# Dynamics of $Q$-Learning in Networked Stochastic Games

Zheng Yuan, Guangchen Jiang, Shuyue Hu, Matjaž Perc, Chen Chu, and Jinzhuo Liu

***Abstract*—Stochastic games form the foundational mathematical framework for describing multiagent interactions and underpin the theoretical foundations of multiagent reinforcement learning (MARL) and optimal decision making. However, previous research has typically focused on either two-agent settings or large-scale well-mixed agent populations, where the considered interaction scenarios were far from realistic. In this article, we consider structured populations where agents can interact with immediate neighbors. By using the pair-approximation method, we develop a new dynamical model to describe the $Q$-learning dynamics in stochastic games on regular graphs. Through comparisons with agent-based simulation results, we validate the accuracy of our dynamical model across various stochastic games, population structures, and algorithm parameters. Our research thus provides both qualitative and quantitative insights into the effects of state transition rules and graph topologies in population dynamics. In particular, we show that, under certain conditions, state transitions can significantly promote the evolution of cooperation in social dilemmas. We also explored the effects of agent degree on cooperation, and unlike previous findings, we show that this can have either positive or negative implications for cooperation depending on the transition rules.**

***Index Terms*—Cooperation, $Q$-learning dynamics, regular graphs, stochastic game.**

Received 26 April 2025; revised 10 November 2025; accepted 27 November 2025. Date of publication 1 January 2026; date of current version 3 June 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62366058 and Grant 6250076060, in part by Yunnan Province XingDian Talent Support Program under Grant YNWR-QNBJ-2020-041, and in part by the Young and Middle-Aged Academic and Technical Leaders Reserve Talent Project of Yunnan Province under Grant 202205AC160034. The work of Matjaž Perc was supported by the Slovenian Research and Innovation Agency (Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije) under Grant P1-0403. *(Corresponding authors: Matjaž Perc; Jinzhuo Liu.)*

Zheng Yuan and Guangchen Jiang are with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: yuanzheng@mail.nwpu.edu.cn; guangchen98.jiang@gmail.com).

Shuyue Hu is with the Department of Basic Theoretical Research, Shanghai Artificial Intelligence Laboratory, Shanghai 201206, China (e-mail: hushuyue@pjlab.org.cn).

Matjaž Perc is with the Faculty of Natural Sciences and Mathematics, University of Maribor, 2000 Maribor, Slovenia, also with the Community Healthcare Center Dr. Adolf Drolc Maribor, 2000 Maribor, Slovenia, also with the Department of Physics, Kyung Hee University, Seoul 02447, Republic of Korea, also with the Complexity Science Hub, 1030 Vienna, Austria, and also with the University College, Korea University, Seoul 02841, Republic of Korea (e-mail: matjaz.perc@gmail.com).

Chen Chu is with the School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China (e-mail: chuchenynufe@hotmail.com).

Jinzhuo Liu is with the School of Software and AI, Yunnan University, Kunming 650504, China, and also with Yunnan Key Laboratory of Software Engineering, Kunming 650504, China (e-mail: jinzhuo.liu@hotmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3641365, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3641365

## I. Introduction

RECENT advances in multiagent reinforcement learning (MARL) have significantly improved the decision-making abilities of agents, driving progress in robotics [1], autonomous systems [2], resource allocation mechanisms [3], social dynamics research [4], [5], [6], [7], and intelligent game agents [8], [9], [10]. Despite these achievements, the theoretical foundations of MARL remain an open question [11].

Understanding the theoretical aspects of multiagent learning has long been a key focus, particularly through the lens of learning dynamics in games [12], [13], [14], [15], [16], [17]. Among various learning algorithms, $Q$-learning is a fundamental method that iteratively estimates the state–action value function for optimal decision making [18]. Together with its variants [19], it has been applied across diverse domains and has had a profound impact [20], [21]. This success has naturally motivated extensive studies on the dynamics of multiagent $Q$-learning. Tuyls et al. [22] introduce the *selection–mutation model*, which formalizes the evolutionary dynamics of $Q$-learning with Boltzmann exploration in two-agent normal-form games. This model establishes a notable connection between MARL and evolutionary game theory (EGT), offering new perspectives for algorithm design in MARL [23]. Expanding beyond two-agent scenarios, Hu et al. [24] demonstrate that, in an infinitely well-mixed population where agents are randomly paired to play two-agent normal-form games, the $Q$-learning dynamics can be described by a Fokker–Planck equation. This has inspired further research on learning dynamics in large-scale multiagent systems [25], [26]. However, these foundational works are limited to static normal-form games with fixed payoff matrices, whereas real-world interactions often occur in dynamic environments [27], [28]. Most MARL algorithms today account for the evolving interplay between agent strategies and their environments [29], [30], [31].

More recently, research has focused on modelling learning dynamics in *stochastic games* (Definition 1), which better reflect real-world interactions by incorporating state transitions that capture how humans both affect and are affected by their environment [32]. Stochastic games provide a powerful framework for investigating adaptive agent behaviors under environmental variability [33], [34]. Hennes et al. [35] introduce state-coupled replicator dynamics to model replicator dynamics in stochastic games, while Chu et al. [36] develop a dynamical model for $Q$-learning in stochastic games under an infinite-agent setting.

Despite these advances in learning dynamics in stochastic games, relatively little attention has been given to structured populations, where graphs offer a more realistic representation of local interactions [37]. To address this gap, we present a new formal model that captures the dynamics of $Q$-learning in stochastic games within large-scale structured populations, that is, networked stochastic games. The *population structure* (Definition 2) is represented as a regular graph, where each agent occupies a node and edges denote pairwise interactions. Each edge corresponds to a two-agent stochastic game with symmetric state transitions. In the networked stochastic games, the learning process unfolds iteratively over discrete time steps. At each time step, every agent selects an action to interact with all immediate neighbors and receives an average reward from all pairwise interactions, where the number of neighbors is determined by the degree of the regular graph. Following action selection, the state transitions in each stochastic game are determined by the joint actions, the current state, and chance. Each agent then updates its strategy using stateless $Q$-learning with Boltzmann exploration, based on the average reward obtained from all interactions.

In contrast to the idealized all-to-all interactions, the vast number of possible local configurations in structured populations significantly complicates their mathematical analysis [38]. Especially in networked stochastic games, an agent's reward is determined not only by its neighbors' actions but also by the states of the stochastic games, rendering the local configurations more complex than those in normal-form games. Moreover, the interplay between agent behaviors and their environment induces population heterogeneity—agents exhibit distinct state distributions and opponent strategies—which critically influences learning dynamics. Modelling the learning dynamics in such settings thus presents the following challenge: *how can we formally describe the local interactions encountered by agents and accurately capture the resulting population heterogeneity?*

To address this challenge, we adopt the novel *pair-approximation method* proposed by Chu et al. [36]. Specifically, we focus on the evolution of a *pair* (Definition 3), defined as the combination of the $Q$-values of an agent pair along with their state. The population state is then comprehensively represented by the probability distribution of such pairs. To formally capture the complex local interactions in networked stochastic games, we introduce the concept of an *interaction configuration* (Definition 4), which specifies the number of pairwise interactions of each type in an agent's neighborhood. Here, the type of an interaction is defined by the neighbor's action and the current game state. Based on the interaction configuration, the immediate payoff received by an agent within a pair can be derived (Lemma 1), which subsequently drives the evolution of its $Q$-values (Lemma 2).

Importantly, the pair-approximation method allows us to capture population heterogeneity that the widely adopted mean-field approximation cannot directly describe [24], [39], [40]. By leveraging the probability distribution of pairs, we can infer agents' heterogeneous state distributions and opponent strategies from their $Q$-values (Lemma 3). In this way, we partition a heterogeneous population into groups of agents with identical $Q$-values. For agents with the same $Q$-values, the influence of each pairwise interaction the agent participates in is approximated by an average effect, which is represented by the average probability of an agent engaging in a given type of pairwise interaction (Lemma 4). Using this average effect, the probability distributions of agents' interaction configurations can be derived (Lemma 5). Finally, we establish a differential equation to track the evolution of environmental states (Lemma 6) and a partial differential equation to characterize the temporal evolution of pair distributions (Theorem 1), offering deeper insights into the dynamics of both agent behavior and environmental state over time.

We validate our model by comparing its predictions with agent-based simulations across various game configurations, state transition rules, network topologies, and learning algorithm parameters. Our model consistently provides an accurate description of multiagent $Q$-learning dynamics. Notably, in a two-state donation game, we demonstrate that environmental stochasticity markedly influences both the evolution of environmental states and the agents' strategy learning. In a stochastic game featuring transitions between stag hunt (SH) and prisoner's dilemma (PD) games, we observe that under a state-independent transition rule, the cooperation rate decreases as the degree of the regular graph increases. In contrast, with a state-dependent transition rule, agents on higher degree regular graphs exhibit more cooperative behavior. Furthermore, in a two-state SH game, we show that state transitions can yield more efficient outcomes compared to repeated normal-form games. These findings complement previous research on stochastic games [33], [41], offering novel insights into resolving real-world social dilemmas and advancing the development of prosocial agents in multiagent systems.

In summary, the main contributions of this article are.

1) We develop the first formal model that describes multiagent $Q$-learning dynamics in stochastic games on graphs, introducing the concept of interaction configuration to formally characterize the features of local interactions.
2) We extend the pair-approximation method to model $Q$-learning dynamics in networked games. This approach effectively captures the strong coupling between agent behaviors and environmental states in stochastic games, and characterizes population heterogeneity.
3) Through extensive experiments, we validate the descriptive power of our dynamical model and reveal that the influence of network topology on cooperative behavior varies depending on the state transition rules.

The remainder of the article is organized as follows. In Section II, we review related work. Section III outlines the key preliminaries. In Section IV, we formally model the dynamics of multiagent $Q$-learning in networked stochastic games. Finally, Section V presents experimental results under various settings to validate our model and explore behavior evolution in networked stochastic games.

## II. Related Work

Numerous studies highlight the importance of modelling reinforcement learning dynamics in multiagent systems, significantly advancing the field. Börgers and Sarin [12] establish a formal link between the behavior of cross learning [42], a fundamental MARL algorithm, and the population dynamics described by replicator dynamics in EGT. Tuyls et al. [22] and Sato and Crutchfield [43] extend this connection to $Q$-learning dynamics, represented through a selection–mutation model. Rodrigues Gomes and Kowalczyk [44] and Wunder et al. [45] further explore $Q$-learning dynamics with $\varepsilon$-greedy exploration in two-agent games. Panozzo et al. [46] propose an extended selection–mutation model for $Q$-learning in extensive-form games, while Deng et al. [47] present state-transition replicator dynamics, deriving ordinary differential equations to capture learning automata dynamics in stochastic games. Building on these dynamics, new algorithms such as FAQ-learning [48] and RESQ-learning [49] have been developed, enhancing our understanding of complex reinforcement learning dynamics in two-agent settings.

Expanding to large-scale agent populations, Hu et al. [24] utilize mean-field theory to develop a Fokker–Planck equation-based model characterizing $Q$-learning dynamics in repeated games. Leung et al. [50] generalize this approach to scenarios where agents have local and incomplete information, while Chu et al. [39] introduce a variant for $Q$-learning dynamics on regular graphs. Beyond regular graphs, Liu et al. [51] propose a framework to formally characterize $Q$-learning dynamics in repeated normal-form games across various nonregular graph topologies, such as random and scale-free graphs. Shi et al. [52] further investigate $Q$-learning dynamics in public goods games on uniformly random hypergraphs, where agents interact within larger groups rather than pairwise. Extending to stochastic games, Chu et al. [36] propose a novel application of the pair-approximation method for modelling $Q$-learning dynamics in well-mixed populations. Apart from $Q$-learning, regret minimization has also received attention. Wang et al. [53] offer an insightful and theoretically grounded characterization of the regret dynamics of individual agents in repeated normal-form games using the master equation approach. Besides, building on these studies, convergence guarantees and equilibrium analysis have attracted increasing interest. Hu et al. [54] provide new convergence results for smooth fictitious play in both network games and network population games. Hu et al. [55] model the real-time regret dynamics in population network games using the continuity equation, providing convergence guarantees for regret minimization in weighted zero-sum and weighted potential population network games.

Therefore, along this line of research on learning dynamics in games, $Q$-learning dynamics in networked stochastic games remain poorly understood. Our work on this scenario not only addresses this gap but can also inform future research on establishing convergence of learning under stochastic games.

## III. Preliminaries

In this section, we introduce the foundational concepts and details related to our multiagent learning framework.

### A. Multiagent Learning Framework in Networked Stochastic Games

Stochastic games effectively capture interactive scenarios in which individual actions wield the power to influence the dynamics of future interactions. Such a dynamic environment, in turn, plays a crucial role in shaping individuals' decision-making processes.

*Definition 1 (Stochastic Game):* A stochastic game with $n_g$ agents and $d$ states is typically defined as a tuple $< \mathcal{N}_g, \mathcal{S}, \mathcal{A}, z, r, \mathbf{x}^1, \ldots, \mathbf{x}^{n_g} >$. In every state $s \in \mathcal{S} := \{s_1, \ldots, s_d\}$, each agent $i \in \mathcal{N}_g := \{1, \ldots, n_g\}$ can select its action $a^i$ from its action set $\mathcal{A}^i(s)$ according to its $\mathbf{x}^i(s)$. The payoff function $r(s, \mathbf{a}) : \prod_{i=1}^{n_g} \mathcal{A}^i(s) \to \mathbb{R}^n$ maps the joint action $\mathbf{a} := [a^1, \ldots, a^{n_g}]^\top$ to an immediate reward for each agent. The transition function $z(s, \mathbf{a}) : \prod_{i=1}^{n_g} \mathcal{A}^i(s) \to \Delta^{d-1}$ determines the change in environmental state under current state $s$ and joint action $\mathbf{a}$, where $\Delta^{d-1}$ is the $(d-1)$-simplex, and $z_{s'}(s, \mathbf{a})$ is the transition probability from state $s$ to $s'$ under joint action $\mathbf{a}$.

For better clarity, we consider the simplest two-agent stochastic game as an illustrative example, which involves two states and two available actions in each state to demonstrate the essential elements of a stochastic game. Each state corresponds to a symmetric normal-form game characterized by a specific payoff matrix. Consequently, a state transition can be interpreted as a change in the type of normal-form game. Due to the fact that we consider the normal-form game to be symmetric, the two agents have the same action sets in a certain state, that is, $\mathcal{A}^1(s) = \mathcal{A}^2(s)$. Then, we can drop the agent index and use $\mathcal{A}(s)$ to represent the set of available actions for all agents in state $s$. Here, we have $\mathcal{A}(s) := \{a_1(s), a_2(s)\}$, where $a_i(s), i \in \{1, 2\}$, is the $i$th action in state $s$. $\mathbf{M}_s$ is the payoff matrix for agents in an arbitrary state $s$ and can be represented as follows:

$$\mathbf{M}_s := \begin{bmatrix} r_{a_1(s)a_1(s)} & r_{a_1(s)a_2(s)} \\ r_{a_2(s)a_1(s)} & r_{a_2(s)a_2(s)} \end{bmatrix}$$

where $r_{a_i(s)a_j(s)}, i, j \in \{1, 2\}$, is the reward that an agent can receive when it takes the $i$th action against its opponent taking the $j$th action in state $s$. Besides, the state transition matrices are given as follows:

$$\mathbf{T}_{s\to s'} := \begin{bmatrix} z_{s'}(s, a_1(s), a_1(s)) & z_{s'}(s, a_1(s), a_2(s)) \\ z_{s'}(s, a_2(s), a_1(s)) & z_{s'}(s, a_2(s), a_2(s)) \end{bmatrix}$$
$$\mathbf{T}_{s\to s} = \mathbf{I}_2 - \mathbf{T}_{s\to s'}$$

where $\mathbf{T}_{s\to s'}$ represents the transition probabilities from an arbitrary state $s$ to another different state $s'$ under different joint actions, and $\mathbf{T}_{s\to s}$ represents the probabilities that the state $s$ will not change given different joint actions. For $i, j \in \{1, 2\}$, $z_{s'}(s, a_i(s), a_j(s))$ denotes the transition probability from state $s$ to $s'$ under the joint action $\mathbf{a} = [a_i(s), a_j(s)]^\top$. $\mathbf{I}_2$ denotes a second-order all-ones matrix. Note that symmetric state transitions mean that the impact of agent behaviors on environmental changes is based solely on their actions, irrespective of their individual identities. Formally, the transition probabilities should satisfy the condition $z_{s'}(s, a_i(s), a_j(s)) = z_{s'}(s, a_j(s), a_i(s))$.

Building upon a specific stochastic game, this article focuses on population-level networked stochastic games. Specifically, we investigate the learning dynamics in a structured population, where each agent plays a separate stochastic game (as defined in Definition 1, with $n_g = 2$ and symmetric state transitions) against each of its neighbors. This pairwise-interaction setup is typical in large-scale population games [24], [36], [39], where the entire population can be viewed as comprising multiple two-agent stochastic games. The neighbors are determined by the population structure, which restricts the range of interactions. This local interaction model mirrors real-world social dynamics, where individuals typically maintain connections within a limited social circle due to cognitive constraints and practical limitations [56], [57]. For instance, in social networks, users predominantly interact with a select group of others, influenced by factors such as geographical proximity, shared interests, and the cognitive cost of maintaining relationships, rather than engaging with the entire user population.

*Definition 2 (Population Structure):* The interactive structure of the population is described by an undirected connected graph $\mathcal{G} := <\mathcal{N}, \mathcal{E}>$ without any loops, where $\mathcal{N} := \{1, \ldots, n\}$ is the set of nodes, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the edge set. Each agent occupies a node, and each edge denoted by $e_{ij} = (i, j), i \neq j$ and $i, j \in \mathcal{N}$ means that agent $i$ is connected to $j$. For the undirected connected graph, $e_{ij}$ and $e_{ji}$ refer to the same edge which indicates $i$ and $j$ are two connected agents (neighbors). Only the two connected agents can interact with each other along the edge.

According to the above definition, we can represent the interactive relationships between agents more explicitly by using an adjacency matrix $\mathbf{B} := [b_{ij}]_{n\times n}$. Based on whether there is an edge between two agents, all diagonal elements $b_{ii}$ of matrix $\mathbf{B}$ equal zero, off-diagonal elements satisfy $b_{ij} = 1(i \neq j)$ if and only if $e_{ij} \in \mathcal{E}$, otherwise $b_{ij} = 0$. Besides, we define the neighbor set of agent $i$ as $\mathcal{N}_i := \{j : e_{ij} \in \mathcal{E}\}$.

In this article, we consider agent populations represented by $k$-regular graphs, where each node has exactly $k$ neighbors. Regular graphs provide a canonical framework for studying population dynamics and the evolution of cooperation, bridging analyses from well-mixed to structured populations [37], [39], [58]. Although heterogeneous networks better capture real-world diversity, their analytical complexity often limits theoretical exploration. As highlighted by Liu et al. [51], incorporating more topological attributes is challenging but essential for modelling dynamics on nonregular graphs. Therefore, most theoretical studies have been conducted on regular graphs [59], [60], [61], [62], [63], [64]. Similarly, in this article, the simple topology of regular graphs enables us to investigate, in a tractable manner, how network structure shapes population dynamics under stochastic games, thereby providing a fundamental theoretical understanding of such systems.

Then, we propose a concurrent learning framework for networked stochastic games, where each agent simultaneously interacts with all $k$ neighbors through pairwise stochastic games. At each time step $t$, every agent selects a single action according to its strategy and applies it uniformly across all interactions. Agent behaviors in these interactions result in both their immediate rewards and state transitions. The immediate reward for each agent is the average of its $k$ interactions, while state transitions follow a uniform rule governed by transition matrices, which map joint actions and current states to future states. Finally, each agent updates its strategy according to the specified learning algorithm.

Note that the assumption that each agent takes a single action across all interactions is common in studies of $Q$-learning dynamics and graphical polymatrix games [13]. It is also widely accepted in research on the evolution of cooperation (e.g., [65], [66], [67], [68]). Although some studies allow agents to choose different actions toward different neighbors (e.g., [69], [70]), this line of research remains limited. Therefore, adopting this assumption ensures that our model remains comparable with a large body of existing work while retaining empirical plausibility—for instance, a social media user often broadcasts identical content to audiences with different social ties, or a self-driving car executes a single maneuver at an intersection when encountering multiple other vehicles.

### B. Q-Learning Algorithm for State-Agnostic Agents

We assume that all agents update their strategies using the well-known $Q$-learning algorithm [18]. Conventionally, $Q$-learning agents condition their actions on the perceived state. For example, under the scenario where an agent takes the same action for all pairwise stochastic games with its neighbors, the standard $Q$-learning algorithm allows the agent to make decisions based on the perceived state of its local neighborhood, which is determined by the game states of its pairwise interactions. This naturally raises further questions of how to represent the perceived state and how this representation, as an input, influences the learning dynamics.

However, to reveal the effects of state transitions and graph topologies on learning dynamics, while the state-aware case involves additional challenges in representing the perceived state and its transition process, a simplified state-agnostic setting where agents make decisions without relying on perceived-state information can provide meaningful insights. Therefore, many classical studies have adopted such a state-agnostic setting (e.g., [41], [61]). Moreover, this setting is also realistic, as in many applications, knowledge about environmental states is at best incomplete. For example, Abou Chakra et al. [71] consider unpredictable environments, and there are real-world situations where agents must make rapid decisions without additional state information, while physical limitations and the high cost of acquiring information can further restrict their knowledge.

Therefore, this article adopts the state-agnostic setting, which is consistent with previous studies [36], and we view the dynamical model based on this setting as a foundational step toward richer extensions and as necessary for understanding the role of information (both its presence and absence) in decision-making [41]. Specifically, state-agnostic agents learn their strategies in a dynamic environment through a stateless version of $Q$-learning. We assume that each pairwise stochastic

game between agents involves $d$ states. For every state, the agent's action set is identical and contains $m$ available actions, that is, $\mathcal{A}(s_1) = \cdots = \mathcal{A}(s_d) = \mathcal{A} = \{a_1, \ldots, a_m\}$. Besides, the state-agnostic agent only needs to maintain a vector of $Q$-values for the $m$ actions. At time step $t$, for an arbitrary agent $i$, its $Q$-value vector $\mathbf{Q}_t^i$ is written as $\mathbf{Q}_t^i = [Q_t^i(a_1), \ldots, Q_t^i(a_m)]^\top$. If agent $i$ takes action $a_j \in \mathcal{A}$ to interact with its $k$ neighbors, and receives an immediate reward $r_t^i(a_j)$, then the $j$th element in its $Q$-value vector $\mathbf{Q}_t^i$ is updated according to (1), while other elements remain unchanged

$$Q_{t+1}^i(a_j) = Q_t^i(a_j) + \alpha\left[r_t^i(a_j) - Q_t^i(a_j)\right] \tag{1}$$

where $\alpha$ is the learning rate.

The strategy of a state-agnostic agent is defined as a probability distribution over its action set. Thus, the strategy of agent $i$ with $Q$-values $\mathbf{Q}_t^i$ at time step $t$ is represented as $\mathbf{x}_t^i = [x_t^i(a_1, \mathbf{Q}_t^i), \ldots, x_t^i(a_m, \mathbf{Q}_t^i)]^\top$, where for each action $a_j \in \mathcal{A}$, $x_t^i(a_j, \mathbf{Q}_t^i)$ denotes the probability that agent $i$ selects action $a_j$. We assume that each agent follows the Boltzmann exploration scheme, which determines $x_t^i(a_j, \mathbf{Q}_t^i)$ as follows:

$$x_t^i(a_j, \mathbf{Q}_t^i) = \frac{e^{\tau Q_t^i(a_j)}}{\sum_{a \in \mathcal{A}} e^{\tau Q_t^i(a)}} \tag{2}$$

where $\tau$ is the Boltzmann exploration temperature. The value of $\tau$ controls the tradeoff between exploration and exploitation in the learning process. If $\tau = 0$, the agent is in complete exploration (taking an action at random). If $\tau \to \infty$, the agent chooses the action with the maximum $Q$-value.

For state-agnostic $Q$-learning agents, our proposed multiagent learning framework is further detailed in Algorithm 1.

**Algorithm 1** Multiagent $Q$-Learning Framework in Networked Stochastic Games

**Require:** A $k$-regular graph $\mathcal{G} = \langle \mathcal{N}, \mathcal{E} \rangle$ representing the population, where $|\mathcal{N}| = n$. Each agent $i \in \mathcal{N}$ has a neighbor set $\mathcal{N}_i = \{j : e_{ij} \in \mathcal{E}\}$. Each edge $e_{ij} \in \mathcal{E}$ corresponds to a stochastic game $\langle \mathcal{N}_g, \mathcal{S}, \mathcal{A}, z, r, \mathbf{x}^1, \mathbf{x}^2 \rangle$ with $|\mathcal{N}_g| = 2$, state set $\mathcal{S} = \{s_1, \ldots, s_d\}$, and an action set $\mathcal{A} = \{a_1, \ldots, a_m\}$ that is identical for all agents and identical across all states. The payoff function is given by $r(s, \mathbf{a})$, and the state transition rule by $z_{s'}(s, \mathbf{a})$. $Q$-learning parameters: learning rate $\alpha$ and Boltzmann exploration temperature $\tau$. A maximum time step $T$.
**Initialization:**
**for** each agent $i \in \mathcal{N}$ **do**
  Initialize its $Q$-values $Q^i(a)$ for all $a \in \mathcal{A}$.
**end for**
**for** each edge $e_{ij} \in \mathcal{E}$ **do**
  Initialize the state of the corresponding stochastic game.
**end for**
Set $t \leftarrow 0$.
**while** $t < T$ **do**
  **for** each agent $i \in \mathcal{N}$ **do**
    Agent $i$ selects an action $a^i \in \mathcal{A}$ according to its strategy $\mathbf{x}^i$.
  **end for**
  **for** each agent $i \in \mathcal{N}$ **do**
    **for** each neighbor $j \in \mathcal{N}_i$ **do**
      Agents $i$ and $j$ play the two-agent stochastic game using their selected actions.
    **end for**
    Agent $i$ receives the immediate reward $r_t^i(a^i)$, averaged over its $k$ pairwise interactions.
  **end for**
  **for** each undirected edge $e_{ij} \in \mathcal{E}$ **do**
    The state $s \in \mathcal{S}$ of the stochastic game between $i$ and $j$ transitions to $s' \in \mathcal{S}$ according to $z_{s'}(s, \mathbf{a})$.
  **end for**
  **for** each agent $i \in \mathcal{N}$ **do**
    Update $Q$-value: $Q_{t+1}^i(a^i) \leftarrow Q_t^i(a^i) + \alpha[r_t^i(a^i) - Q_t^i(a^i)]$.
  **end for**
  $t \leftarrow t + 1$
**end while**

## IV. Formal Model for Multiagent $Q$-Learning Dynamics in Stochastic Games on Regular Graphs

In this section, we formally model the $Q$-learning dynamics for state-agnostic agents playing stochastic games within the concurrent learning framework described in Section III-A.

### A. Q-Value Dynamics for Individual Agents Within a Pair

In stochastic games, agents' decision-making not only yields payoffs that prompt them to update their strategies but also induces changes in the environmental states. These micro-level alterations subsequently drive the macro-level population dynamics. Consequently, our modelling approach initially focuses on the $Q$-value updates of agents and the state transitions resulting from pairwise interactions. Unlike normal-form games, characterizing state transitions presents a novel challenge. Since all stochastic games within the population occur in pairwise fashion, the joint actions of a pair of agents determine the state transitions between them, which in turn influence the agents' payoffs. Therefore, the pair approximation method, as adopted by Chu et al. [36], naturally serves as a framework for modelling these dynamics in stochastic games. Following their intuitive and effective framework, we first define the pair as follows.

*Definition 3 (Pair):* A pair is defined by the $Q$-values of two connected agents and their state, represented by a tuple $< \mathbf{Q}_t^1, s, \mathbf{Q}_t^2 >$, where $\mathbf{Q}_t^1 = [Q_t^1(a_1), \ldots, Q_t^1(a_m)]^\top$ and $\mathbf{Q}_t^2 = [Q_t^2(a_1), \ldots, Q_t^2(a_m)]^\top$ are the $Q$-value vectors of the two agents distinguished by the superscripts 1 and 2, and $s$ is the current state of the stochastic game the two agents play.

Now, we aim to model the $Q$-value dynamics of individual agents within a pair. To do so, we first need to calculate the average reward an agent receives from interacting with its neighbors in stochastic games. For each pairwise interaction, the reward an agent receives depends on both its own action and its neighbor's action, as well as the current game state. Formally, at time step $t$, in a pairwise interaction between an agent and a neighbor in state $s \in \mathcal{S}$, if the agent takes action

$a_i$ and the neighbor takes action $a_j$, the agent's reward is

$$r_t\left(a_i \left| s, a_j\right.\right) = \mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j \tag{3}$$

where $\mathbf{e}_i$ is the unit vector with the $i$th element equal to 1 and all other elements equal to 0. Furthermore, to compute the average payoff of an agent in stochastic games with its $k$ neighbors on a $k$-regular graph, we need to account for both the actions of each neighboring agent and the associated game states. Drawing inspiration from the concept of *neighbor configuration* employed to capture the behaviors of an agent's neighbors in networked normal-form games [39], we introduce a novel concept of *interaction configuration* for determining an agent's average payoff in networked stochastic games. Since our focus lies on the payoff of an agent within a pair, we can leverage the fact that for any agent in a pair, the strategy of one neighbor and the corresponding game state are already known. Thus, we only need to consider the configuration of the remaining $k-1$ uncertain pairwise interactions. The interaction configuration is formally defined as follows:

*Definition 4 (Interaction Configuration):* At time step $t$, for an arbitrary agent $i$ (where $i \in \{1,2\}$) in the pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$ on a $k$-regular graph, let $k^i_{(s_h,a_j),t}$ represent the number of pairwise interactions where agent $i$ interacts in state $s_h$ with a neighbor who takes action $a_j$. Note that the interaction in the focal pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$ is excluded from this count. The interaction configuration for agent $i$ is given by

$$\mathbf{w}_t^i = \left[k^i_{(s_1,a_1),t}, \ldots, k^i_{(s_1,a_m),t}, \ldots, k^i_{(s_d,a_1),t}, \ldots, k^i_{(s_d,a_m),t}\right]^\top$$

where $0 \le k^i_{(s_h,a_j),t} \le k-1$ for all $h \in \{1,\ldots,d\}$ and $j \in \{1,\ldots,m\}$, and the constraint $\sum_{h=1}^{d}\sum_{j=1}^{m} k^i_{(s_h,a_j),t} = k-1$ holds.

The interaction configuration formalizes the complex local interactions of an agent in networked stochastic games, where the agent may interact with different neighbors in various states. Based on this configuration, we can derive the following lemma to compute the immediate reward for an individual agent within a given pair.

*Lemma 1:* For a pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$ on a $k$-regular graph, if agents 1 and 2 within this pair select actions $a_i$ and $a_j$, respectively, and have interaction configurations $\mathbf{w}_t^1$ and $\mathbf{w}_t^2$, then the immediate reward received by agent 1 is given by

$$r_t^1\left(a_i \left| \mathbf{w}_t^1, s, a_j\right.\right) = \frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \mathbf{e}_i^\top \mathbf{M}' \mathbf{w}_t^1\right) \tag{4}$$

and the immediate reward received by agent 2 is given by

$$r_t^2\left(a_j \left| \mathbf{w}_t^2, s, a_i\right.\right) = \frac{1}{k}\left(\mathbf{e}_j^\top \mathbf{M}_s \mathbf{e}_i + \mathbf{e}_j^\top \mathbf{M}' \mathbf{w}_t^2\right) \tag{5}$$

where the matrix $\mathbf{M}' = [\mathbf{M}_{s_1} \mid \mathbf{M}_{s_2} \mid \cdots \mid \mathbf{M}_{s_d}]$, obtained by concatenating the payoff matrices for all $d$ states column-wise, has $m$ rows and $d \times m$ columns.

*Proof:* For an arbitrary pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$ on a $k$-regular graph, the immediate reward that an agent in this pair can receive depends on the rewards from all of the agent's pairwise interactions. This includes both the interaction in the focal pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$ and the other $k-1$ interactions, which are formalized by the agent's interaction configuration. For example, consider agent 1 in this pair. Given agent 1's interaction configuration $\mathbf{w}_t^1$ and agent 2's action $a_j$, and using (3), which derives the agent's reward from a specific pairwise interaction, we can calculate the immediate payoff when agent 1 chooses action $a_i$ as follows:

$$\begin{aligned} r_t^1\left(a_i \left| \mathbf{w}_t^1, s, a_j\right.\right) &= \frac{1}{k}\left(r_t\left(a_i \left| s, a_j\right.\right) + r_t\left(a_i \left| \mathbf{w}_t^1\right.\right)\right) \\ &= \frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \sum_{h=1}^{d}\sum_{l=1}^{m} k^1_{(s_h,a_l),t} r_t\left(a_i | s_h, a_l\right)\right) \\ &= \frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \sum_{h=1}^{d}\sum_{l=1}^{m} k^1_{(s_h,a_l),t} \mathbf{e}_i^\top \mathbf{M}_{s_h} \mathbf{e}_l\right) \\ &= \frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \mathbf{e}_i^\top \left[\mathbf{M}_{s_1} \middle| \mathbf{M}_{s_2} \middle| \ldots \middle| \mathbf{M}_{s_d}\right] \mathbf{w}_t^1\right) \\ &= \frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \mathbf{e}_i^\top \mathbf{M}' \mathbf{w}_t^1\right) \end{aligned} \tag{6}$$

where $\mathbf{M}' = [\mathbf{M}_{s_1} \mid \mathbf{M}_{s_2} \mid \cdots \mid \mathbf{M}_{s_d}]$ is the matrix obtained by concatenating the payoff matrices corresponding to all $d$ states column-wise. Similarly, given agent 2's interaction configuration $\mathbf{w}_t^2$ and the action $a_i$ taken by agent 1, we can derive agent 2's immediate payoff $r_t^2(a_j|\mathbf{w}_t^2, s, a_i)$. □

After receiving the immediate reward from game interactions, each state-agnostic agent updates its $Q$-value vector to perform strategy learning. We derive the following lemma to capture the dynamics of $Q$-values for each agent within an arbitrary pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$.

*Lemma 2:* At time step $t$, for an arbitrary pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$, where agent 1 with $Q$-values $\mathbf{Q}_t^1$ and agent 2 with $Q$-values $\mathbf{Q}_t^2$ interact in state $s$, suppose agent 1 takes action $a_i$ with interaction configuration $\mathbf{w}_t^1$, and agent 2 takes action $a_j$ with interaction configuration $\mathbf{w}_t^2$. Then, the rate at which agent 1 updates its $Q$-value for action $a_i$ can be expressed as follows:

$$v_i^1\left(\mathbf{Q}_t^1, a_i, \mathbf{w}_t^1, a_j \middle| s\right) = \alpha\left[\frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \mathbf{e}_i^\top \mathbf{M}' \mathbf{w}_t^1\right) - Q_t^1\left(a_i\right)\right] \tag{7}$$

the rate at which agent 2 updates its $Q$-value for action $a_j$ is

$$v_j^2\left(\mathbf{Q}_t^2, a_j, \mathbf{w}_t^2, a_i \middle| s\right) = \alpha\left[\frac{1}{k}\left(\mathbf{e}_j^\top \mathbf{M}_s \mathbf{e}_i + \mathbf{e}_j^\top \mathbf{M}' \mathbf{w}_t^2\right) - Q_t^2\left(a_j\right)\right]. \tag{8}$$

*Proof:* Consider agent 1 in a given pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2\rangle$, where agent 1 chooses action $a_i$ with interaction configuration $\mathbf{w}_t^1$ and agent 2 chooses action $a_j$ with interaction configuration $\mathbf{w}_t^2$. By (1), we can derive the difference equation for agent 1's $Q$-value for the $i$th action. Substituting the reward information from (4) into this difference equation, in the continuous-time limit, we obtain the following equation that governs the dynamics of the change in the $Q$-value for action $a_i$ when agent 1 takes action $a_i$ at time step $t$

$$\begin{aligned} v_i^1\left(\mathbf{Q}_t^1, a_i, \mathbf{w}_t^1, a_j \middle| s\right) &= Q_{t+1}^1\left(a_i\right) - Q_t^1\left(a_i\right) \\ &= \alpha\left[r_t^1\left(a_i \middle| \mathbf{w}_t^1, s, a_j\right) - Q_t^1\left(a_i\right)\right] \\ &= \alpha\left[\frac{1}{k}\left(\mathbf{e}_i^\top \mathbf{M}_s \mathbf{e}_j + \mathbf{e}_i^\top \mathbf{M}' \mathbf{w}_t^1\right) - Q_t^1\left(a_i\right)\right]. \end{aligned} \tag{9}$$

Similarly, using (1) and (5), we can derive $v_j^2(\mathbf{Q}_t^2, a_j, \mathbf{w}_t^2, a_i|s)$, which expresses the rate at which agent 2 increases or decreases its $Q$-value for action $a_j$. □

The asynchronous update property of the $Q$-learning algorithm ensures that at any given time step, an agent updates only the $Q$-value for the action it has taken, rather than updating all $Q$-values. Therefore, when agent 1 in a given pair $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2 \rangle$ takes action $a_i$ at time step $t$ with interaction configuration $\mathbf{w}_t^1$, and agent 2 takes action $a_j$, the rate of change of agent 1's $Q$-value vector $\mathbf{Q}_t^1$ can be represented as: $\mathbf{v}^1(\mathbf{Q}_t^1, a_i, \mathbf{w}_t^1, a_j|s) = [v_1^1(\mathbf{Q}_t^1, a_i, \mathbf{w}_t^1, a_j|s), \ldots, v_m^1(\mathbf{Q}_t^1, a_i, \mathbf{w}_t^1, a_j|s)]^\top$, where the $i$th element is calculated using (7), and all other elements are zero. Similarly, for agent 2 with interaction configuration $\mathbf{w}_t^2$, when agent 2 takes action $a_j$ and agent 1 takes action $a_i$, the rate of change of agent 2's $Q$-value vector $\mathbf{Q}_t^2$ is represented as follows: $\mathbf{v}^2(\mathbf{Q}_t^2, a_j, \mathbf{w}_t^2, a_i|s) = [v_1^2(\mathbf{Q}_t^2, a_j, \mathbf{w}_t^2, a_i|s), \ldots, v_m^2(\mathbf{Q}_t^2, a_j, \mathbf{w}_t^2, a_i|s)]^\top$, where the $j$th element is calculated using (8), and all other elements are zero.

### B. State Distribution and Opponent Strategies for an Individual Q-Learner

Agents encounter different types of pairwise interactions with varying probabilities during the interaction process. These interactions are distinguished by the game state and neighbor behavior. While we have derived the payoff calculation for an agent in a specific pairwise interaction, the probability of an agent participating in any given interaction is still unclear. To address this, we now aim to derive the agent's state distribution and the opponent strategy under each state, enabling us to compute the probabilities of the agent engaging in various pairwise interactions.

It is important to note that in networked stochastic games, state transitions lead to significant population heterogeneity, where different agents may encounter distinct state distributions and opponent strategies. This heterogeneity breaks the homogeneity assumption of mean-field theory commonly used in modelling population dynamics [24], [39]. However, the pair approximation method we employ has the advantage of modelling this heterogeneity. By utilizing the probability distribution of pairs, we can derive the different state distributions and opponent strategies faced by agents based on their own $Q$-values.

Let $p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t)$ represent the proportion of pairs in the population where agents 1 and 2 are in state $s$ at time step $t$ and have $Q$-values $\mathbf{Q}_t^1$ and $\mathbf{Q}_t^2$, respectively. Using the compatibility conditions of the pair-approximation method [72], the marginal probability of finding agents with $Q$-values $\mathbf{Q}_t^1$ at time step $t$ is obtained by integrating the joint distribution $p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t)$ over agent 2's $Q$-values and summing over all states

$$p\left(\mathbf{Q}_t^1, t\right) = \sum_{s\in\mathcal{S}} \int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) dV_2 \tag{10}$$

where we define $dQ_t^1(a_1)\ldots dQ_t^1(a_m)$ (resp.$dQ_t^2(a_1)\ldots dQ_t^2(a_m)$) as $dV_1$ (resp.$dV_2$) in this article. We further establish the following lemma to give the state distribution and opponent strategy in each state for an individual agent from its $Q$-values.

*Lemma 3:* For an individual agent with $Q$-values $\mathbf{Q}_t^1$ at time step $t$, the probability of participating in a pairwise interaction occurring in state $s$ is given by

$$p\left(s\middle|\mathbf{Q}_t^1, t\right) = \frac{\int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) dV_2}{\sum_{s\in\mathcal{S}} \int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) dV_2}. \tag{11}$$

As for the opponent strategy for an agent interacting in state $s$, we denote the conditional probability of this agent encountering a neighbor who takes action $a_j$ as $p(a_j|\mathbf{Q}_t^1, s, t)$. Then, $p(a_j|\mathbf{Q}_t^1, s, t)$ is given as follows:

$$p\left(a_j\middle|\mathbf{Q}_t^1, s, t\right) = \frac{\int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) \frac{e^{\tau Q_t^2(a_j)}}{\sum_{a\in\mathcal{A}} e^{\tau Q_t^2(a)}} dV_2}{\int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) dV_2}. \tag{12}$$

*Proof:* According to the conditional probability formula, for an agent with $Q$-values equal to $\mathbf{Q}_t^1$, the probability of participating in a pairwise stochastic game in state $s$ is

$$p\left(s\middle|\mathbf{Q}_t^1, t\right) = \frac{p\left(\mathbf{Q}_t^1, s, t\right)}{p\left(\mathbf{Q}_t^1, t\right)} \tag{13}$$

where the joint probability $p(\mathbf{Q}_t^1, s, t)$ represents the proportion of pairs in which agent 1 has $Q$-values $\mathbf{Q}_t^1$ and the state is $s$. This joint probability can be derived by integrating the probability distribution of pairs $\langle \mathbf{Q}_t^1, s, \mathbf{Q}_t^2 \rangle$ with respect to the random vector $\mathbf{Q}_t^2$. Then, by computing the denominator $p(\mathbf{Q}_t^1, t)$ on the right-hand side of (13) using (10), we obtain (11).

In a similar way, given that a focal agent with $Q$-values $\mathbf{Q}_t^1$ interacts in state $s$, the probability for this agent encountering an opponent who takes action $a_j$ is calculated by

$$\begin{aligned} p\left(a_j\middle|\mathbf{Q}_t^1, s, t\right) &= \frac{p\left(\mathbf{Q}_t^1, s, a_j, t\right)}{p\left(\mathbf{Q}_t^1, s, t\right)} \\ &= \frac{\int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) x_t^2\left(a_j, \mathbf{Q}_t^2\right) dV_2}{p\left(\mathbf{Q}_t^1, s, t\right)} \end{aligned} \tag{14}$$

where $x_t^2(a_j, \mathbf{Q}_t^2)$ is the probability that agent 2 in the pair $< \mathbf{Q}_t^1, s, \mathbf{Q}_t^2 >$ takes action $a_j$, and $x_t^2(a_j, \mathbf{Q}_t^2)$ is obtained by (2). □

Building on Lemma 3, we introduce the following lemma to characterize the probability of an agent participating in an arbitrary type of pairwise interaction.

*Lemma 4:* For an arbitrary agent with $Q$-values equal to $\mathbf{Q}_t^1$ at time step $t$, the probability that this agent finds itself engaged in an interaction in state $s_h$ with an opponent taking action $a_l$ is given by

$$p\left(s_h, a_l\middle|\mathbf{Q}_t^1, t\right) = \frac{\int \cdots \int p\left(\mathbf{Q}_t^1, s_h, \mathbf{Q}_t^2, t\right) \frac{e^{\tau Q_t^2(a_l)}}{\sum_{a\in\mathcal{A}} e^{\tau Q_t^2(a)}} dV_2}{\sum_{s\in\mathcal{S}} \int \cdots \int p\left(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t\right) dV_2}. \tag{15}$$

*Proof Sketch:* To derive the probability of an agent with $Q$-values $\mathbf{Q}_t^1$ engaging in a specific interaction $(s_h, a_l)$, we first determine the probability that the agent plays in state $s_h$, denoted as $p(s_h|\mathbf{Q}_t^1, t)$. Next, we obtain the conditional probability that this agent encounters a neighbor taking action $a_l$ in state $s_h$, expressed as $p(a_l|\mathbf{Q}_t^1, s_h, t)$. Multiplying these two probabilities yields

$$p\left(s_h, a_l\middle|\mathbf{Q}_t^1, t\right) = p\left(s_h\middle|\mathbf{Q}_t^1, t\right) \cdot p\left(a_l\middle|\mathbf{Q}_t^1, s_h, t\right).$$

Thus, we can prove this lemma by using (11) and (12).

### C. Probability Distribution of Interaction Configuration

The interaction configuration can take various values depending on the size of the action set $\mathcal{A}$, the state set $\mathcal{S}$, and the degree of the regular graph. At time step $t$, let $p(\mathbf{w}_t^i|\mathbf{Q}_t^i)$ denote the probability that an arbitrary agent $i$ within a given pair, possessing $Q$-values $\mathbf{Q}_t^i$, assumes a specific interaction configuration $\mathbf{w}_t^i$. Based on Lemma 4, the pair approximation method partitions the population into groups of agents sharing identical $Q$-values and applies a mean-field approximation within each group. Consequently, for a given agent, all pairwise interactions within its local neighborhood are represented by the same average effect [i.e., (15)]. Under this approximation, we make the following assumption:

*Assumption 1:* For an agent within a pair with specific $Q$-values, all different pairwise interactions contributing to a particular local configuration are represented by the same average effect under (15). Accordingly, these pairwise interactions are assumed to be statistically independent.

Building on this, the following lemma derives the probability distribution over an individual agent's possible interaction configurations.

*Lemma 5:* The interaction configuration of an individual agent within a given pair follows a multinomial distribution. Thus, the probability that an arbitrary agent $i$, having $Q$-values $\mathbf{Q}_t^i$, has a certain interaction configuration $\mathbf{w}_t^i$ at time step $t$, denoted by $p(\mathbf{w}_t^i|\mathbf{Q}_t^i)$, is given as follows:

$$p\left(\mathbf{w}_t^i\middle|\mathbf{Q}_t^i\right) = (k-1)!\prod_{h=1}^{d}\prod_{l=1}^{m}\frac{p\left(s_h,a_l\middle|\mathbf{Q}_t^i,t\right)^{k^i_{(s_h,a_l),t}}}{k^i_{(s_h,a_l),t}!}. \tag{16}$$

*Proof:* Consider an arbitrary agent $i$ in a given pair, whose neighborhood consists of $k-1$ interaction slots and whose interaction configuration is denoted by $\mathbf{w}_t^i$. Let $k^i_{(s,a),t}$ be the number of interactions occurring in state $s$ with a neighbor selecting action $a$. First, we note that the number of ways to assign $k^i_{(s,a),t}$ interactions of type $(s,a)$ to the $k-1$ available slots is given by the binomial coefficient $\binom{k-1}{k^i_{(s,a),t}}$. As the remaining $k-1-k^i_{(s,a),t}$ slots must then be allocated among the other interaction types, and this process is carried out sequentially for all interaction types, the total number of spatial arrangements is described by the multinomial coefficient $\binom{k-1}{k^i_{(s_1,a_1),t}}\binom{k-1-k^i_{(s_1,a_1),t}}{k^i_{(s_1,a_2),t}},\ldots,\binom{k^i_{(s_d,a_m),t}}{k^i_{(s_d,a_m),t}} = ((k-1)!)/\left(\prod_{h=1}^{d}\prod_{l=1}^{m}k^i_{(s_h,a_l),t}!\right)$.

Next, based on Assumption 1, for a particular spatial arrangement conforming to the configuration $\mathbf{w}_t^i$, the probability is $\prod_{h=1}^{d}\prod_{l=1}^{m}p(s_h,a_l|\mathbf{Q}_t^i,t)^{k^i_{(s_h,a_l),t}}$.

Since every arrangement corresponding to $\mathbf{w}_t^i$ is equally likely, the overall probability for the agent $i$ to have configuration $\mathbf{w}_t^i$ is given by multiplying the number of arrangements (the multinomial coefficient) with the probability of any specific arrangement. Thus, the interaction configuration of an individual agent follows a multinomial distribution. □

### D. Evolutionary Dynamics of the Population State

Over time, the strategy learning of agents and state transitions jointly drive the evolution of the population. Using the pair-approximation method, the probability distribution of pairs provides sufficient insights into agent behaviors and the population's environmental information. Consequently, the population state can be effectively represented by this distribution. We denote the proportion of pairs $\langle\mathbf{Q}_t^1,s,\mathbf{Q}_t^2\rangle$ in the population as $p(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t)$. By tracking the temporal evolution of this distribution (i.e., deriving $\partial p(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t)/\partial t$), we can model the population dynamics.

Next, we present an effective step-by-step approach to solving the population state evolution. Specifically, we first denote the proportion of state $s$ in the population at time $t$ as $p(s,t)$, which can also be interpreted as the proportion of pairs in the population engaged in interactions occurring in state $s$. Then, for clarity, we represent the $Q$-values of an agent pair within $\langle\mathbf{Q}_t^1,s,\mathbf{Q}_t^2\rangle$ as $\mathbf{P}=[\mathbf{Q}_t^1,\mathbf{Q}_t^2]$. We use the function $p(\mathbf{P},t\mid s)$ to describe the distribution of agent pairs' $Q$-values within the state space $s$, where agents interact in state $s$. The function $p(\mathbf{P},t\mid s)$ can also be understood as the proportion of agent pairs with $Q$-values $\mathbf{P}$ among all agent pairs interacting in state $s$. Finally, since the joint probability $p(\mathbf{P},s,t)$ can be derived by multiplying the marginal probability $p(s,t)$ by the conditional probability $p(\mathbf{P},t\mid s)$, we have

$$\begin{aligned}\frac{\partial p(\mathbf{P},s,t)}{\partial t} &= \frac{\partial\left[p(s,t)\,p(\mathbf{P},t|s)\right]}{\partial t}\\ &= p(s,t)\frac{\partial p(\mathbf{P},t\mid s)}{\partial t}+p(\mathbf{P},t|s)\frac{dp(s,t)}{dt}\end{aligned} \tag{17}$$

where $p(\mathbf{P},t\mid s)=p(\mathbf{P},s,t)/p(s,t)$, and we have

$$p(s,t)=\int\cdots\int p\left(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t\right)\,dV_1\,dV_2. \tag{18}$$

Equation (17) states that to solve the evolution of the population state, we can proceed in two steps: first, we derive the temporal evolution of the *environmental state* (see Lemma 6), which is defined as the distribution of states in the population. Then, we determine the evolution of the probability distribution of agent pairs' $Q$-values within each state space.

*Lemma 6:* For a networked population where each agent plays stochastic games with its neighbors, the state transitions in each pairwise interaction at the microscopic level, determined by agent strategies and the state transition rule, drive the macroscopic evolution of the environmental state of the population. Consequently, the evolution of $p(s,t)$ can be described by the following differential equation:

$$\begin{aligned}\frac{dp(s,t)}{dt} &= \sum_{s'\in\mathcal{S}}\int\cdots\int p\left(\mathbf{Q}_t^1,s',\mathbf{Q}_t^2,t\right)\\ &\quad\times\sum_{i=1}^{m}\sum_{j=1}^{m}x_t^1\left(a_i,\mathbf{Q}_t^1\right)x_t^2\left(a_j,\mathbf{Q}_t^2\right)z_s\left(s',a_i,a_j\right)dV_1dV_2\\ &\quad-\int\cdots\int p\left(\mathbf{Q}_t^1,s,\mathbf{Q}_t^2,t\right)dV_1dV_2\end{aligned} \tag{19}$$

where $z_s(s',a_i,a_j)$ represents the transition probability from state $s'$ to $s$ for a pair of agents, given that agents 1 and 2 take actions $a_i$ and $a_j$, respectively.

*Proof:* At time step $t$, the probability $p(s,t)$ evolves due to interactions transitioning into and out of state $s$. The rate

of change of $p(s,t)$ is thus determined by the net balance between the inflow and outflow of interactions associated with state $s$. The inflow term aggregates all interactions that transition from any other state $s' \neq s$ into state $s$, and is given by $\sum_{s'\in\mathcal{S},\, s'\neq s} \int \cdots \int p(\mathbf{Q}_t^1, s', \mathbf{Q}_t^2, t) \sum_{i=1}^m \sum_{j=1}^m x_t^1(a_i, \mathbf{Q}_t^1) x_t^2(a_j, \mathbf{Q}_t^2) z_s(s', a_i, a_j)\, dV_1 dV_2$, while the outflow term encompasses all interactions that leave state $s$ for other states, given by $\sum_{s'\in\mathcal{S},s'\neq s} \int \cdots \int p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t) \sum_{i=1}^m \sum_{j=1}^m x_t^1(a_i, \mathbf{Q}_t^1) x_t^2(a_j, \mathbf{Q}_t^2) z_{s'}(s, a_i, a_j)\, dV_1 dV_2$. To express both inflow and outflow in a unified form, we rewrite the summations over $s' \neq s$ as summations over all $s' \in \mathcal{S}$. Specifically, the inflow term can be reformulated as $\sum_{s'\in\mathcal{S}} \int \cdots \int p(\mathbf{Q}_t^1, s', \mathbf{Q}_t^2, t) \sum_{i=1}^m \sum_{j=1}^m x_t^1(a_i, \mathbf{Q}_t^1) x_t^2(a_j, \mathbf{Q}_t^2) z_s(s', a_i, a_j) dV_1 dV_2 - \int \cdots \int p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t) \sum_{i=1}^m \sum_{j=1}^m x_t^1(a_i, \mathbf{Q}_t^1) x_t^2(a_j, \mathbf{Q}_t^2) z_s(s, a_i, a_j)\, dV_1 dV_2$. Note that the above minus term represents the proportion of agent pairs that remain in state $s$. Therefore, adding this term to the outflow term yields exactly the total proportion of pairs currently in state $s$, that is, $\int \cdots \int p(\mathbf{Q}_t^1, s, \mathbf{Q}_t^2, t)\, dV_1 dV_2$. Therefore, by algebraically combining the inflow and outflow terms, we obtain the compact expression for the macroscopic dynamics of the environmental state as given in (19). □

We then formalize the evolution of the distribution of pairs in the following theorem.

*Theorem 1:* The time evolution of the probability distribution of pairs in the population is governed by the following partial differential equation:

$$\begin{aligned}
&\frac{\partial p(\mathbf{P}, s, t)}{\partial t} \\
&= \sum_{s'\in\mathcal{S}} \left[ p(s', t) \left( p(\mathbf{P}, t | s') + \frac{\partial p'(\mathbf{P}, t | s')}{\partial t} \right) \right. \\
&\qquad \left. \times \sum_{i=1}^m \sum_{j=1}^m x_t^1(a_i, \mathbf{Q}_t^1)\, x_t^2(a_j, \mathbf{Q}_t^2)\, z_s(s', a_i, a_j) \right] \\
&\quad - p(\mathbf{P}, s, t)
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
&\frac{\partial p'(\mathbf{P}, t | s)}{\partial t} \\
&= -\sum_{i=1}^m \sum_{j=1}^m \sum_{\mathbf{w}_t^1 \in \mathcal{W}} \left[ v_i^1(\mathbf{Q}_t^1, a_i, \mathbf{w}_t^1, a_j | s) x_t^2(a_j, \mathbf{Q}_t^2) \right. \\
&\qquad \left. \times \frac{\partial \left[ x_t^1(a_i, \mathbf{Q}_t^1) p(\mathbf{w}_t^1 | \mathbf{Q}_t^1) p(\mathbf{P}, t | s) \right]}{\partial Q_t^1(a_i)} \right] \\
&\quad - \sum_{i=1}^m \sum_{j=1}^m \sum_{\mathbf{w}_t^2 \in \mathcal{W}} \left[ v_j^2(\mathbf{Q}_t^2, a_j, \mathbf{w}_t^2, a_i | s) x_t^1(a_i, \mathbf{Q}_t^1) \right. \\
&\qquad \left. \times \frac{\partial \left[ x_t^2(a_j, \mathbf{Q}_t^2) p(\mathbf{w}_t^2 | \mathbf{Q}_t^2) p(\mathbf{P}, t | s) \right]}{\partial Q_t^2(a_j)} \right].
\end{aligned} \tag{21}$$

*Proof:* Modelling the evolution of $p(\mathbf{P}, s, t)$ also entails characterizing the evolution of $p(\mathbf{P}, t \mid s)$. To this end, assume there are $d$ distinct $Q$-value spaces, each corresponding to a state $s \in \mathcal{S}$. The $Q$-value space of state $s$ is a $2m$-dimensional Euclidean space, referred to as the state space $s$. At time $t$, each agent pair interacting in state $s$ occupies a position in this space based on their $Q$-values, so that $p(\mathbf{P}, t \mid s)$ also denotes the proportion of agent pairs at $\mathbf{P}$ within state $s$. Over time, agent pairs may transition from $\mathbf{P}$ to another position $\mathbf{P}'$ within the same state space or to $\mathbf{P}'$ in a different one. Modelling these transitions across state spaces is challenging. However, for state-agnostic $Q$-learners—whose strategy updating and state transitions are independent—the modelling process can be decomposed into two stages for tractability: first, we assume fixed states and model how agents' behavior alters the $Q$-value distribution within a state space. Then, we incorporate state transitions by updating state information according to the transition rule.

When assuming no state transitions occur, the evolution of the $Q$-value distribution in a state space is solely due to internal position changes induced by $Q$-value updates.

Following the master equation approach presented in [53], and by categorizing the transitions in agent pair positions based on their distinct joint actions and various interaction configurations, (21) can be derived.

Based on the derivation of (21), we now consider the actual dynamics introduced by state transitions. For all agent pairs reaching position P in space $s$ without considering state transitions, in practice only a fraction remain in space $s$, as state transitions cause some of them to be redirected to different spaces. Similarly, agent pairs from other spaces may transition into space $s$ and arrive at position P. Based on this analysis, we update the state information according to the behavior of agent pairs driving their transitions across different state spaces, yielding the actual rate of change in the $Q$-value distribution of agent pairs within each state space

$$\begin{aligned}
&\frac{\partial p(\mathbf{P}, t | s)}{\partial t} \\
&= \frac{1}{p(s,t)} \left[ \sum_{s'\in\mathcal{S}} \left[ p(s', t) \left( p(\mathbf{P}, t | s') + \frac{\partial p'(\mathbf{P}, t | s')}{\partial t} \right) \right. \right. \\
&\qquad \left. \times \sum_{i=1}^m \sum_{j=1}^m x_t^1(a_i, \mathbf{Q}_t^1)\, x_t^2(a_j, \mathbf{Q}_t^2)\, z_s(s', a_i, a_j) \right] \\
&\qquad \left. - p(s,t)\, p(\mathbf{P}, t | s) - p(\mathbf{P}, t | s) \frac{dp(s,t)}{dt} \right].
\end{aligned} \tag{22}$$

By substituting (22) into (17), we can prove Theorem 1 □

Through Theorem 1, we can track the temporal evolution of the distribution of pairs in the population. Furthermore, at each time step, we can derive important information about expected agent behaviors from the probability distribution of pairs. Specifically, we can calculate the expected $Q$-value of action $a_i$ by

$$\mathbb{E}[Q_t(a_i)] = \int Q_t(a_i)\, p(Q_t(a_i), t)\, dQ_t(a_i) \tag{23}$$

where $p(Q_t(a_i), t)$ represents the proportion of agents in the population whose $Q$-value of action $a_i$ equals $Q_t(a_i)$, and is

$$\begin{aligned}
p(Q_t(a_i), t) = \sum_{s\in\mathcal{S}} \int \cdots \int p(\mathbf{Q}_t, s, \mathbf{Q}_t^2, t) \\
dQ_t(a_1) \ldots dQ_t(a_{i-1})\, dQ_t(a_{i+1}) \ldots dQ_t(a_m)\, dV_2.
\end{aligned} \tag{24}$$

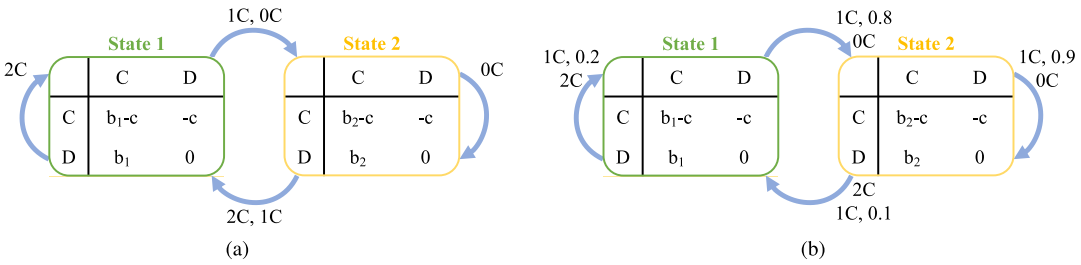



Fig. 1. Illustration of two-state donation games with different state transition rules. In both (a) and (b), each agent selects an action from the set $\mathcal{A} = \{a_1, a_2\} = \{\text{cooperate (C)}, \text{defect (D)}\}$, where $a_1$ corresponds to cooperation (C) and $a_2$ to defection (D). In state $i \in \{1, 2\}$, choosing C entails a cost $c$ to the agent and provides a benefit $b_i$ to the partner, whereas choosing D incurs no cost. Blue arrows indicate possible transitions between states. (a) Arrow labels indicate how the number of cooperators determines the transition. For example, in state 1, if one agent cooperates and the other defects (1C), they transition to state 2. Transitions occur with probability 0 or 1. (b) Arrow labels indicate the number of cooperators and the corresponding transition probability. In both cases, parameters are set to $b_1 = 1.8$, $b_2 = 0.3$, and $c = 0.1$.

Besides, the expected strategy for action $a_i$ is given by

$$\mathbb{E}\left[x_t\left(a_i\right)\right] = \int \ldots \int p\left(\mathbf{Q}_t, t\right) x_t\left(a_i, \mathbf{Q}_t\right) dQ_t\left(a_1\right) \ldots dQ_t\left(a_m\right). \tag{25}$$

## V. Experiments

In this section, we verify that the dynamic model developed in Section IV accurately captures the learning dynamics of structured populations in stochastic games by comparing it with agent-based simulation results across various experimental scenarios. In addition, we analyze how state transition rules and underlying population structures influence agent behavior.

### A. Environmental Stochasticity Leads to Varying Population Dynamics

Stochastic games can be classified into two categories based on the nature of state transitions: deterministic and probabilistic. To evaluate our dynamical model, we first validate its accuracy under both types of transition rules.

We conduct experiments using a two-state donation game, where agents engage in donation games with different payoff values based on their current state. This game is a special instance of the PD, commonly used to explore the evolution of cooperation in social dilemmas. Fig. 1 illustrates the payoff matrix for the donation game in each state and outlines the deterministic and probabilistic state transition rules considered. Note that even with deterministic transitions, the game is still called a "stochastic game," as this represents a special case of the general framework, and the next state may still depend on chance events if the agents' strategies are stochastic.

In this article, unless otherwise specified, we follow previous studies [36], [39] and set the learning rate $\alpha = 0.4$ and the Boltzmann exploration temperature $\tau = 2$ for both agent-based simulations and theoretical analyses. For the agent-based simulations, the population size $n$ is set to 10 000, with 100 simulations run for each setting to mitigate randomness. We assume that the initial state for each agent pair is determined randomly. As for the initialization of agents' $Q$-values, following related classical works [24], [51], we consider that the $Q$-values of agents are sampled from Beta distributions. In our experiments on the two-state donation game, without loss of generality, we assume that the initial $Q$-values for the first

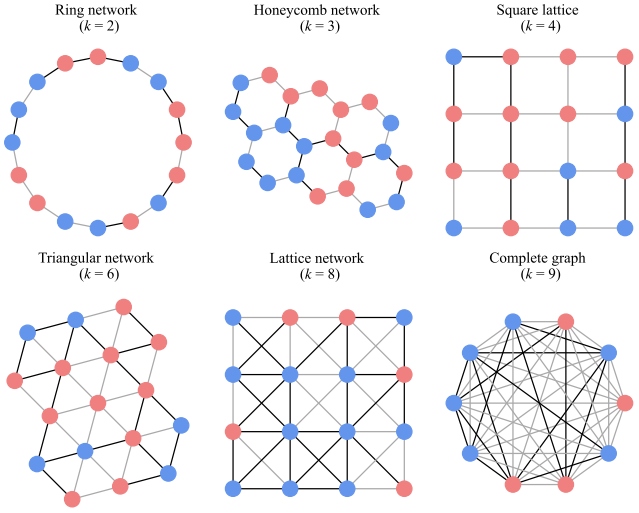



Fig. 2. Examples of regular graphs. For clarity, we show cropped portions (rather than full layouts) of regular graphs with periodic boundary conditions, where not all wrap-around edges of boundary nodes are depicted, although every node in the full graph has the same degree. Different node colors indicate the distinct actions chosen by agents, while different edge colors represent different states.

action $a_1$ (C) and the second action $a_2$ (D) follow different Beta distributions: $\text{Beta}(20, 80, r_{\min}, r_{\max})$ for action C and $\text{Beta}(10, 90, r_{\min}, r_{\max})$ for action D. The first two parameters of the Beta distribution enable us to flexibly control the shape of the probability density function, while the latter two define the support as $[r_{\min}, r_{\max}]$, where $r_{\min} = -0.1$ and $r_{\max} = 1.8$ denote the minimum and maximum payoffs for agents in the stochastic game. Besides, we consider agents distributed on a regular graph with degree $k = 4$, using the widely adopted square lattice with *periodic boundary conditions* [73], [74], [75]. In addition to the square lattice, other commonly used regular graph topologies include the ring network ($k = 2$), honeycomb network ($k = 3$), triangular network ($k = 6$), and lattice network ($k = 8$). Examples of these regular graphs are illustrated in Fig. 2.

Fig. 3 displays the mean values from the agent-based simulations as solid lines, while the shaded areas represent results derived from our theoretical model. Agent behaviors are described by the average $Q$-values and average strategies.

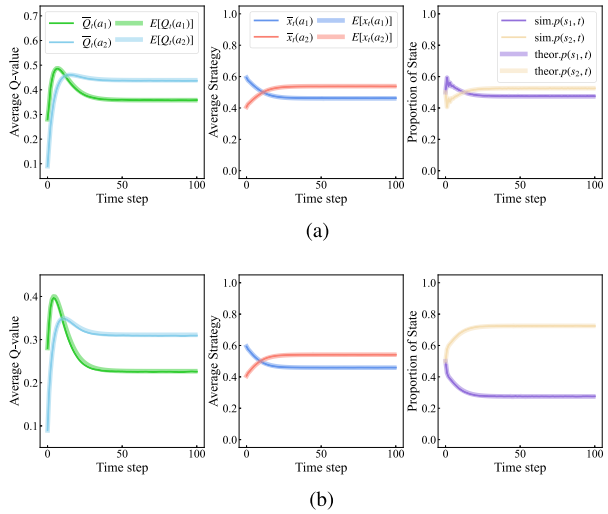



Fig. 3. Evolution of agent behaviors and that of the environmental state of the population under (a) deterministic and (b) probabilistic state transition rules. In (a) and (b), we sequentially illustrate the temporal evolution of the average $Q$-values, average strategies, and environmental state of the population, comparing results from both the simulation and dynamical model.

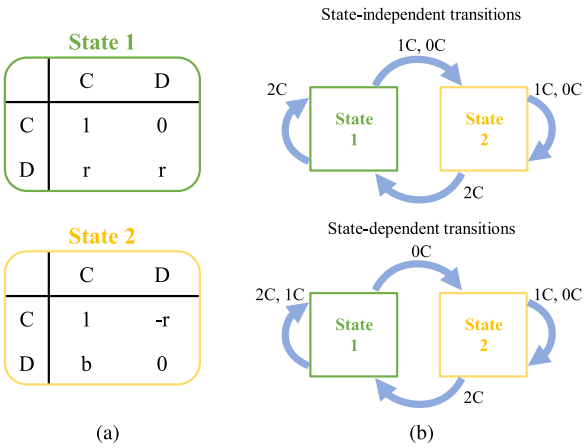



Fig. 4. Two-state stochastic game with transitions between an SH game and a PD game. The parameter values for the payoff matrices are set to $b = 1.2$ and $r = 0.1$. (a) Payoff matrices. (b) State transition rules.

For simulations, the average $Q$-value of an action is computed as the mean of all agents' $Q$-values for that action, while the average strategy of an action is the mean probability across all agents of selecting that action. For theoretical predictions, the results are obtained using (23) and (25). We observe that, under both deterministic and probabilistic transition rules, the time evolution of agent behaviors and the environmental state predicted by our theoretical model closely align with the agent-based simulation results. Notably, the dynamics of the environmental state can differ significantly between the two transition patterns. Under the probabilistic transition rule, the proportion of agent pairs interacting in state 2 is considerably higher than under the deterministic rule.

These findings demonstrate that environmental stochasticity introduced by probabilistic transitions can substantially influence population dynamics compared to deterministic ones. This suggests that carefully designing transition probabilities can effectively steer agent behavior to resolve social dilemmas.

### B. Impact of Network Structure on Agent Behavior

We now aim to validate the applicability of our theoretical model in capturing the learning dynamics of structured populations across various graph topologies, and to investigate how population structure influences dynamics in stochastic games.

We consider a two-state stochastic game where the states correspond to an SH game and a PD game, respectively. In either state, agents select actions from the set $\mathcal{A} = \{a_1, a_2\} = \{\text{cooperate (C)}, \text{defect (D)}\}$. Fig. 4 presents the payoff matrices for both games and outlines two different transition rules. A state transition rule is termed state-independent if the next state's determination relies solely on the agents' joint action at the current time step, independent of their current state. In the example shown in Fig. 4, under a typical state-independent rule, if either agent defects in the current step, they transition to state 2; otherwise, they stay in state 1. In contrast, the state-dependent transition rule takes into account both the agents' joint actions and their current state to determine the next state. As illustrated in Fig. 4, under the state-dependent rule, the two agents remain in their current state at the next time step when selecting the joint action "1C." In this stochastic game, which transitions between an SH game and a PD game, Fig. 5 demonstrates that our theoretical model accurately captures the time evolution of average $Q$-values for populations with various structures, under both state-independent and state-dependent transition rules. We also compare the results of structured populations with those of a well-mixed population, which is represented by a complete graph with degree $k = 99$. Notably, in this experiment, we set $Q_0(a_1) \sim \text{Beta}(60, 60, -0.1, 1.2)$ and $Q_0(a_2) \sim \text{Beta}(5, 5, -0.1, 1.2)$. This setting for the Beta distributions differs from that used in the previous experiment, allowing us to further confirm that the applicability of our model is not constrained by the choice of initial conditions.

To further explore the impact of population structure on cooperative behaviors, we present the dynamics of cooperation rates for different structured populations, as obtained from our theoretical model and agent-based simulations (Fig. 6). Note that the theoretical results for the well-mixed population with degree $k = 99$ are obtained from the dynamical model of Chu et al. [36]. Notably, the effects of population structure on agent behavior differ significantly depending on the transition rules. As illustrated in Fig. 6(a), under the state-independent transition rule, increasing the degree of the regular graph leads to a decrease in the cooperation rate. Conversely, Fig. 6(b) indicates that under the state-dependent transition rule, agents interacting on a regular graph with a higher degree exhibit more cooperative behavior. Overall, agent populations achieve higher cooperation rates under the state-dependent rule compared to the state-independent rule.

### C. Important Role of State Transitions in Promoting the Evolution of Cooperation

The emergence and evolution of cooperation in multiagent systems has been a longstanding and prominent research

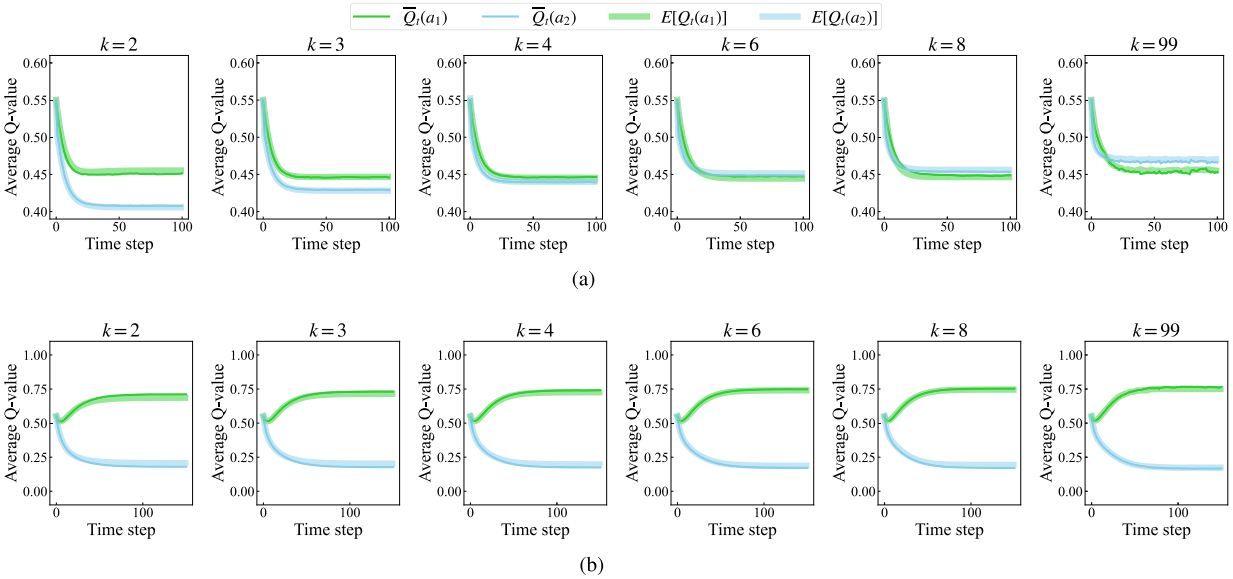



Fig. 5. Evolution of the average $Q$-values of agent populations with various population structures. (a) State-independent transitions. (b) State-dependent transitions.

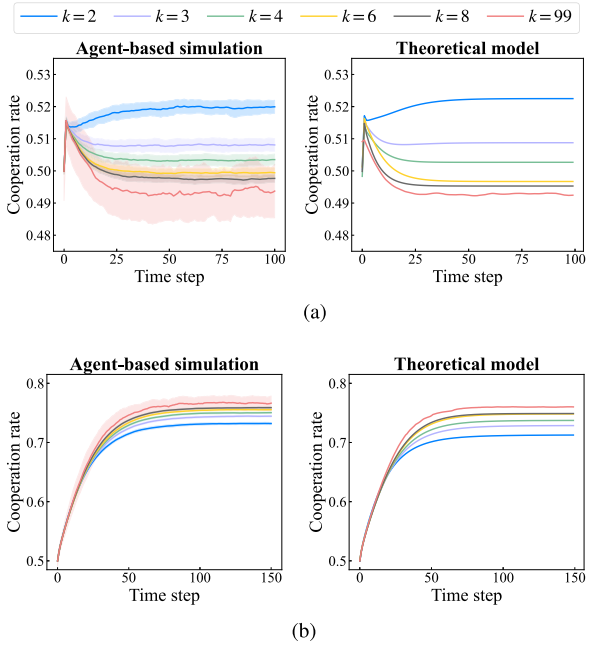



Fig. 6. Effect of population structure on agents' cooperative behavior under different transition rules. (a) State-independent transition rule. (b) State-dependent transition rule.

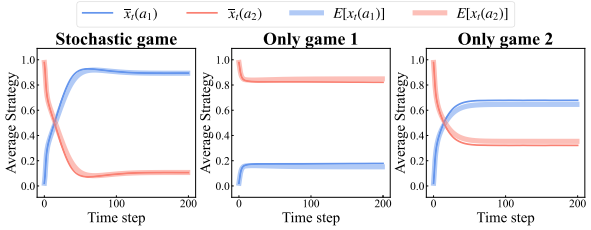



Fig. 7. Stochastic games are the catalysts for fostering cooperation in structured populations. In this experiment, we distribute the population over a four-regular graph and set $\alpha = 0.5$ and $\tau = 3$. This parameter setting differs from that used in previous experiments and is intended to demonstrate that the applicability of our model is not constrained by specific algorithmic parameters. Additionally, we initialize $Q_0(a_1) \sim \text{Beta}(10, 60, 0, 1.8)$ and $Q_0(a_2) \sim \text{Beta}(60, 10, 0, 1.8)$.

direction [21], [76], leading to the development of numerous effective mechanisms for promoting cooperation and the discovery of many intriguing phenomena. In stochastic games, understanding how state transition mechanisms influence the evolution of cooperation among state-agnostic $Q$-learners is of great importance and represents a significant contribution.

To demonstrate the efficacy of stochastic games, we conduct experiments on a two-state SH game, where state 1 is considered more valuable than state 2. In state 1, mutual cooperation yields a reward of 1.8 for each agent. If one agent cooperates while the other defects, the cooperator receives 0, while the defector always earns 0.8. In state 2, mutual cooperation results in a lower reward of 1 for each agent. As in state 1, unilateral cooperation yields 0 for the cooperator, while the defector receives a fixed reward of 0.4. State transitions follow a state-independent rule [see Fig. 4(b)], where mutual cooperation promotes transitions to the more profitable state, and defection leads to the less valuable one. This setup illustrates the "tragedy of the commons," highlighting how prosocial behavior can safeguard shared resources, whereas negative behaviors lead to environmental degradation and diminished benefits.

In Fig. 7, we visualize the strategy learning dynamics in the two-state SH game, comparing these outcomes with those from two associated normal-form games. Consequently, the strategy learning dynamics of three populations are presented: the first population engages in stochastic games with transitions between the two SH games, while the other two populations play only the more valuable SH game (game 1) and the less valuable SH game (game 2), respectively. The SH game features two Nash equilibria: the risk-dominant equilibrium $(D, D)$ and the efficient equilibrium $(C, C)$, which maximizes social welfare. Understanding how agents can avoid the inefficient equilibrium and coordinate towards the efficient outcome

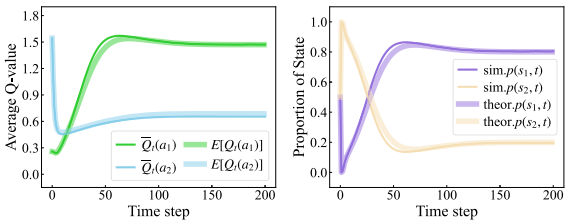



Fig. 8. Evolution of the average $Q$-values and the environmental state of the population in the two-state SH games.

is crucial across various fields. In our experiments, agents that always play the more valuable game 1 fail to reach the efficient equilibrium, achieving only an 18% cooperation rate. Conversely, the population that always plays the less valuable game 2 attains a cooperation rate of approximately 68%. Remarkably, when agents engage in stochastic games, the cooperation rate rises to nearly 90%, indicating that the state transition mechanism can positively influence the evolution of cooperation in social dilemmas.

Fig. 8 presents the dynamics of the average $Q$-values and the environmental state of the population in the two-state SH game. As shown, the initial distribution of $Q$-values in this experiment results in a much higher initial average $Q$-value for defection than for cooperation. This helps demonstrate that the higher cooperation rate observed under the state transition mechanism, compared with the two repeated normal-form games, is not due to a cooperation-favoring initial condition. By contrast, even under an initial condition that disfavors cooperation, agents can learn to cooperate, leading to more frequent interactions in the more valuable state.

Further experiments examining how algorithm parameters affect the learning dynamics are analyzed in detail in the Supplementary Material.

## VI. Conclusion

In this article, we propose a formal model for multiagent $Q$-learning dynamics in stochastic games on regular graphs. We use interaction configurations to characterize the fixed local interactions of agents, which determine the rewards driving their strategy learning processes. By leveraging the pair-approximation method, we derive state distributions and opponent strategies for agents with different $Q$-values and obtain probability distributions of interaction configurations for different agents. Considering the influence of these configurations on agents and the underlying environmental dynamics, we derive a partial differential equation to track the evolution of the pair distribution. To validate our model, we conduct experiments across various stochastic games, exploring different game configurations, state transition rules, population structures, and algorithm parameters. Our findings reveal that population structure can significantly affect the evolution of cooperation among state-agnostic $Q$-learners, depending on the transition rules. Notably, we demonstrate that under certain conditions, populations with state transitions converge to more prosocial outcomes that cannot be achieved in corresponding single normal-form games.

In future work, we will explore the applicability of our model to asymmetric games and alternative learning algorithms. Moreover, developing a dynamical model for agents that condition their decisions on environmental states presents a promising direction for further study, offering deeper insights into how agents' ability to perceive and utilize environmental information influences the evolution of cooperative behavior.

## References

[1] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3389–3396.

[2] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2021.

[3] N. Naderializadeh, J. J. Sydir, M. Simsek, and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 6, pp. 3507–3523, Jun. 2021.

[4] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas," in *Proc. 16th Conf. Auto. Agents MultiAgent Syst.*, 2017, pp. 464–473.

[5] J. Perolat, J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel, "A multi-agent reinforcement learning model of common-pool resource appropriation," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates, 2017, pp. 3646–3655.

[6] S. Zheng, A. Trott, S. Srinivasa, D. C. Parkes, and R. Socher, "The AI economist: Taxation policy design via two-level deep multiagent reinforcement learning," *Sci. Adv.*, vol. 8, no. 18, p. 2607, May 2022.

[7] M. Rios, N. Quijano, and L. F. Giraldo, "Understanding the world to solve social dilemmas using multi-agent reinforcement learning," 2023, *arXiv:2305.11358*.

[8] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[9] D. Silver et al., "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[10] N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker," *Science*, vol. 365, no. 6456, pp. 885–890, Aug. 2019.

[11] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, "Evolutionary dynamics of multi-agent learning: A survey," *J. Artif. Intell. Res.*, vol. 53, pp. 659–697, Aug. 2015.

[12] T. Börgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *J. Econ. Theory*, vol. 77, no. 1, pp. 1–14, Nov. 1997.

[13] Y. K. Cheung, "Multiplicative weights updates with constant step-size in graphical constant-sum games," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 3532–3542.

[14] V. Boone and G. Piliouras, "From Darwin to Poincaré and Von Neumann: Recurrence and cycles in evolutionary and algorithmic game theory," in *Proc. Int. Conf. Web Internet Econ.*, 2019, pp. 85–99.

[15] J. P. Bailey and G. Piliouras, "Multi-agent learning in network zero-sum games is a Hamiltonian system," in *Proc. 18th Int. Conf. Auto. Agents MultiAgent Syst.*, 2019, pp. 233–241.

[16] S. Leonardos, G. Piliouras, and K. Spendlove, "Exploration-exploitation in multi-agent competition: Convergence with bounded rationality," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, 2021, pp. 26318–26331, Art. no. 2015.

[17] L. Flokas, E. V. Vlatakis-Gkaragkounis, and G. Piliouras, "Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*. Red Hook, NY, USA: Curran Associates, 2019, pp. 10450–10461, Art. no. 938.

[18] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[19] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE Access*, vol. 7, pp. 133653–133667, 2019.

[20] J. Clifton and E. B. Laber, "Q-learning: Theory and applications," *Annu. Rev. Statist. Appl.*, vol. 7, no. 1, pp. 279–301, 2020.

[21] S. Fatima, N. R. Jennings, and M. Wooldridge, "Learning to resolve social dilemmas: A survey," *J. Artif. Intell. Res.*, vol. 79, pp. 895–969, Mar. 2024.

[22] K. Tuyls, K. Verbeeck, and T. Lenaerts, "A selection-mutation model for q-learning in multi-agent systems," in *Proc. 2nd Int. Joint Conf. Auto. Agents Multiagent Syst.*, Jul. 2003, pp. 693–700.

[23] K. Tuyls and S. Parsons, "What evolutionary game theory tells us about multiagent learning," *Artif. Intell.*, vol. 171, no. 7, pp. 406–416, May 2007.

[24] S. Hu, C.-W. Leung, and H.-F. Leung, "Modelling the dynamics of multiagent Q-learning in repeated symmetric games: A mean field theoretic approach," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 12102–12112.

[25] S. Hu, C.-W. Leung, H.-F. Leung, and H. Soh, "The evolutionary dynamics of independent learning agents in population games," 2020, *arXiv:2006.16068*.

[26] S. Hu, C.-W. Leung, H.-f. Leung, and H. Soh, "The dynamics of Q-learning in population games: A physics-inspired continuity equation model," in *Proc. 21st Int. Conf. Auto. Agents Multiagent Syst.*, 2022, pp. 615–623.

[27] G. Hardin, "The tragedy of the commons: The population problem has no technical solution; it requires a fundamental extension in morality," *Science*, vol. 162, no. 3859, pp. 1243–1248, Dec. 1968.

[28] W. Uddin Mondal, V. Aggarwal, and S. V. Ukkusuri, "On the near-optimality of local policies in large cooperative multi-agent reinforcement learning," 2022, *arXiv:2209.03491*.

[29] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 5567–5576.

[30] Q. Long, Z. Zhou, A. Gupta, F. Fang, Y. Wu, and X. Wang, "Evolutionary population curriculum for scaling multi-agent reinforcement learning," 2020, *arXiv:2003.10423*.

[31] J. Hao et al., "Exploration in deep reinforcement learning: From single-agent to multiagent domain," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 8762–8782, Jul. 2024.

[32] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci. USA*, vol. 39, no. 10, pp. 1095–1100, 1953.

[33] C. Hilbe, Š. Šimsa, K. Chatterjee, and M. A. Nowak, "Evolution of cooperation in stochastic games," *Nature*, vol. 559, no. 7713, pp. 246–249, Jul. 2018.

[34] F. Huang, M. Cao, and L. Wang, "Learning enables adaptation in cooperation for multi-player stochastic games," *J. Roy. Soc. Interface*, vol. 17, no. 172, Nov. 2020, Art. no. 20200639.

[35] D. Hennes, K. Tuyls, and M. Rauterberg, "State-coupled replicator dynamics," in *Proc. 8th Int. Conf. Auto. Agents Multiagent Syst.*, vol. 26, 2009, pp. 789–796.

[36] C. Chu, Z. Yuan, S. Hu, C. Mu, and Z. Wang, "A pair-approximation method for modelling the dynamics of multi-agent stochastic games," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 5, pp. 5565–5572.

[37] M. A. Nowak and R. M. May, "Evolutionary games and spatial chaos," *Nature*, vol. 359, no. 6398, pp. 826–829, May 1992.

[38] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, no. 5805, pp. 1560–1563, Dec. 2006.

[39] C. Chu, Y. Li, J. Liu, S. Hu, X. Li, and Z. Wang, "A formal model for multiagent Q-learning dynamics on regular graphs," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 194–200.

[40] W. U. Mondal, M. Agarwal, V. Aggarwal, and S. V. Ukkusuri, "On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC)," *J. Mach. Learn. Res.*, pp. 1–46, 2021.

[41] M. Kleshnina, C. Hilbe, Š. Šimsa, K. Chatterjee, and M. A. Nowak, "The effect of environmental information on evolution of cooperation in stochastic games," *Nature Commun.*, vol. 14, no. 1, p. 4153, Jul. 2023.

[42] J. G. Cross, "A stochastic learning model of economic behavior," *Quart. J. Econ.*, vol. 87, no. 2, pp. 239–266, May 1973.

[43] Y. Sato and J. P. Crutchfield, "Coupled replicator equations for the dynamics of learning in multiagent systems," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 67, no. 1, Jan. 2003, Art. no. 015206.

[44] E. Rodrigues Gomes and R. Kowalczyk, "Dynamic analysis of multi-agent Q-learning with $\mu$-greedy exploration," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 369–376.

[45] M. B. Wunder, M. L. Littman, and M. Babes, "Classes of multiagent Q-learning dynamics with epsilon-greedy exploration," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 1167–1174.

[46] F. Panozzo, N. Gatti, and M. Restelli, "Evolutionary dynamics of Q-learning over the sequence form," in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014, pp. 2034–2040.

[47] C. Deng, Z. Rong, L. Wang, and X. Wang, "Modeling replicator dynamics in stochastic games using Markov chain method," in *Proc. 20th Int. Conf. Auto. Agents MultiAgent Syst.*, 2021, pp. 420–428.

[48] M. Kaisers and K. Tuyls, "Frequency adjusted multi-agent Q-learning," in *Proc. 9th Int. Conf. Auto. Agents Multiagent Syst.*, 2010, pp. 309–316.

[49] D. Hennes, M. Kaisers, and K. Tuyls, "RESQ-learning in stochastic games," in *Proc. Adapt. Learn. Agents Workshop AAMAS*, 2010, p. 8.

[50] C.-W. Leung, S. Hu, and H.-F. Leung, "Modelling the dynamics of multi-agent Q-learning: The stochastic effects of local interaction and incomplete information," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 384–390.

[51] J. Liu, G. Jiang, C. Chu, Y. Li, Z. Wang, and S. Hu, "A formal model for multiagent Q-learning on graphs," *Sci. China Inf. Sci.*, vol. 68, no. 9, Sep. 2025, Art. no. 192206.

[52] J. Shi, C. Liu, and J. Liu, "Hypergraph-based model for modeling multi-agent Q-learning dynamics in public goods games," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 6, pp. 6169–6179, Nov. 2024.

[53] Z. Wang, C. Mu, S. Hu, C. Chu, and X. Li, "Modelling the dynamics of regret minimization in large agent populations: A master equation approach," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 534–540.

[54] S. Hu, H. Soh, and G. Piliouras, "The best of both worlds in network population games: Reaching consensus and convergence to equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 77149–77173.

[55] D. Hu et al., "Regret minimization in population network games: Vanishing heterogeneity and convergence to equilibria," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 12, pp. 20146–20156, Dec. 2025.

[56] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *J. Human Evol.*, vol. 22, no. 6, pp. 469–493, Jun. 1992.

[57] R. I. M. Dunbar, "Coevolution of neocortical size, group size and language in humans," *Behav. Brain Sci.*, vol. 16, no. 4, pp. 681–694, Dec. 1993.

[58] X. Li et al., "Punishment diminishes the benefits of network reciprocity in social dilemma experiments," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 1, pp. 30–35, Jan. 2018.

[59] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, "A simple rule for the evolution of cooperation on graphs and social networks," *Nature*, vol. 441, no. 7092, pp. 502–505, May 2006.

[60] H. Ohtsuki and M. A. Nowak, "Evolutionary stability on graphs," *J. Theor. Biol.*, vol. 251, no. 4, pp. 698–707, Apr. 2008.

[61] Q. Su, A. McAvoy, L. Wang, and M. A. Nowak, "Evolutionary dynamics with game transitions," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 51, pp. 25398–25404, 2019.

[62] Z. Sun, X. Chen, and A. Szolnoki, "State-dependent optimal incentive allocation protocols for cooperation in public goods games on regular networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 6, pp. 3975–3988, Jun. 2023.

[63] C. Wang, M. Perc, and A. Szolnoki, "Evolutionary dynamics of any multiplayer game on regular graphs," *Nature Commun.*, vol. 15, no. 1, p. 5349, Jun. 2024.

[64] Q. Wang, X. Chen, N. He, and A. Szolnoki, "Evolutionary dynamics of population games with an aspiration-based learning rule," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 8387–8400, May 2025.

[65] F. C. Santos, M. D. Santos, and J. M. Pacheco, "Social diversity promotes the emergence of cooperation in public goods games," *Nature*, vol. 454, no. 7201, pp. 213–216, Jul. 2008.

[66] M. Perc and A. Szolnoki, "Coevolutionary games—A mini review," *Biosystems*, vol. 99, no. 2, pp. 109–125, Feb. 2010.

[67] F. Battiston et al., "Networks beyond pairwise interactions: Structure and dynamics," *Phys. Rep.*, vol. 874, pp. 1–92, Aug. 2020.

[68] M. Jusup et al., "Social physics," *Phys. Rep.*, vol. 948, pp. 1–148, Feb. 2022.

[69] L. Wardil and J. K. L. da Silva, "Distinguishing the opponents promotes cooperation in well-mixed populations," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 81, no. 3, Mar. 2010, Art. no. 036115.

[70] D. Jia et al., "Social networking agency and prosociality are inextricably linked in economic games," *Nature Hum. Behav.*, vol. 9, no. 12, pp. 1–12, Sep. 2025.

[71] M. Abou Chakra, S. Bumann, H. Schenk, A. Oschlies, and A. Traulsen, "Immediate action is the best strategy when facing uncertain climate change," *Nature Commun.*, vol. 9, no. 1, p. 2566, Jul. 2018.

[72] G. Szabó and G. Fáth, "Evolutionary games on graphs," *Phys. Rep.*, vol. 446, nos. 4–6, pp. 97–216, Jul. 2007.

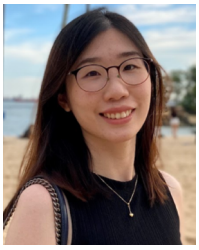
[73] C. P. Roca, J. A. Cuesta, and A. Sánchez, "Effect of spatial structure on the evolution of cooperation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 80, no. 4, Oct. 2009, Art. no. 046106.
[74] A. Szolnoki and M. Perc, "Conditional strategies and the evolution of cooperation in spatial public goods games," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 85, no. 2, Feb. 2012, Art. no. 026104.
[75] C. Hauert and G. Szabó, "Spontaneous symmetry breaking of cooperation between species," *PNAS Nexus*, vol. 3, no. 9, p. 326, Sep. 2024.
[76] C. Mu et al., "Multi-agent, human-agent and beyond: A survey on cooperation in social dilemmas," *Neurocomputing*, vol. 610, Dec. 2024, Art. no. 128514.

**Zheng Yuan** received the B.E. degree from South China University of Technology, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China.

His research interests include multiagent reinforcement learning, complex networks, and game theory.

**Guangchen Jiang** received the B.E. and M.E. degrees in software from Yunnan University, Kunming, China, in 2021 and 2024, respectively. He is currently pursuing the Ph.D. degree with the School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China.

His research interests include game theory and multiagent systems.



**Shuyue Hu** received the Ph.D. degree from Chinese University of Hong Kong, Hong Kong, China, in 2019.

She is currently a Researcher at Shanghai Artificial Intelligence Laboratory, Shanghai, China. Her research interests include multiagent systems, game theory, and large language models.

**Matjaž Perc** received the Ph.D. degree in physics from the University of Maribor, Maribor, Slovenia, in 2006.

He is currently a Professor of physics at the University of Maribor, a Staff Researcher at the Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, an Adjunct Professor at Kyung Hee University and Korea University, Seoul, South Korea and the External Faculty Member at the Complexity Science Hub, Vienna, Austria.

Dr. Perc is a member of Academia Europaea and European Academy of Sciences and Arts, and among the top 1% most cited physicists according to 2020, 2021, 2022, 2023, and 2024 Clarivate Analytics data. In 2019, he became a Fellow of the American Physical Society. He is also the 2015 recipient of the Young Scientist Award for Socio and Econophysics from the German Physical Society, and the 2017 USERN Laureate. In 2018, he received the Zois Award, which is the Highest National Research Award in Slovenia. Since 2021, he has been the Vice-Dean of Natural Sciences at the European Academy of Sciences and Arts.

**Chen Chu** received the Ph.D. degree from Dongbei University of Finance and Economics, Dalian, China, in 2015.

He is currently a Professor with Yunnan University of Finance and Economics, Kunming, China. His research interests include artificial intelligence, network science, and multiagent learning in games.

**Jinzhuo Liu** received the Ph.D. degree from Yunnan University, Kunming, China, in 2013.

She is currently an Associate Professor with Yunnan University. Her research interests include artificial intelligence, game theory, and reinforcement learning.