_i \in \mathbb{R}.$$

The dynamics take place over a sequence of discrete time

Let $x_i(k)$ denote the strategy of agent i at time k, and denote the number of neighbors of agent i playing A and B at time k by $n_i^A(k)$ and $n_i^B(k)$, respectively.

Best-response dynamics

The total payoffs to agent i are accumulated over all neighbors, and are equal to $a_i n_i^A(k) + b_i n_i^B(k)$ when $x_i(k) = A$, or $c_i n_i^A(k) + d_i n_i^B(k)$ when $x_i(k) = B$.

One or more random agent at a time becomes active and chooses a single action to play against all neighbors.

The myopic **best response update rule** dictates that the active agent at time k updates at $k+1$ as follows

$$x_i(k+1) = \begin{cases} A, & \text{if } a_i n_i^A + b_i n_i^B > c_i n_i^A + d_i n_i^B \\ B, & \text{if } a_i n_i^A + b_i n_i^B < c_i n_i^A + d_i n_i^B \\ z_i, & \text{if } a_i n_i^A + b_i n_i^B = c_i n_i^A + d_i n_i^B \end{cases}.$$

where z_i is fixed to either A, B or $x_i(k)$.

Coordinating and anti-coordinating agents

Let $\delta_i := a_i - c_i + d_i - b_i$, $\gamma_i := d_i - b_i$ and $\tau_i := \frac{\gamma_i}{\delta_i}$ for $\delta_i \neq 0$.

If $\delta_i > 0$, the update rule is given by:

Threshold models

$$x_i(k+1) = \begin{cases} A & \text{if } n_i^A(k) > \tau_i \deg_i \\ B & \text{if } n_i^A(k) < \tau_i \deg_i \\ z_i & \text{if } n_i^A(k) = \tau_i \deg_i \end{cases}$$

which is the update rule of **coordinating agents**.

If $\delta_i < 0$, the update rule is given by:

$$x_i(k+1) = \begin{cases} A & \text{if } n_i^A(k) < \tau_i \deg_i \\ B & \text{if } n_i^A(k) > \tau_i \deg_i \\ z_i & \text{if } n_i^A(k) = \tau_i \deg_i \end{cases}$$

which is the update rule of **anti-coordinating agents**.

Research goal

Let $\Gamma := (\mathbb{G}, \tau, \{+, -, \pm\})$ denote a *network game*, which consists of the network $\mathbb{G}$, a vector of agent thresholds $\tau = (\tau_1, \ldots, \tau_n)^\top$, and one of $+$, $-$, or $\pm$, corresponding to the cases of all coordinating, all anti-coordinating, or a mixture of both types of agents, respectively.

To analyze the asymptotic behavior of $(\mathbb{G}, \tau, +)$, $(\mathbb{G}, \tau, -)$ and $(\mathbb{G}, \tau, \pm)$ under the best response dynamics.

When coordinator and anti-coordinators coexist

Consider a game with only two agents, one coordinator and the other anti-coordinator with both thresholds equal to 0.5

(coordinator, anti-coordinator)

$$(A, \underset{\circ}{A}) \rightarrow (\underset{\circ}{A}, B) \rightarrow (B, \underset{\circ}{B}) \rightarrow (\underset{\circ}{B}, A) \rightarrow (A, A)$$

When coordinator and anti-coordinators coexist

Consider a game with only two agents, one coordinator and the other anti-coordinator with both thresholds equal to 0.5

(coordinator, anti-coordinator)

$$(A, A) \rightarrow (A, B) \rightarrow (B, B) \rightarrow (B, A) \rightarrow (A, A)$$

This network will cycle and never reach an equilibrium!

When updates take place synchronously

Consider a game with only **two anti-coordinator** with both thresholds equal to 0.5, who update synchronously

$$(A, B) \rightarrow (A, B)$$

$$(B, A) \rightarrow (B, A)$$

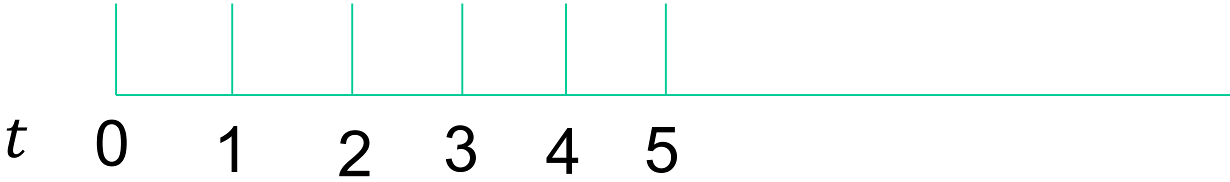
$$(A, A) \rightarrow (B, B) \rightarrow (A, A)$$

Again, this network may cycle and never reach an equilibrium!

When can we expect convergence to an equilibrium?

Asynchronous updating

Time instants:



Asynchronous clock

At each time, each agent updates its action with a positive probability which is less than 1.

- This setup makes the decision making process probabilistic

Basic convergence results

***A*-coordinating**: any agent who updates to Strategy *A* would also do so if some agents currently playing B were instead playing *A*

Theorem: For a network of A-coordinating agents, when agents update asynchronously following the best response rule, the network game dynamics converge.

Outline

- Convergence of game dynamics
- Controlling games through “incentives”
  - Uniform reward
  - Targeted reward
  - Budgeted targeted reward
- Controlling games through “targeted” agents
  - Formulation as a Markovian decision process
  - Q-learning
  - Ergodic condition

Incentive-based control of A-coordinating networks

Suppose we can offer an incentive *r* for taking a particular action.

$$\begin{array}{c c} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{pmatrix} a+r & b+r \\ c & d \end{pmatrix}, \end{array} \qquad a, b, c, d, r \in \mathbb{R}$$

How much would it cost to have all agents converge to A?

Cases:

- Uniform incentives
- Targeted incentives
- Targeted incentives subject to a budget constraint

Uniform incentive-based control

All agents receive the same incentive

$$\begin{array}{c|cc} & A & B \\ \hline A & a_i + r_0 & b_i + r_0 \\ B & c_i & d_i \end{array}, \qquad a_i, b_i, c_i, d_i \in \mathbb{R}$$

Find the minimum value of the uniform incentive such that the entire network converges to *A*?

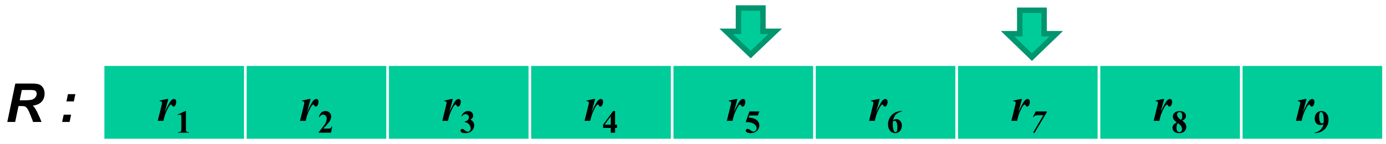
- ***A*-coordinating**: any agent who updates to Strategy *A* would also do so if some agents currently playing B were instead playing *A*
- ***A*-monotone**: Offering incentives to play *A* will never lead to an agent to switch away from *A*
- **Uniquely-convergent**: Offering incentives leads to a unique equilibrium

Theorem: Every network of *A*-coordinating agents is *A*-monotone and uniquely convergent.

Uniform incentive-based control

Proposition:
One can construct a finite set ***R*** that contains *r**

Because of the *A*-monotone property, one can carry out the binary search:



$\boldsymbol{R}$: r_1 r_2 r_3 r_4 r_5 r_6 r_7 r_8 r_9

Theorem:
Within finite steps, binary search solves the uniform reward problem

Targeted incentive-based control

Suppose it’s possible to offer different rewards to individual agents:

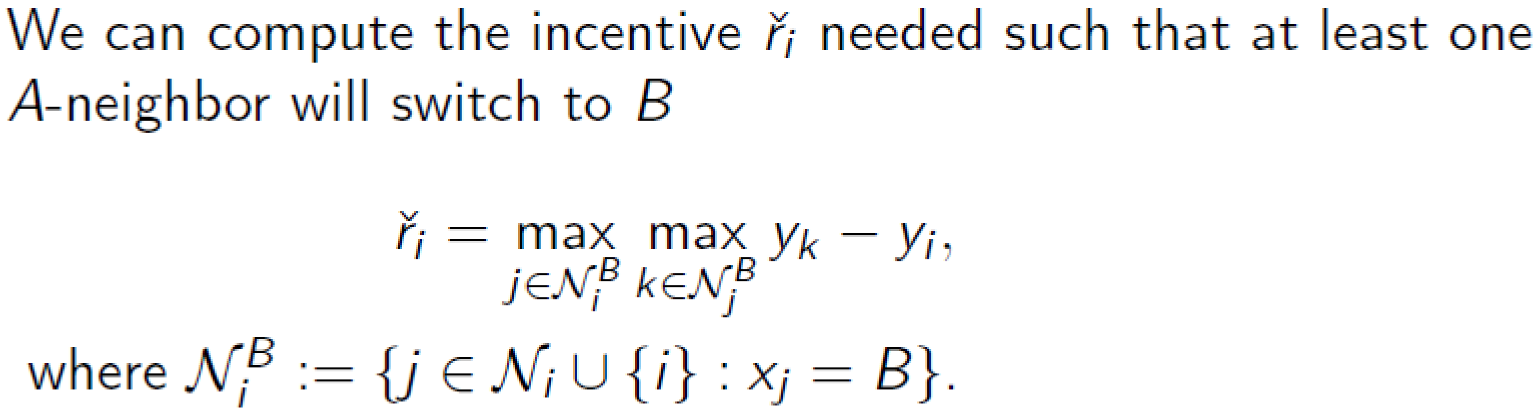
$$\begin{array}{ccc} & A & B \\ A & a_i + r_i & b_i + r_i \\ B & c_i & d_i \end{array}, \quad a_i, b_i, c_i, d_i \in \mathbb{R},\ r_i \in \mathbb{R}_{\geq 0}$$

Problem 1: Find $\mathbf{r} = (r_1, \ldots, r_n)$ that minimizes $\sum_i \mathbf{r}$ such that the entire network converges to A.

Problem 2 (budget constraint): Find $\mathbf{r}$ that maximizes the number of agents who converge to A subject to $\sum_i \mathbf{r} \leq \rho$.

Targeted incentive-based control

Computationally complex to solve exactly (conjectured to be NP)



We can compute the incentive $\check{r}_i$ needed such that at least one A-neighbor will switch to B

$$\check{r}_i = \max_{j \in \mathcal{N}_i^B} \max_{k \in \mathcal{N}_j^B} y_k - y_i,$$

where $\mathcal{N}_i^B := \{j \in \mathcal{N}_i \cup \{i\} : x_j = B\}$.

Algorithm: Iteratively choose agents to switch until the desired equilibrium is reached or the budget limit is exceeded.

Targeted incentive-based control

How should we choose these agents?

Several possibilities: max degree, min required incentives, etc.

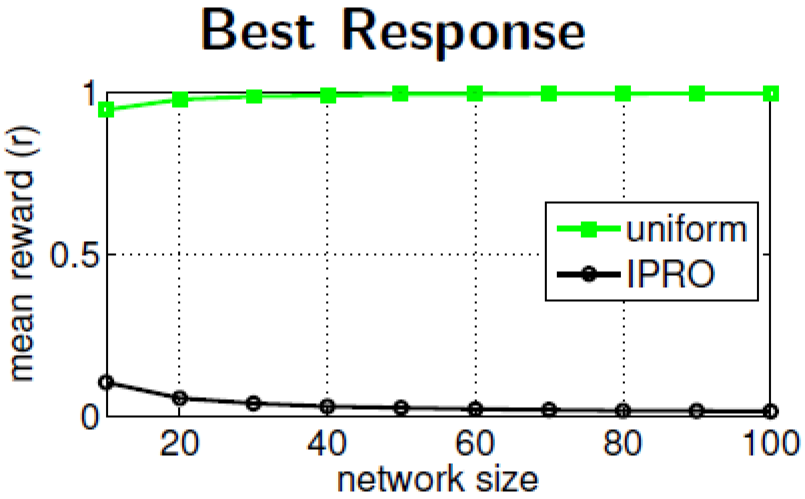
Approach: Iteratively maximize a benefit-to-cost ratio

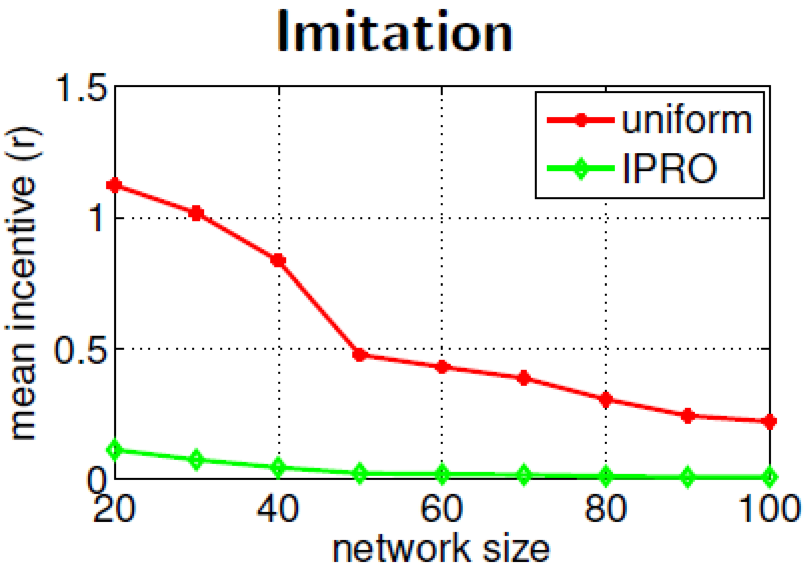
Benefit = # of agents who switch to *A*, cost = incentive

$$\max_i \frac{\Delta\Phi(x)^\alpha}{\check{r}_i^\beta}, \quad \text{where } \Delta\Phi(x) = \Phi(x(t_2)) - \Phi(x(t_1)),$$

$$\Phi(x) = \sum_{i=1}^{n} n_i^A(x), \qquad \alpha \text{ and } \beta \text{ are design parameters.}$$

Simulation results: Uniform vs. Targeted incentives





Imitation dynamics

Agents adopt the strategy of their highest-earning neighbors

$\mathcal{S}_i^M(t)$ = set of strategies earning the maximum payoff in the neighborhood of agent i:

$$\mathcal{S}_i^M(t) = \left\{ x_j(t) \,\middle|\, y_j(t) = \max_{k \in \mathcal{N}_i \cup \{i\}} y_k(t) \right\}.$$

Imitation update rule:

$$x_i(t+1) = \begin{cases} A & \mathcal{S}_i^M(t) = \{A\} \\ B & \mathcal{S}_i^M(t) = \{B\} \\ x_i(t) & \mathcal{S}_i^M(t) = \{A, B\} \end{cases}$$

Stochastic games: The environment changes ***stochastically***

The possible environmental states

$$\mathcal{S} \triangleq \{s^1, s^2, \ldots, s^M\}$$

Assumption: the dynamics of the the enviorenmental state form an irreducible and aperiodic Markov chain

The payoff table (or function) of the game

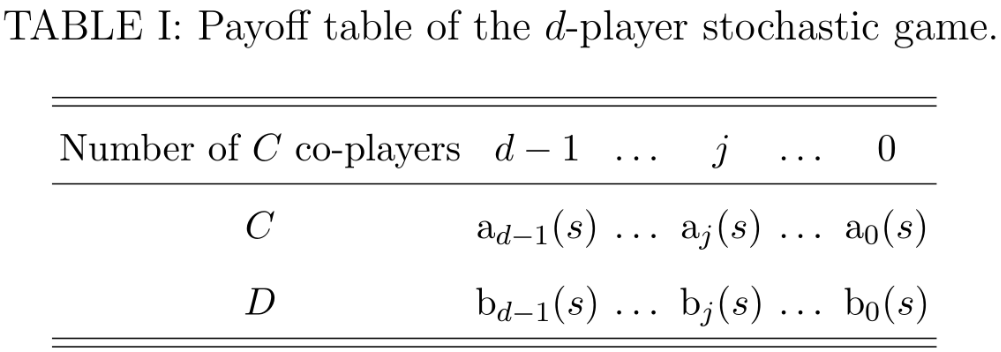


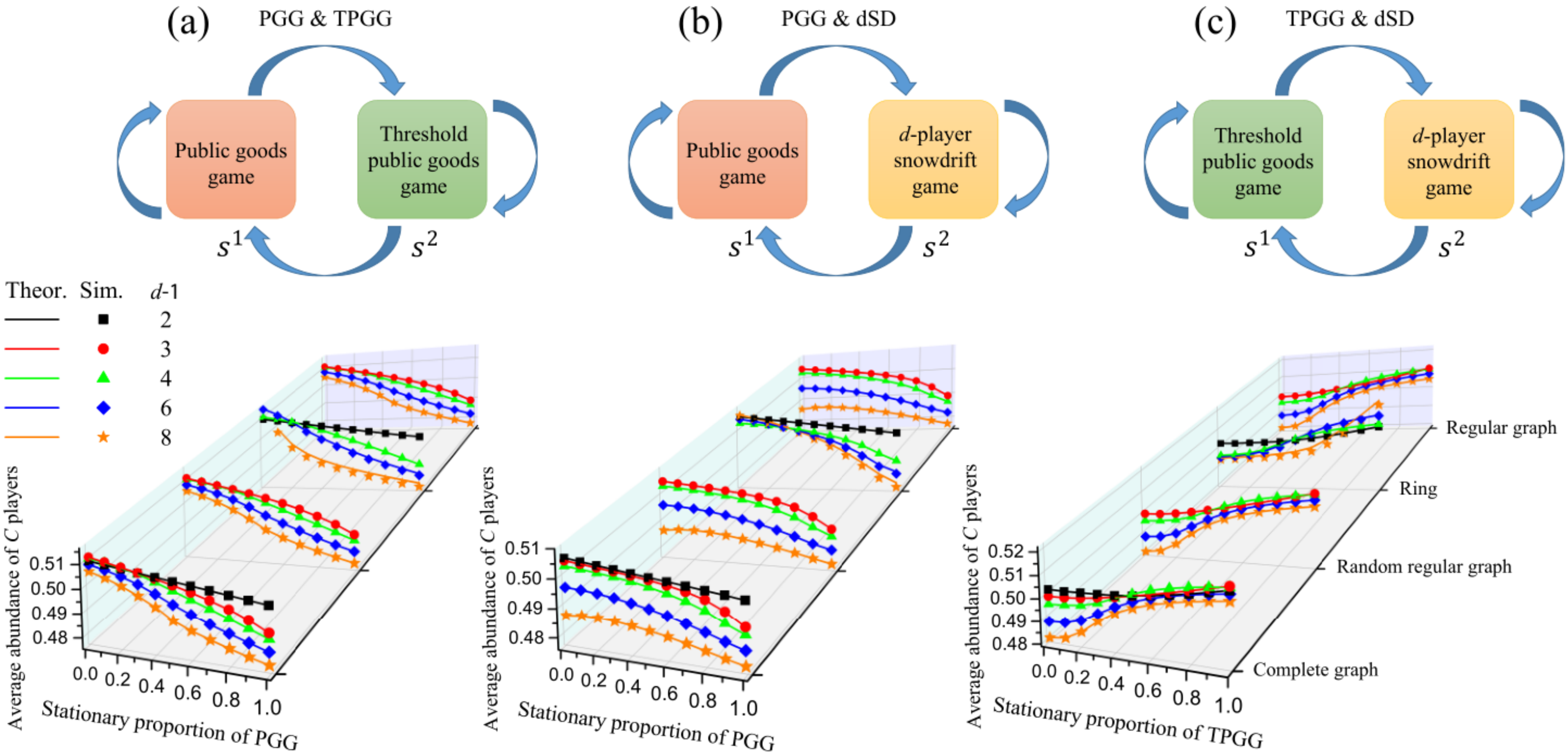
TABLE I: Payoff table of the d-player stochastic game.

Number of C co-players	$d-1$	$\ldots$	j	$\ldots$	0
C	$\mathrm{a}_{d-1}(s)$	$\ldots$	$\mathrm{a}_j(s)$	$\ldots$	$\mathrm{a}_0(s)$
D	$\mathrm{b}_{d-1}(s)$	$\ldots$	$\mathrm{b}_j(s)$	$\ldots$	$\mathrm{b}_0(s)$

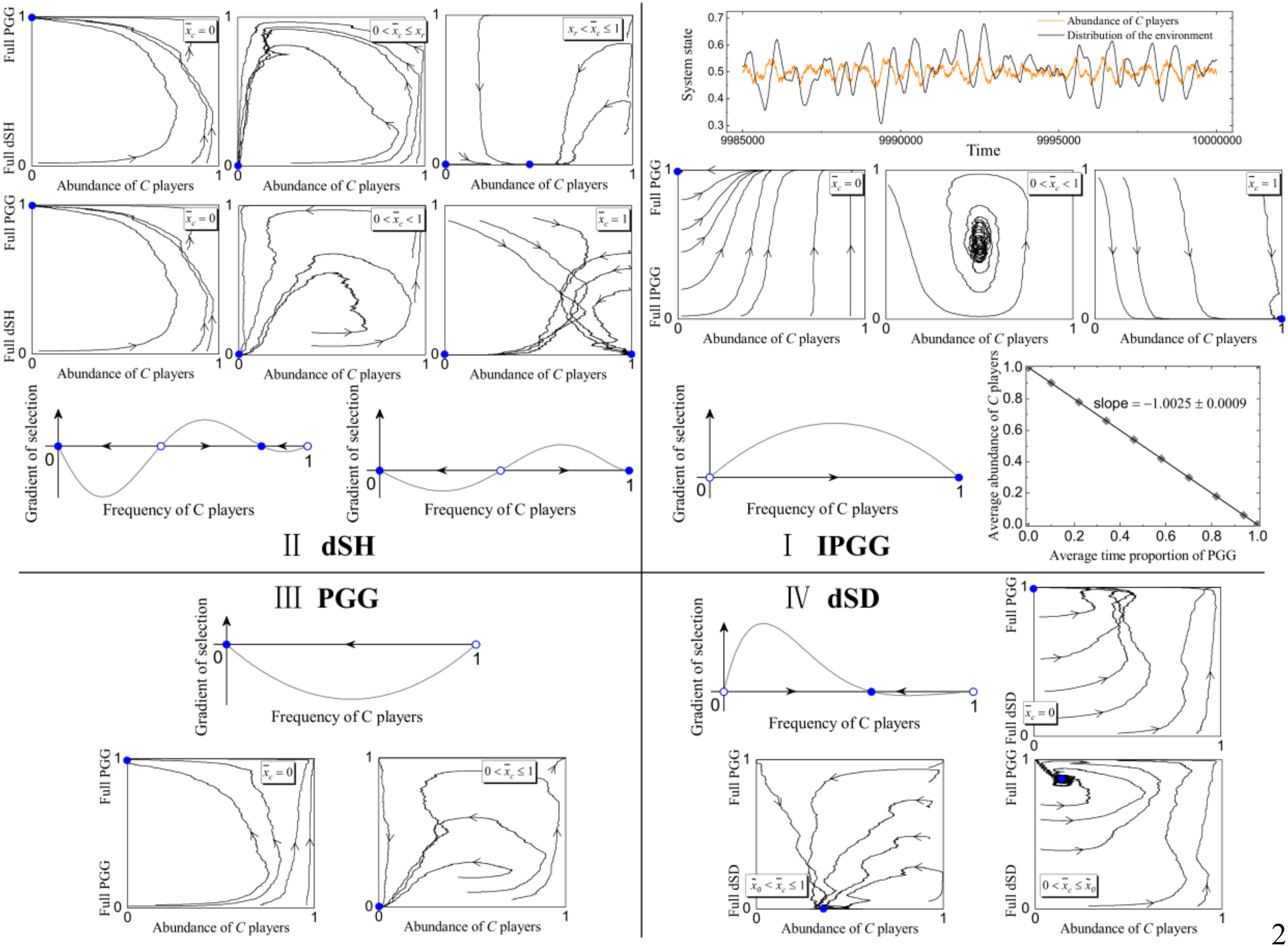
where $\mathrm{a}_j(s)$ and $\mathrm{b}_j(s) \in \mathbb{R},\ \forall s \in \mathcal{S}, j \in \mathcal{J}, a \in \mathcal{A}$

Decision-making: Strategic updating is asynchronous and guided by a learning process

◆ Condition for the prevalence of cooperating agents



Dominance, bistability, and coexistence obtained from replicator dynamics are recovered.



Outline

- Convergence of game dynamics

- Controlling games through “incentives”
  - Uniform reward
  - Targeted reward
  - Budgeted targeted reward

- Controlling games through “targeted” agents
  - Formulation as a Markovian decision process
  - Q-learning
  - Ergodic condition

Markov chain representation

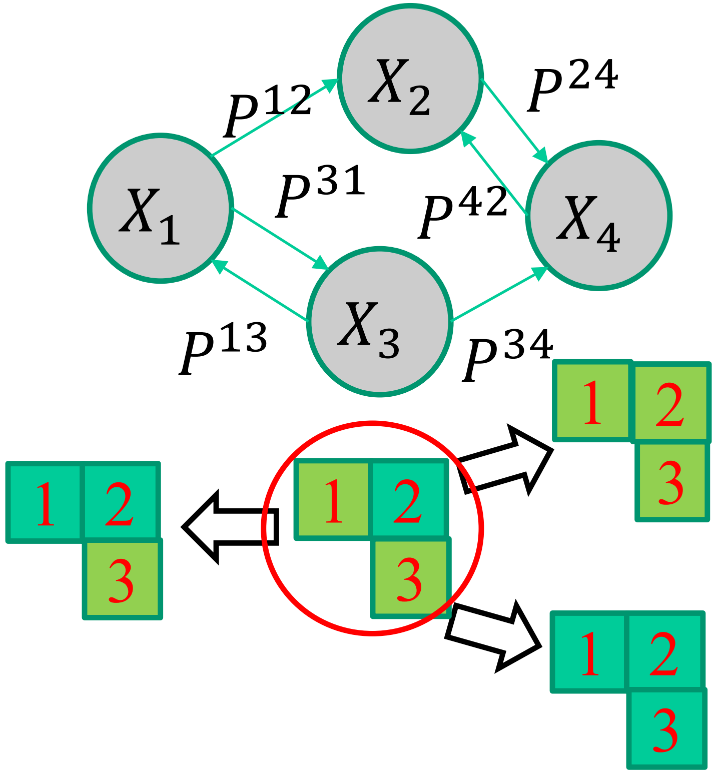
$$X(t) \xrightarrow{\text{asynchronous best response}} X(t+1)$$

Markov chain

$$X = (x^1, x^2, \cdots, x^N)$$

- **Markov chain**

  Best response decision making rule guarantees that the current (joint) action depends only on the previous action



Control action

Control Method

Targeted agent: allowed to be controlled and adopt a given action.
Ordinary agent: follow the asynchronous best response rule

Control Objective

- Avoid *"price of anarchy" "social dilemma": a*gents' selfish decision making often leads to undesired outcomes.

- Some action is preferred.
  Maximize the sum of agents' payoffs

Formulation as a Markovian Decision Process (I)

State

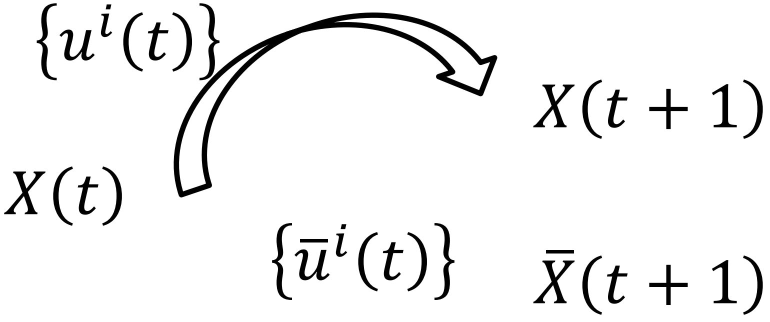
Joint action of all the players X

Control Action

The joint action of all the targeted players u

Utility function

Equip a corresponding utility function for each state $r(X)$



Formulation as a Markovian Decision Process (II)

Policy

a stationary deterministic policy π: $\mathcal{X} \rightarrow \mathcal{U}$

Objective function

expected discounted objective function

$$J_\pi(X) \coloneqq E_{\pi, X(0)=X} \sum_{t=0}^{\infty} \beta^t \, r\big(X(t), u(t)\big)$$

State value

V^*: $\mathcal{X} \rightarrow \mathcal{R}$, which calculates the quality of each state,

$$V^*(X) \coloneqq max_{\pi \in \Pi} \, J_\pi(X)$$

Q value

Q^*: $\mathcal{X} \times \mathcal{U} \rightarrow \mathcal{R}$, which calculates for each state-action pair,

$$Q^*(X, u) \coloneqq r(X, u) + \beta \textstyle\sum_{Y \in \mathcal{X}} Pr(Y|X, u) \, V^*(Y)$$

Optimal policy

$u^*(X)$=$argmx_{u \in \mathcal{U}} \, Q^*(X, u)$

Q-learning and ergodic condition

Q-learning

a recursive learning algorithm for computing the Q-values

$Q_{k+1}(X,u) = Q_k(X,u)$, if $(X,u) \neq \big(X(t),u(t)\big)$

$$Q_{k+1}(X,u) = Q_k(X,u) + \boxed{\gamma_k}\{\boxed{[r(X,u) + \beta max_{v\in\mathcal{U}} Q_k(\bar{X},v)]} - \boxed{Q_k(X,u)}\}$$

if $(X,u) = \big(X(t),u(t)\big)$

Learning rate

estimate

Old value

Ergodicity condition

every state-action pair (X,u) occurs infinitely often

Ergodicity condition is necessary for Q-learning to converge to the optimal Q-value

Ergodic condition

- **Condition 1**: One state is always reachable from another state
- **Condition 2**: All-*A* and All-*B* are mutually reachable

Condition 1 and condition 2 are equivalent for unbiased population and uniformly biased population.

Ergodicity condition

every state-action pair (X, u) occurs infinitely often

Ergodicity condition is necessary for Q-learning to converge to the optimal Q-value

Conclusions and outlook

Conclusions:

- Under the coordination game model, networks of all coordinating or all anti-coordinating agents will almost surely reach an equilibrium in finite time.
- Incentive-driven control helps to drive coordinating networks towards desired equilibria
- Targeted control can be realized through Q-learning

Outlook:

- More analytical results for stochastic games
- Efficient learning algorithms design

Some selected publications from my group on related topics

“Optimal control of robust team stochastic games,” F. Huang, M. Cao and L. Wang, under review, 2023

“Learning enables adaptation in cooperatition for multi-player stochastic games,” F. Huang, M. Cao and L. Wang. *Journal of the Royal Society Interface*, 2020

“Control using Q-learning for networked coordination games,” J. Bo and M. Cao. *Asian Control Conference*, 2022, journal version under review, 2023

“Controlling networks of imitative agents,” P. Ramazi, J. Riehl and M. Cao. *IEEE Transactions on Network Science and Engineering*, 2023.

“The lower convergence tendency of imitators compared to best responders,” P. Ramazi, J. Riehl and M. Cao. *Automatica*, 2022.

“Asynchronous decision-making dynamics under best-response update rule in finite heterogeneous populations,” P. Ramazi and M. Cao. *IEEE Transactions on Automatic Control*, 63(3), 742-751, 2018.

“Networks of conforming or nonconforming individuals tend to reach satisfactory decisions,” P. Ramazi, J. R. Riehl, and M. Cao. *Proceedings of the National Academy of Science of USA (PNAS)*, 113(46), pp12985-12990, 2016

“A survey on the analysis and control of evolutionary matrix games,” J. R. Riehl, P. Ramazi and M. Cao. *Annual Reviews in Control*, 45(6), 87-106, 2018

http://www.rug.nl/staff/m.cao m.cao@rug.nl