SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2025, Vol. 68, Iss. 9, 192206:1–192206:14
https://doi.org/10.1007/s11432-024-4289-6

A formal model for multiagent Q-learning on graphs

Jinzhuo LIU[1,2], Guangchen JIANG[1], Chen CHU[2,3,4], Yong LI[5], Zhen WANG[2,6*] & Shuyue HU[7*]

[1]*School of Software, Yunnan University, Kunming 650504, China*
[2]*School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China*
[3]*Department of Statistics, School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, China*
[4]*Yunnan Key Laboratory of Service Computing, Yunnan University of Finance and Economics, Kunming 650221, China*
[5]*School of Information Science & Engineering, Yunnan University, Kunming 650500, China*
[6]*School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China*
[7]*Shanghai Artificial Intelligence Laboratory, Shanghai 201206, China*

Received 4 March 2024/Revised 29 June 2024/Accepted 14 August 2024/Published online 16 June 2025

Abstract Understanding the dynamics of multi-agent learning has long been an important research topic. Existing research has focused mostly on 2-agent games or well-mixed populations. However, in real-world multi-agent systems, agents often interact in spatially or socially structured networks (or graphs). In this paper, we examine the dynamics of multi-agent Q-learning on graphs. Combining mean-field theory and combinatorics analysis, we present a new analytical approach to formally describe the time evolution of Q-values in the system with a topological structure. Through extensive numerical simulations, we show that our theory consistently provides an accurate depiction of the Q-learning dynamics across different typical games, initial conditions, and various graph structures, encompassing regular graphs, scale-free graphs, and random graphs. Moreover, we show that when comparing regular graphs to other types of graphs with the same average degree, the differences in the system evolution are largely attributed to the behaviors and Q-values of agents with lower degrees.

Keywords multiagent, Q-learning dynamics, game theory, graph theory

Citation Liu J Z, Jiang G C, Chu C, et al. A formal model for multiagent Q-learning on graphs. Sci China Inf Sci, 2025, 68(9): 192206, https://doi.org/10.1007/s11432-024-4289-6

1 Introduction

Recent years have witnessed a substantial advancement in the capability of learning agents, leading to their wide applications in a variety of multi-agent systems (MASs), such as traffic systems [1], swarm robots [2–4], and wireless networks [5, 6]. However, in contrast to the solid theoretical foundation of learning under single-agent settings [7–10], learning under multi-agent settings (or multi-agent learning) is still far from being well understood [11–16].

In pioneering work, Tuyls et al. [17] and Sato and Crutchfield [18] showed that Q-learning dynamics [19] in 2-player repeated normal-form games can be viewed as a selection-mutation mechanism in the evolutionary game theory [20]. This finding connected arguably the most well-known reinforcement learning (RL) model to evolutionary game theory, and consequently has inspired many subsequent studies to investigate multi-agent learning using evolutionary game-theoretic (EGT) approaches [21–28]. Kianercy and Galstyan [29] characterized the fixed point structure of Q-learning in diverse 2-player games. Kaisers and Tuyls [30] observed that the EGT approach may predict outcomes that deviate from the actual Q-learning dynamics and consequently proposed a new algorithm that better aligned with the EGT prediction. Leonardo et al. [31] showed that by tuning the exploration parameter, there are phase transitions between multiple equilibria for Q-learning in certain n-player games. Hu et al. [32] focused on a well-mixed population and showed that the multi-agent Q-learning dynamics can be characterized by a Fokker-Planck equation.

* Corresponding author (email: w-zhen@nwpu.edu.cn, hushuyue@pjlab.org.cn)

Previous work on multi-agent learning dynamics (see [22] for a comprehensive review) has focused either on finite-player (mostly 2-player) games or well-mixed populations. However, different from the conventional settings of 2-player games and well-mixed populations, real-world MASs can exhibit topological structures—agents interact in spatially or socially structured networks (or graphs) [33–37]. For example, within a social network, individuals tend to communicate primarily with the friends they know rather than random strangers. Similarly, in traffic networks, vehicles interact based on their spatial proximity rather than uniformly with all cars in the system. Representing each individual (or agent) as a vertex and each potential interaction between two individuals as an edge, a multiagent system (MAS) with a topological structure can be formally represented by a graph. However, despite its widespread applicability, multi-agent learning on graphs has received limited attention in existing theoretical investigations.

Motivated by this gap, we study multi-agent learning on graphs in this work with a particular focus on Q-learning [19]—arguably the most well-known learning model. Previous studies, such as Hu et al. [32] have utilized mean-field theory (MFT) to model interactions in large, well-mixed populations, and approximate the effects of other agents on the single agent by an averaged effect. This approach was further developed by Leung et al. [38], who applied it to Q-learning within social learning models. While effective for infinite and well-mixed populations, these models often overlook the local impact among agents in graph-structured populations, where agents primarily interact with nearby neighbors.

Multi-agent learning on graphs typically features a large number of agents as well as a potentially unequal degree distribution (e.g., the degree distribution exhibits a scale-free property) [39, 40]. While the former naturally causes analytical techniques for learning in 2-player games infeasible [17, 18, 22], the latter breaks the mean-field assumption, which assumes each individual can be treated equally as the same mean-field, used in the previous analyses for learning in well-mixed populations [32, 38]. Therefore, multi-agent learning on graphs presents a new challenge: how can one effectively capture the correlation between the strategies of many agents and the degree distribution of the underlying graph structure?

To address this challenge, we present a new analytical approach that combines the MFT and combinatorics analysis. This entails (i) encoding possible joint actions of an agent's neighbors into a neighbor configuration (Definition 2), (ii) reformulating the agent's reward function based on the neighbor configuration (Lemma 1), (iii) employing a mean-field approximation to estimate the probability of each neighbor configuration that the agent may encounter (Lemma 2), and (iv) deriving the expected change in Q-values for the agent (Lemma 4). Based on these key steps, we show that our approach leads to a Fokker-Planck equation that governs the probability distribution of Q-values on graphs (Theorem 1). Moreover, we show that as the degree of each vertex (or agent) approaches infinity, our derived Fokker-Planck equation can be reduced to the Fokker-Planck equation for well-mixed populations (Corollary 1); that is to say, our theory, which considers multi-agent learning on graphs, can be seen as a generalization of the previous approach [32] that assumes well-mixed populations and overlooks the possibility of underlying topology of MASs.

We numerically validate our theory on five distinct symmetric games and various graph structures, encompassing regular graphs with diverse degrees, scale-free graphs, as well as two categories of random graphs in our experiments. We show that our theory consistently provides an accurate depiction of Q-learning dynamics across different games, initial conditions, and graph structures. In particular, we show that our model yields a more accurate prediction of learning on regular and random graphs than on scale-free graphs. We analyze that this is because, within scale-free graphs, many substructures hug and resist interactions with others. Moreover, we show that in regular graphs, as agents' (or nodes') degrees increase, their behavior and Q-values converge to that of the mean-field (i.e., the averaged behavior and the averaged Q-values). Last but not least, we show that when comparing regular graphs to other types of graphs with the same average degree, the differences in the system evolution are largely attributed to the behaviors and Q-values of agents with lower degrees.

2 Related work

2.1 Modeling the dynamics of Q-learning in games

Tuyls et al. [17] developed a system of replicator equations to model the policy dynamics of two Q-learning agents in repeated 2-player matrix games. The importance of replicator dynamics in RL algorithms, such as Q-learning, is evident as it is linked to the gradient of expected rewards [41]. Subsequently, some studies

extended this approach beyond 2-player games to analyze the dynamics of a small-sized population of Q-learning agents [42]. Kianercy and Galstyan's study [43] presents a comprehensive portrayal of the fixed point structure across multiple 2-player 2-action games and investigates how this structure responds to noise resulting from exploration. Subsequent studies generalized this EGT approach to other learning algorithms, such as FAQ-learning, infinitesimal gradient ascent, and regret minimization. Researchers developed formal models for the dynamics of two agents under these learning algorithms [30,44]. Recently, Hu et al. [32,45] developed a Fokker-Planck equation-based model to characterize the population dynamics of infinitely many Q-learning agents in an MAS. Building upon this work, Leung et al. [38] examined local interactions between agents, in which an agent is paired with another agent in the population to play a 2-player symmetric game at each time step. However, the underlying topological graphs of the MASs considered in these studies [32, 38, 45, 46] essentially consist of well-mixed populations, where in each round of gameplay, every agent interacts with another agent randomly selected from the rest of the population.

2.2 Simulation of Q-learning on graphs

Additionally, some researchers have focused on exploring the impact of graph structure on the interaction behavior between agents by carrying out simulations on MASs with a specific underlying topological graph. Villatoro et al. [47, 48] investigated the impact of different graphs on agents' behavior in a Q-learning MAS, where agents' rewards depend on the past actions of others and studied the effects of certain substructures on the diffusion of agent behavior. They pointed out that different features of the graphs where agents are located can result in different convergence times to reach a social convention. These experiments have revealed evidence that demonstrates how certain abstract substructures within certain topological graphs, such as a scale-free graph, can result in longer convergence times or even prevent full convergence to society-wide conventions. If agents can interact with a majority of the population, it can prevent the formation of local conventions. Sen et al. [49] examined the effect of graph structure, such as scale-free and ring networks. In scale-free networks, hub nodes (agents), agents with numerous neighbors, have more influence than others. While ring (regular) networks converge faster for fewer actions, scale-free networks can converge faster for a larger number of actions. Yu et al. [50] allowed each agent to consider the actions of all its neighbors and investigated the effects of underlying graphs. Nevertheless, these studies are limited to simulations and lack a proper theory. This work complements these simulation-based studies by providing a feasible approach to theoretically study the dynamics of multi-agent learning on graphs.

3 Preliminaries

3.1 Multiagent systems on the graph

Consider an MAS on the graph. The interactions between agents are subject to certain limitations, which we abstract as a graph $\mathcal{G}$ that satisfies Definition 1. Each vertex represents an agent that can interact with another one with the 2-player game $\mathtt{G}$ via an edge e_{ij} in graph $\mathcal{G}$ and get the reward r_t^i averaged over all games with its neighbors.

Definition 1 (Graph)**.** A graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ consist of a set $\mathcal{N}$ of vertices (agents) and a set $\mathcal{E}$ of edges in which each edge e_{ij} connects 2 vertices i, j. The graph we considered in this paper is an undirected connected graph without any loops, whose degree distribution is given by a function $\rho(\cdot)$.

We use an adjacency matrix $\boldsymbol{B} := [b_{ij}]_{n\times n}$ to represent the agents' interaction topological graph, while the edge set $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ governs the interactions between two agents. For the adjacency matrix $\boldsymbol{B}$, all diagonal elements b_{ii} are zero, while off-diagonal elements satisfy $b_{ij} = 1\,(i \neq j)$ if and only if $e_{ij} \in \mathcal{E}$; otherwise, $b_{ij} = 0$. Both the edge $e_{ij}\,(i \neq j)$ and $e_{ji}\,(i \neq j)$ indicate the same 2-player game, means that agent i, i.e., vertex i, can interact with its neighbor j.

The behavior of an agent interacting with its neighbors leads naturally to the concepts of agent i's neighbor set and degree. In this paper, the neighbor set $\mathcal{N}_i := \{j : e_{ij} \in \mathcal{E}, i \neq j\}$ is the set of agent i's neighbors, and the matrix $\boldsymbol{K} := \operatorname{diag}\left\{k^1, \ldots, k^n\right\}$ is a diagonal matrix, where $k^i = \sum_{j=1}^n b_{ij} = |\mathcal{N}_i|$ is the agent i's degree, represents the degree of all agents in the population. Also, we use a discrete probability function $\rho(\cdot)$, where $\rho(\text{degree} = k) \geqslant 0$ and $\sum_k \rho(\text{degree} = k) = 1$, to represent the degree distribution of individuals in the whole population. For the ease of notation, we simplify $\rho(\text{degree} = k)$ as $\rho(k)$.

Algorithm 1 A framework for multiagent Q-learning on a graph.

Require: A graph-structured agent population $\mathcal{G}(\mathcal{N}, \mathcal{E})$, a symmetric normal-form game $\mathtt{G}$, a maximum time step T.

while $t < T$ **do**
  for each agent $i \in \mathcal{N}$ **do**
    Agent i selects an action $a \in \mathcal{A}$ in accordance with its strategy;
  end for
  for each agent $i \in \mathcal{N}$ **do**
    for each neighbor $z \in \mathcal{N}_i$ **do**
      Agents i and z play game $\mathtt{G}$ using their respective selected actions;
    end for
    Agent i receives a payoff $r_t^i(a)$ that is averaged over the k^i games it plays;
  end for
  for each agent $i \in \mathcal{N}$ **do**
    Agent i updates the Q-value of its corresponding action a:
    $q_{t+1}^i(a) \leftarrow q_t^i(a) + \alpha[r_t^i(a) - q_t^i(a)]$;
  end for
  $t \leftarrow t + 1$;
end while

3.2 Interbehavior betweens agents

For an MAS on a graph that satisfies Definition 1, agents obtain rewards by playing the 2-player symmetric game.

The normal-form game denoted by $\mathtt{G} := \langle \mathcal{A}, \boldsymbol{R} \rangle$ in which $\mathcal{A} := \{a_1, \ldots, a_m\}$ is the available action set, $\boldsymbol{R} \in \mathbb{R}^{m \times m}$ is a reward matrix that is symmetric for the two players:

$$\boldsymbol{R} := \left[r_{a_i a_j} \right]_{1 \leqslant i,j \leqslant m},$$

where $r_{a_i a_j}$ denotes the reward when an agent taking action a_i plays the game against a neighbor taking action a_j via an existing edge e_{ij}. This action can be encoded into an one-hot vector $\boldsymbol{a}_j = [0, \ldots, 1, \ldots, 0]^{\mathrm{T}}$, i.e., the unit vector with the j-th component equaling to 1 and others 0. In every time step t, the agent i receives a reward $r_j^i(a_j)$ according to the actions it and its neighbors take. For one game $\mathtt{G}$, the payoff of using action a_j against action a_y is

$$r(a_j \mid a_y) = \boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{R} \boldsymbol{a}_y. \tag{1}$$

We denote the joint action of neighbors of agent i as $\boldsymbol{a}^{-i}$. For the agent with more than one neighbor, an arbitrary agent's reward is averaged over all k^i games with its k^i neighbors at time step t, i.e.,

$$r_t^i\left(a_j \mid \boldsymbol{a}^{-i}\right) = \frac{1}{k^i} \sum_{\forall z^i \in \mathcal{N}_i, a_y \in \mathcal{A}} r_t^i\left(a_j \mid a_y^{z^i}\right), \tag{2}$$

where $a_y^{z^i}$ is the action a_y which chosen by agent i's neighbor z^i.

Since the game $\mathtt{G}$ is symmetric, the reward for an arbitrary agent in $\mathcal{G}$ playing a particular action depends only on the other agent's action, not the identity of the opponent; therefore, the game $\mathtt{G}$ is the same for all agents in $\mathcal{G}$. The agent can use mixed strategies, where the strategy of agent i at time step t is $\boldsymbol{x}_t^i := \left[x_t^i(a_1), \ldots, x_t^i(a_m)\right]^{\mathrm{T}}, \boldsymbol{x}_t^i \in \mathbb{R}^m, x_t^i \geqslant 0, \sum_j^m x_t^i(a_j) = 1$, where $x_t^i(a_j)$, $\forall a_j \in \mathcal{A}$ denotes the probability of agent i choosing action a_j (pure strategy) at time t.

Overall, we assume that for every time step t, each agent i chooses an action a_j in accordance with its own strategy $\boldsymbol{x}_t^i$ and plays game $\mathtt{G}$ with its k_i neighbors. Since normal-form games involve only a single state and do not require state transitions, we followed the previous studies [32, 45] and adopted stateless Q-learning. Based on the average payoff $r_t^i(a_j)$, the agent's strategy is revised by means of the stateless Q-learning method. The interaction framework is summarized in Algorithm 1.

3.3 *Q*-learning and Boltzmann explorations

In this subsection, we will detail the Q-learning and Boltzmann exploration used in this paper.

3.3.1 *Q-learning*

As mentioned above, since the game between agents is a repeated and normal-form game, no state transition occurs. There, the Q-learning framework under which the agents learn their actions is the

stateless version [51], meaning that each agent i possesses an individual vector of estimated Q-values $\boldsymbol{q}_t^i := [q_t^i(a_1), \ldots, q_t^i(a_m)]^{\mathrm{T}}$.

An agent needs to consider only its own current Q-value in each round without considering its history.

To evaluate action a_j and maximize the reward of future decisions, the corresponding Q-value $q_t^i(a_j)$ is updated as follows after action a_j is executed:

$$q_{t+1}^i(a_j) \leftarrow q_t^i(a_j) + \alpha \left[r_t^i(a_j) - q_t^i(a_j) \right], \tag{3}$$

where α is the learning rate.

3.3.2 *Boltzmann exploration*

Boltzmann exploration is a typical strategy for decision-making when faced with uncertainty and is one of the most commonly used tools in RL [52]. To address the fundamental dilemma of the exploration-exploitation trade-off, we use a Boltzmann probability with a certain parameter to model each agent's strategy: $x_t^i(a_j) := \exp\left(\beta q_t^i(a_j)\right) / \sum_{\forall a \in \mathcal{A}} \exp\left(\beta q_t^i(a)\right)$, where β is the Boltzmann exploration temperature. When β is relatively small, the agent tends to be more exploratory and will try some low-reward behaviors. If $\beta = 0$, this is equivalent to complete exploration (that is, the action will be selected completely at random). When β is relatively large, the agent tends to exploit its experience. If $\beta \to \infty$, the agent will become greedy (that is, it will tend to choose only the action that maximizes $q_t^i(a)$).

4 Modeling the *Q*-learning dynamics on a graph

In this section, we formally model the Q-learning dynamics on a graph. We start by discussing the limitations of the previous approach, focusing on well-mixed populations in Subsection 4.1. In Subsection 4.2, we present the key steps towards developing our main result—a continuity equation for describing the Q-learning dynamics on a graph (Theorem 1). Then, we discuss some important observations on our main result.

4.1 Limitations of the previous approaches

MFT is an effective approach of evolutionary game theory to transform the intractable many-body problem in a high-dimensional stochastic model into an effective one-body problem in a simpler model, thus obtaining a description of the system behavior at a low computational cost.

In previous work, Hu et al. [32] considered the case of an infinitely large, well-mixed population and used MFT to approximate the effects of other agents on the single agent by an averaged effect. Specifically, they approximated the immediate reward of an agent by the expected reward against the average strategy of the entire population. Then, they employed a probability distribution to depict the Q-value distribution of an MAS and established a continuity equation to characterize the evolutionary dynamics of that Q-value distribution. This idea has been later applied to Q-learning under the social learning model; Leung et al. [38] proposed the use of Brownian motions to model the stochasticity that arises from social learning and developed a Fokker-Planck equation (a variant of the continuity equation with diffusion terms) to formally describe the learning dynamics.

The key idea of using MFT for modeling Q-learning dynamics is that for different agents, as they use the same action, they change when facing the same situation. So, for an arbitrary agent in the well-mixed population, the environment it confronts, i.e., the actions chosen by its neighbors, is highly similar to that faced by any other in the population. However, for the population on a graph, limited by the graph structure $\mathcal{G}(\mathcal{N}, \mathcal{E})$, agents cannot interact with all others in the population, but only with local neighbors. So, for an arbitrary agent in the population, the strategies adopted by its neighbors are not the same or may be drastically different. In general, these previous studies only consider infinite and well-mixed populations and give relatively less consideration to the local impact among agents. In other words, such a situation [32] does not hold for all MASs on graphs.

4.2 Our mean-field theoretic approach

Let us start with an arbitrary agent i in the population with a degree of k^i, with a fixed set of neighbors $\mathcal{N}_i$. The individual's payoff depends on its action and the actions of all its neighbors, i.e., the configuration

of the neighbors. Before developing our dynamic model, we define the neighbor configuration and rewrite the payoff.

Definition 2 (Neighbor configuration)**.** Let $k^i_{a_j}$ be the number of agent i's neighbors who adopt action $a_j \in \mathcal{A}$. A neighbor configuration of agent i is $\gamma^i_t = [k^i_{a_1,t}, \ldots, k^i_{a_m,t}]^{\mathrm{T}}$ such that $k^i_{a_j,t} \geqslant 0$, $j = 1, \ldots, m$ and $\sum_{j=1}^m k^i_{a_j,t} = k^i$. The set of all possible neighbor configurations for agent i is denoted by Γ^i.

For example, suppose that agent i chooses between 2 actions (a_1 and a_2) and has 2 neighbors ($k^i = 2$). The set Γ^i consists of three possible neighbor configurations, namely $[2, 0]^{\mathrm{T}}$ (both the neighbors choose a_1), $[1, 1]^{\mathrm{T}}$ (one neighbor chooses a_1 and the other chooses a_2), and $[0, 2]^{\mathrm{T}}$ (both the neighbors choose a_2). Please note that in the scenario of $[1, 1]^{\mathrm{T}}$, which agents take action a_1 does not change the rewards of the focal agent. We can rewrite the agent's reward and obtain the following lemma by utilizing the above neighbor configuration definition and building upon (2).

Lemma 1. The reward of agent i who adopts action $a_j \in \mathcal{A}$ against its k^i neighbors can be expressed as

$$r_t\left(a_j \mid \boldsymbol{a}^{-i}\right) = r_t\left(a_j \mid \gamma^i_t\right) = \frac{1}{k^i} \boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{R} \gamma^i_t. \tag{4}$$

Proof. See Appendix A in Supporting information.

This lemma shows that the reward of agent i at time t depends on its own action a_j as well as its current neighbor configuration γ^i_t. As individual agents choose their actions independently and have their own neighborhoods, the current neighbor configurations generally differ across different agents. Using MFT, the likelihood of each neighbor configuration, denoted by $\lambda(\gamma^i_t)$, can be calculated based on the average strategy of the population. We show this in the following lemma.

Lemma 2. With mean-field approximation, the probability of the occurrence of a neighbor configuration $\gamma^i_t \in \Gamma^i$ for agent i is

$$\lambda\left(\gamma^i_t\right) = \frac{k^i!}{\prod_{l=1}^m k^i_{a_l}!} \prod_{j=1}^m \overline{x}_t(a_j)^{k^i_{a_j}}, \tag{5}$$

where $\overline{x}_t\left(a_j\right) = \frac{1}{n} \sum_{\forall z \in \mathcal{N}} \frac{\exp\left(\beta q^z_t(a_j)\right)}{\sum_{\forall a \in \mathcal{A}} \exp\left(\beta q^z_t(a)\right)}$ is the probability of all agents in the population selecting action a_j and the average strategy of the population is $\overline{\boldsymbol{x}}_t = [\overline{x}_t\left(a_1\right), \ldots, \overline{x}_t\left(a_m\right)]$.

Proof. See Appendix B in Supporting information.

Recall that upon receiving a reward, agent i updates its Q-value for the action in use and does not update the Q-values for the other actions. The action choice is made according to the mixed strategy and is thus stochastic. From the modeler's perspective, the action choices of individual agents are generally intractable given the infinitely large number of agents. Here, we track the expected change of Q-values for individual agents, factoring into the stochasticity that arises from their action choices. Let $f^i_j\left(\boldsymbol{q}^i_t\right)$ be the expected change in the Q-value of action a_j for agent i at time t, given the current Q-values $\boldsymbol{q}^i_t$. We derive its form in Lemma 3.

Lemma 3. The expected change in the Q-value of action $a_j \in \mathcal{A}$ for agent i at time t is given by

$$f^i_j\left(\boldsymbol{q}^i_t\right) = \alpha x^i_t\left(a_j\right) \left[\sum_{\gamma^i_t \in \Gamma^i} \left[\lambda\left(\gamma^i_t\right) r_t\left(a_j \mid \gamma^i_t\right)\right] - q^i_t(a_j) \right]. \tag{6}$$

Proof. See Appendix C in Supporting information.

Observe that the expected change $v^i_j\left(\boldsymbol{q}^i_t\right)$ in the Q-value depends on agent i and particularly agent i's current neighbor configurations γ^i_t. Note that by Definition 2, their sets of all possible neighbor configurations are the same for agents with the same degree. That is, $\Gamma^i = \Gamma^{z^i}$ if the degrees $k^i = k^{z^i}$ for agents i and z^i. Thus, for individual agents with degree k, we can write their set of all possible neighbor configurations as $\Gamma\left(k\right)$. Leveraging on this, we show that for every agent in the population, the expected changes in the Q-value of action $a_j \in \mathcal{A}$ can be unified into the same form $v_j(\boldsymbol{q}_t)$, regardless of the current neighbor configurations of individual agents.

Lemma 4. The expected change in the Q-value of action $a_j \in \mathcal{A}$ for any agent at time t is given by

$$v_j(\boldsymbol{q}_t) = \alpha x_t\left(a_j\right) \left[\sum_{k=1}^{\tilde{k}} \sum_{\gamma \in \Gamma(k)} \left[\rho(k) \lambda\left(\gamma\right) r_t\left(a_j \mid \gamma\right)\right] - q_t(a_j) \right], \tag{7}$$

where $\rho(k)$ is the degree distribution of the graph $\mathcal{G}$, $\tilde{k}$ is the maximum degree of the graph $\mathcal{G}$, and $\Gamma(k)$ is the set of all possible neighbor configurations given degree k with generic neighbor configuration γ.

Proof. See Appendix D in Supporting information.

By Lemma 4, at every time step t, the expected changes in the Q-values for any agent in the population admits the same form $v_j(\boldsymbol{q}_t)$. This allows us to represent the system state by a probability distribution over the set of Q-values $\boldsymbol{q}_t \in \mathbb{R}^m$, and track the system evolution by tracking the time evolution of the probability distribution of Q-values. We define $\boldsymbol{Q}_t = [Q_{1,t}, Q_{2,t}, \ldots, Q_{m,t}]^{\mathrm{T}} \in \mathbb{R}^m$ to be a vector of random variables denoting the Q-values of a randomly selected agent in the population at time t, and with a slight abuse of notation, $p(\boldsymbol{q}_t, t)$ to be the probability density function. Intuitively, $p(\boldsymbol{q}_t, t)$ can be understood as the fraction of agents having the Q-value $\boldsymbol{q}_t$ in the population at time t. We are now in a good position to present our main result about the time evolution of the system state.

Theorem 1. In the continuous time limit, the probability distribution of the Q-values in the population is governed by the following Fokker-Planck equation:

$$\frac{\partial p(\boldsymbol{q}_t, t)}{\partial t} = -\sum_{j=1}^{m} \left[\frac{\partial}{\partial q_t(a_j)} \Big[p(\boldsymbol{q}_t, t)\, v_j(\boldsymbol{q}_t) \Big] \right], \tag{8}$$

where $v_j(\boldsymbol{q}_t) = \alpha x_t(a_j)[\sum_{k=1}^{\tilde{k}} \rho(k) \sum_{\gamma \in \Gamma(k)} [[\frac{(k-1)!}{\prod_{l=1}^{m} k_{a_l}!} \prod_{l=1}^{m} \overline{x}_t(a_l)^{k_{a_l}}] \boldsymbol{a}_j^{\mathrm{T}} \boldsymbol{R} \boldsymbol{\gamma}] - q_t(a_j)]$, $\tilde{k}$ is the maximum degree of agents in a population, and $\overline{x}_t(a_l) = \int \frac{\exp(\beta q_t(a_l))}{\sum_{\forall a \in \mathcal{A}} \exp(\beta q_t(a))}\, p(\boldsymbol{q}_t, t) \mathrm{d}\boldsymbol{q}_t$.

Proof sketch. We are interested in the time evolution of the probability density function $p(\boldsymbol{q}_t, t)$. Let $\delta \in (0, 1]$ be the amount of time that passes between two repetitions of the games. We define $\theta(\boldsymbol{Q}_t)$ to be a test function of Q-values, and define a quantity Y_t to be the expected change of the test function during time δ, i.e., $Y_t = \frac{\mathbb{E}[\theta(\boldsymbol{Q}_{t+\delta})] - \mathbb{E}[\theta(\boldsymbol{Q}_t)]}{\delta}$.

We show that with the Taylor series, θ can be expanded as

$$\mathbb{E}\left[\theta\left(\boldsymbol{Q}_{t+\delta}\right)\right] = \mathbb{E}\left[\theta\left(\boldsymbol{Q}_t\right)\right] + \delta \mathbb{E}\left[\sum_{j=1}^{m} v_j(\boldsymbol{q}_t)\right] + \frac{1}{2}\delta^2 \mathbb{E}\left[\sum_{j=1}^{m} v_j(\boldsymbol{q}_t)^2 \frac{\partial^2 \theta(\boldsymbol{q}_t)}{\partial q_t(a_j)^2}\right] + \delta^2 \mathbb{E}\left[o\left(\sum_{j=1}^{m} v_j(\boldsymbol{q}_t)^2\right)\right]. \tag{9}$$

Observe that we can recover the quantity Y_t by moving the first term of the right-hand side to the left-hand side and dividing both sides by δ. Then, taking the limit of Y with $\delta \to 0$ (assuming the continuous-time limit), the contributions of the third and fourth terms on the right-hand side of (9) become negligible. After that, we use integration by parts and leverage the property that $p(\boldsymbol{q}_t, t)$ approaches 0 in the limit, yielding

$$\int \theta(\boldsymbol{q}_t) \frac{\partial p(\boldsymbol{q}_t, t)}{\partial t} \mathrm{d}\boldsymbol{q}_t = -\int \theta(\boldsymbol{q}_t) \sum_{j=1}^{m} \left[\frac{\partial}{\partial q_t(a_j)} \Big[p(\boldsymbol{q}_t, t)\, v_j(\boldsymbol{q}_t) \Big] \right] \mathrm{d}\boldsymbol{q}_t.$$

As this holds for any test function $\theta(\boldsymbol{q}_t)$, we obtain (8) in the theorem.

Then, We conduct further research on Theorem 1 to explore more fundamental reasons and get Corollary 1.

Corollary 1. Consider a population of agents on a graph and the agents' degree $k \to \infty$. Let us assume a population with all agents with the degree of k, and the distribution of Q-values, $p(\boldsymbol{q}_t, t, k)$, is a function of k. The Q-value distribution $p(\boldsymbol{q}_t, t, k)$ obtained by this paper converge to the one $\hat{p}(\boldsymbol{q}_t, t)$ obtained by the mean-filed approach [32,45], i.e., $p(\boldsymbol{q}_t, t, k) \xrightarrow[\mathcal{D}]{k \to \infty} \hat{p}(\boldsymbol{q}_t, t)$, where $\mathcal{D}$ means converging in distribution and $p(\boldsymbol{q}_t, t, k) = \int \frac{\partial p(\boldsymbol{q}_t, t, k)}{\partial t}\, \mathrm{d}t$.

Proof. See Appendix E in Supporting information.

In Corollary 1, we know that agents with different degrees and the same initial distribution have different Q-value distributions after some steps. Agents' Q-value distributions are not the same for regular graphs with different degrees. As the degree increases, the Q-value distribution converges to the mean-field result [32,45]. Agents with the same degree in a random graph behave similarly to the same degree agents in the regular graph regarding their actions. In this scenario, the Q-value of agents with larger degrees is also close to the mean-field result. The Q-value distribution of high-degree agents is very similar to the mean-field result.

	a_1	a_2
a_1	2, 2	0, 0
a_2	0, 0	4, 4

(a)

	a_1	a_2
a_1	1, -1	-1, 1
a_2	-1, 1	1, -1

(b)

	a_1	a_2
a_1	3, 3	0, 5
a_2	5, 0	1, 1

(c)

	a_1	a_2
a_1	3, 3	2, 3.5
a_2	3.5, 2	1, 1

(d)

	a_1	a_2
a_1	2, 2	3, -5
a_2	-5, 3	4, 4

(e)

Figure 1 Payoff matrices for (a) CG, (b) MP, (c) PD, (d) SD, and (e) SH.

5 Experimental validation

We consider 5 widely used symmetric normal-form games [53]—coordination game (CG), matching pennies (MP), prisoner's dilemma (PD), snowdrift game (SD, also known as Chicken), and stag hunt (SH)—to verify the effectiveness of our model in different settings. The action sets and payoff matrices for these five games are shown in Figure 1.

For the agent-based simulations and the dynamics model, throughout the entire article, the exploration temperature β is set to 2, and the learning rate α is set to 0.4. To avoid finite-size effects and randomness, we average several independent samples for each setting. Unless otherwise specified, we set the initial Q-values to a beta distribution between all agents.

5.1 The evolution of agent behaviors on regular graphs

We conduct simulations for the regular graph case based on the following interaction structure: a random regular graph with all nodes of degree k, no self-loops, and multiple edges. We also explore the accuracy of our method under three different initial Q-value distributions, as well as the evolution of these games. Three different beta distributions are as $\mathrm{Beta}\left(2,8,Q_{\min},Q_{\max}\right)$, $\mathrm{Beta}\left(5,5,Q_{\min},Q_{\max}\right)$, and $\mathrm{Beta}\left(8,2,Q_{\min},Q_{\max}\right)$ where $Q_{\min}=\max\left(\min\left(r_{a_1,a_1},r_{a_1,a_2}\right),\min\left(r_{a_2,a_1},r_{a_2,a_2}\right)\right)$ and $Q_{\max}=\min\left(\max\left(r_{a_1,a_1},r_{a_1,a_2}\right),\max\left(r_{a_2,a_1},r_{a_2,a_2}\right)\right)$.

Unless otherwise specified, the experimental results of the theoretical calculations are obtained for all regular graphs based on finite-difference method integration with an interval of 0.01. The simulation results are based on 5000 experiments repeated on graphs' population of 1000 agents. Our experimental results are displayed in Figure 2.

We illustrate the Q-value and policy of the population as a function of the time step for a random regular graph of degree $k=4$ with three different initial Q-value distributions. The dots represent the results of the agent-based simulations, and the lines represent the theoretical results derived from our dynamics model. By comparing the average Q-value $q(a_i)$ and the average strategy $\bar{x}(a_i)$ obtained by our model with the corresponding simulation results, we find that our model can capture the evolution process even in the case of considering local interactions.

In addition, Our model also reflects the differences between the evolution processes in different games. We found that for the CG and MP, due to different initial Q-value distributions and strategies, the final strategy of the agents will approximate two distinct Nash equilibrium. For the remaining three games (PD, SD, SH), regardless of the initial Q-value distribution, the evolutionary results eventually approximate a Nash equilibrium. Take PD for example, defection (a_2) becomes the dominant strategy, which is consistent with the prediction of traditional game theory.

5.2 Differences of populations with varying graph structures

We conducted relevant experiments to investigate the performance differences of populations with varying graph structures and degree distributions across different games.

Figure 3 illustrates the performance of our method on populations with four different graph structures. To avoid the influence of other factors, we constrained the graphs to have an average degree of 4. We selected three non-regular graphs with an average degree of approximately 4, namely Barabási-Albert (BA) scale-free network [54], Erdős-Rényi (ER) random graph [55], and random geometric graphs (RGG) [56]. Additionally, we included a regular graph with a degree of 4. We generated 20 different graphs using the same parameters for experimentation for each non-regular graph type and recorded the results. Due to construction constraints with a guaranteed average degree of the graphs, here the ER graph is a population of 400 agents and the RGG graph consists of 100 agents.

In Figure 3(a), we display the theoretical and experimental results of the SD on these various graphs. Although we conducted experiments for all five games, due to space limitations, we chose to present

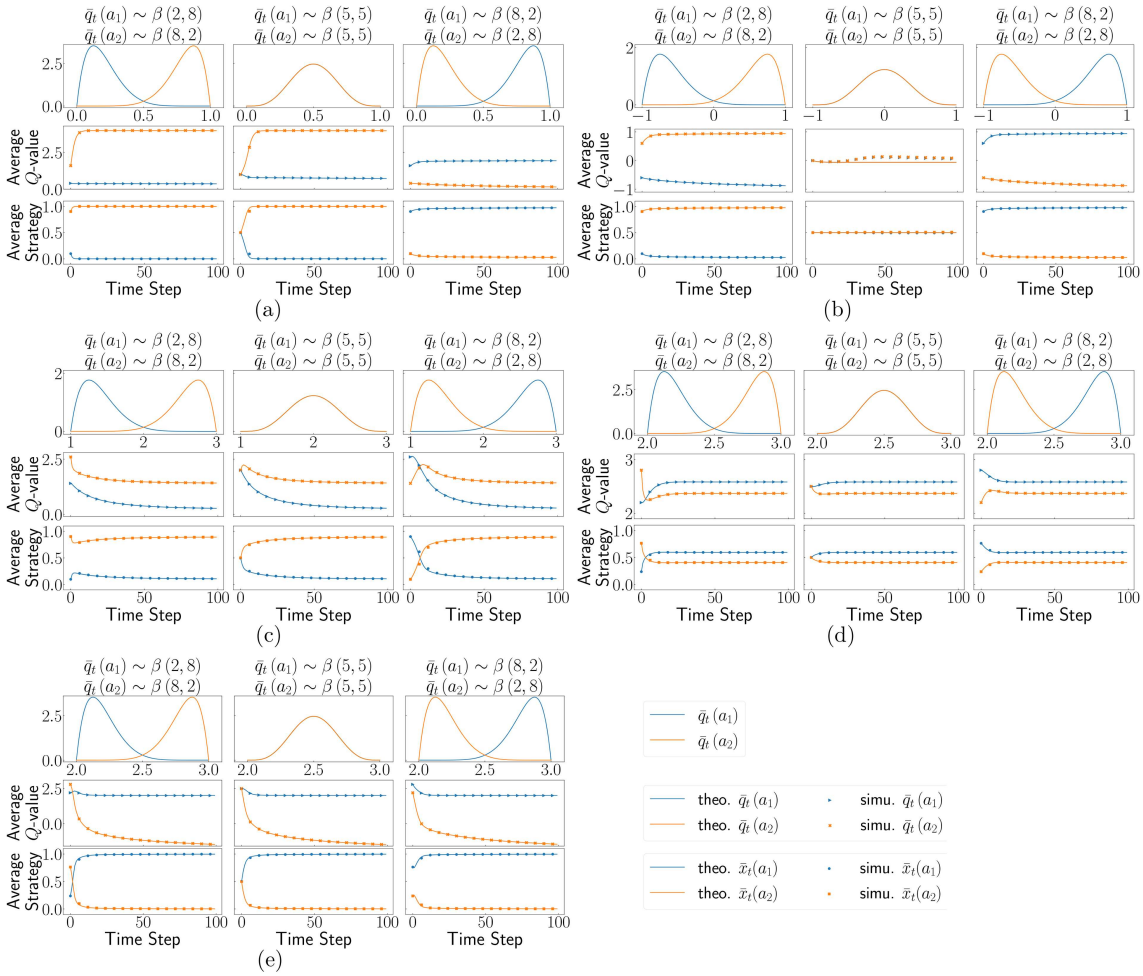

Figure 2 Time series of the average Q-values and strategies on a regular graph of degree $k = 4$ under 3 different initial Q-value distributions. To simplify, we will ignore the last two parameters of the Beta distribution and represent Beta (alpha, beta, $Q_{\min}$, $Q_{\max}$) as Beta (alpha, beta). (a) CG; (b) MP; (c) PD; (d) SD; (e) SH.

the results of the SD, which showed the most significant performance differences. In this figure, the line segments represent the theoretical calculation results. In contrast, the filled polygons represent the actual experimental results, that the width of each polygon corresponds to twice the standard deviation of the bias between the graphs, and the center represents the average value of the calculated results.

The theoretical and simulation results of the agents on ER network and RGG are similar. This observation aligns with their degree distributions, and the differences between the results of these two graphs and the regular graph can be attributed to the presence of agents with different degrees in the graphs. Correspondingly, the results in the BA network show a more significant discrepancy compared to the ER network and RGG, which is consistent with their degree distributions.

Furthermore, we also observed a significant discrepancy between the BA network's simulation and theoretical calculation results compared to those on ER networks and RGG. The reasons are twofold. First, agents' degrees are generally not the same. Second, by the preferential attachment of the BA networks, agents are more likely to be linked to those with high degrees, causing their actual neighbor configurations to deviate slightly from the mean-field neighbor configuration. This increases the discrepancies between theoretical predictions and experimental results. As a result, there are slight discrepancies between the results of the theoretical calculation and the results obtained through simulation. Moreover, the divergence between theoretical and experimental results on random graphs stabilizes after about 15-time steps, rather than increasing continuously. This suggests that the model effectively captures the dynamics of populations in most random graphs, despite some initial discrepancies.

In addition, according to Delgado's [57] and Sen et al.'s [48, 58] research, certain structural aspects within the graph contribute to specificities in the actions of some agents compared to the entire population.

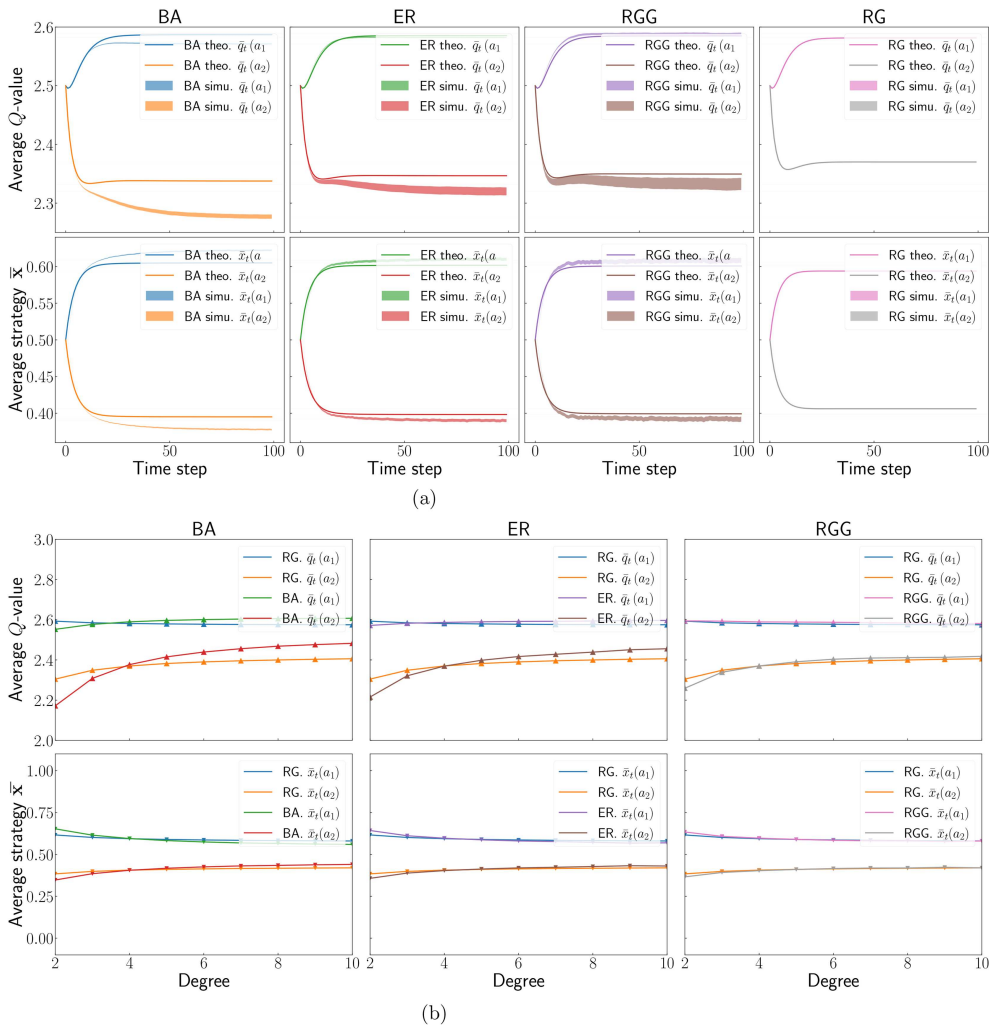

Figure 3 Experiment results of the SD for the populations in three graph structures: BA, ER, and RGG. (a) We present the simulation and theoretical calculation results of the SD for populations in the three graph structures with an average degree of around 4. Additionally, we provide a comparison with the results from a regular graph population with a degree of 4. (b) We extracted the average Q-values and strategies of agents with varying degrees within the populations of these three graph structures.

This, too, contributes to the observed differences between simulation and theoretical results in the BA network.

In Figure 3(b), we examined the Q-values and strategies of agents with degrees ranging from 2 to 10 in the populations on these three graphs. We compared these results with the corresponding agents in regular graphs with the same degrees. The findings demonstrate that the differences in Q-values and strategies due to varying degrees in the BA network are more significant than in the other two graphs. This observation indirectly validates that specific unique graph structures in the BA network can lead to these differences, as observed in our experiments. Furthermore, we can deduce from Figure 3(b) that the behavioral performance of agents of the same degree tends to be the same in different populations, and small-degree agents and those in scale-free graphs differed slightly from others in their behavior and Q-values.

5.3 Variation in population strategies and Q-values induced by regular graph structures of different degrees

Furthermore, to explore the distinctions among individuals with varying degrees, we conducted experiments on regular graphs with different degrees, as depicted in Figure 4. Similarly, we conducted experiments for the five aforementioned games. However, due to space limitations, we have chosen to present only the results of the SD. We can observe that populations with different degrees in regular graph struc-

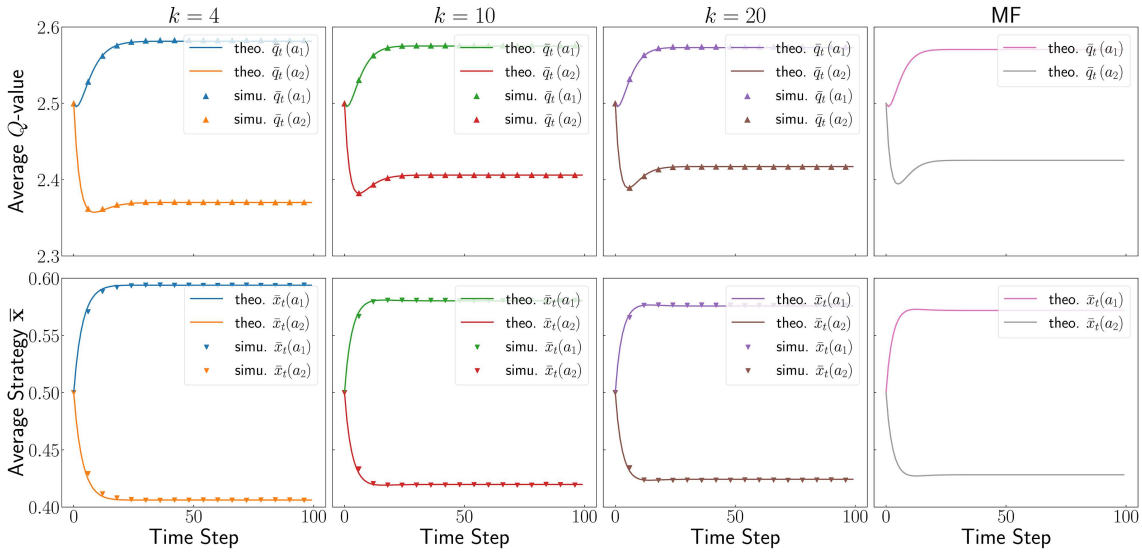

Figure 4 Simulation and theoretical calculation results for populations with different degrees in the SD with regular graph structures, where MF represents results obtained using the mean-field method.

tures also exhibit slight variations in experimental results. As the degree increases, the experimental results for the regular graph approach are closer to those obtained by the mean-field method. Among them, the population with a degree of 4 shows the most significant discrepancy compared to populations with other degrees. Besides, the experimental results for the population with a degree of 20 closely approach the result obtained through the mean-field method.

Furthermore, to delve deeper into the variations among populations with different degrees in regular graph structures across the PD, SH, and SD games, we have depicted the changes between these three distinct game results using line plots and violin plots, as illustrated in Figure 5. For Figure 5(a), the lines in the left-side line plot represent theoretical calculation results, while the points represent results from multiple simulations. The right-side violin plot illustrates the data distribution and its range of experimental results for populations on graphs with various degrees. Figures 5(b)–(d) show line charts of MASs' strategies and their degrees. It is clear that in SH, an agent's degree has minimal impact on strategy selection, whereas in SD, it significantly influences their strategic choices. Moreover, the strategies of agents are more significantly affected only when the degree is small ($\lessapprox 10$).

It can be observed that for different games, as the probability of each action approaches 0.5, the range of strategy distribution becomes larger. This is because, in the case of two-action games, their strategies can be considered as Bernoulli distributions, with the corresponding variance $\Delta^2 = \overline{x}_t(a_1)(1 - \overline{x}_t(a_1))$. The closer an action is to 0.5, the more significant its variance. Moreover, the bias of these actions between the result of regular graphs of various degrees and the result obtained by the mean-field results exponentially decreases as the degree increases. In addition, the behavior of agents in the small degree regular graphs has more significant changes as a result of degree changes than those on the large degree regular graphs. These observations align with what we discovered in Corollary 1.

6 Conclusion and future work

This paper proposes a formal model of the Q-learning dynamics on a graph. By capturing the influence of different neighbor configurations on the learning of agents, we derive a system of differential equations to describe the evolution of the Q-value distribution. To verify the accuracy of our theoretical model, we conduct a series of agent-based simulations on regular, random graphs and scale-free graphs. Under five typical symmetric normal-form games, the behavioral evolution is consistent with the prediction of our theoretical model. With an increase in degree, agents' actions tend to concentrate more on the results of a graphless (mean-field) structure. In addition, our model captures the effect of the degree on agents' behavior on graphs and unveils the relationship between the strategies of many agents and the degree distribution of the underlying graph structure.

Our current model has limitations in detailing the evolutionary dynamics of MASs with complex topo-

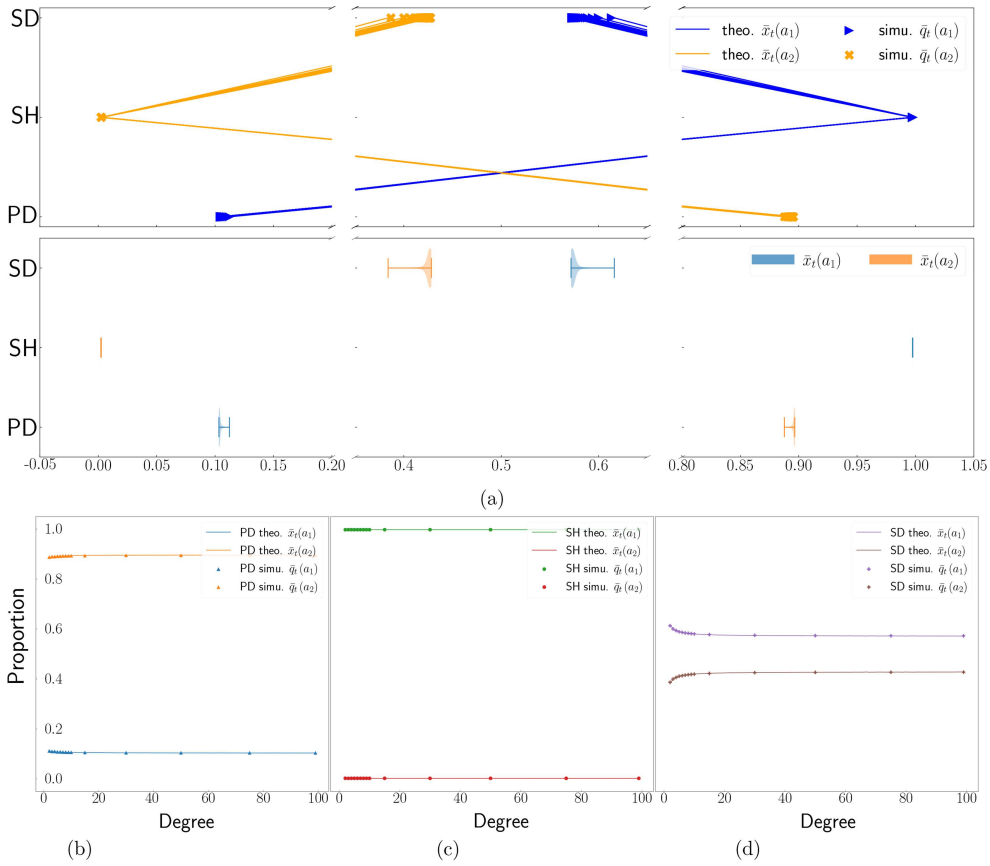

Figure 5 (a) We conducted experiments on agent populations on regular graphs (degree 2–100) for the PD, SH, and SD games, and represented their strategy distributions using violin plots. (b)–(d) Line charts of MASs' strategies and their degrees, which clearly show that in SD, strategies are most affected by degree, while in SH, the impact is minimal. Furthermore, the strategies of agents are significantly more affected only when their degrees are small ($\lessapprox 10$).

logical structures. For instance, clusters of agents may evolve strategies that diverge from the entire population. To address this, we plan to incorporate more topological attributes into our model. Future work will also extend our approach to include more complex game scenarios such as asymmetric and stochastic games, and explore the dynamics of games on coupling networks. Additionally, we will implement various learning algorithms to improve the robustness and applicability of our models. These enhancements will enable us to explore a wider range of strategic interactions and network topologies, thus providing deeper insights into the dynamics of MASs.

Acknowledgements This work was supported by National Science Fund for Distinguished Young Scholarship of China (Grant No. 62025602), National Natural Science Foundation of China (Grant Nos. U22B2036, 62366058, 11937815, 62066045, 11971421, 11931015, 62261136549), Excellent Youths Project for Basic Research of Yunnan Province (Grant No. 202101AW070015), Fok Ying-Tong Education Foundation, China (Grant No. 171105), Yunnan Province XingDian Talent Support Program (Grant Nos. YNWR-QNBJ-2020-041, YNWR-YLXZ-2018-020), Foundation of Yunnan Key Laboratory of Service Computing (Grant No. YNSC23117), and Tencent Foundation and XPLORER PRIZE.

Supporting information Appendixes A–F. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhou M, Jiarui J, Zhang W, et al. Multi-agent reinforcement learning for order-dispatching via order-vehicle distribution matching. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019. 2645–2653

2 Liu S, Lever G, Wang Z, et al. From motor control to team play in simulated humanoid football. Sci Robot, 2022, 7: eabo0235

3 Hüttenrauch M, Šošić A, Neumann G. Deep reinforcement learning for swarm systems. J Mach Learn Res, 2019, 20: 1–31

4 Blais M A, Akhloufi M A. Reinforcement learning for swarm robotics: an overview of applications, algorithms and simulators. Cogn Robot, 2023, 3: 226–256

5 Derakhshan F, Yousefi S. A review on the applications of multiagent systems in wireless sensor networks. Int J Distrib Sens Netws, 2019, 15: 155014771985076

6 Naderializadeh N, Sydir J J, Simsek M, et al. Resource management in wireless networks via multi-agent deep reinforcement learning. IEEE Trans Wireless Commun, 2021, 20: 3507–3523

7 Bosse T, Treur J. Formal interpretation of a multi-agent society as a single agent. J Artif Soc Social Simul, 2006, 9: 1–6

8 Yoon M G. Single agent control for multi-agent dynamical consensus systems. IET Control Theor Appl, 2012, 6: 1478–1485

9 Hotz V J, Miller R A. Conditional choice probabilities and the estimation of dynamic models. Rev Economic Studies, 1993, 60: 497–529

10 Shum M. Econometric Models for Industrial Organization. Singapore: World Scientific, 2016

11 Bailey J P, Piliouras G. Multi-agent learning in network zero-sum games is a hamiltonian system. In: Proceedings of International Conference on Autonomous Agents and Multiagent Systems, 2019

12 Engel M, Piliouras G. A stochastic variant of replicator dynamics in zero-sum games and its invariant measures. Phys D-Nonlinear Phenom, 2023, 456: 133940

13 Hussain A, Belardinelli F, Piliouras G. Beyond strict competition: approximate convergence of multi agent Q-learning dynamics. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence, 2023. 135–143

14 Czechowski A, Piliouras G. Non-chaotic limit sets in multi-agent learning. Auton Agent Multi-Agent Syst, 2023, 37: 29

15 Liu L, Chen X, Szolnoki A. Coevolutionary dynamics via adaptive feedback in collective-risk social dilemma game. eLife, 2023, 12: e82954

16 Wang S, Yao W, Cao M, et al. Evolutionary dynamics under periodic switching of update rules on regular networks. IEEE Trans Netw Sci Eng, 2024, 11: 1337–1346

17 Tuyls K, Verbeeck K, Lenaerts T. A selection-mutation model for Q-learning in multi-agent systems. In: Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, 2003. 693–700

18 Sato Y, Crutchfield J P. Coupled replicator equations for the dynamics of learning in multiagent systems. Phys Rev E, 2003, 67: 015206

19 Watkins C J C H, Dayan P. Q-learning. Mach Learn, 1992, 8: 279–292

20 Sandholm W H. Population Games and Evolutionary Dynamics. Cambridge: MIT Press, 2010

21 Panozzo F, Gatti N, Restelli M. Evolutionary dynamics of Q-learning over the sequence form. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2014

22 Bloembergen D, Tuyls K, Hennes D, et al. Evolutionary dynamics of multi-agent learning: a survey. J Artif Intell Res, 2015, 53: 659–697

23 Gomes E R, Kowalczyk R. Dynamic analysis of multiagent Q-learning with ε-greedy exploration. In: Proceedings of the 26th Annual International Conference on Machine Learning, 2009. 369–376

24 Wunder M, Littman M L, Babes M. Classes of multiagent Q-learning dynamics with epsilon-greedy exploration. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010. 1167–1174

25 Tuyls K, Parsons S. What evolutionary game theory tells us about multiagent learning. Artif Intell, 2007, 171: 406–416

26 Weibull J W. Evolutionary Game Theory. Cambridge: MIT Press, 1997

27 Phelps S, Parsons S, McBurney P. An evolutionary game-theoretic comparison of two double-auction market designs. In: Proceedings of AAMAS 2004 Workshop on Agent-Mediated Electronic Commerce VI, 2005. 101–114

28 Ponsen M, Tuyls K, Kaisers M, et al. An evolutionary game-theoretic analysis of poker strategies. Entertain Comput, 2009, 1: 39–45

29 Kianercy A, Galstyan A. Dynamics of Boltzmann Q learning in two-player two-action games. Phys Rev E, 2012, 85: 041145

30 Kaisers M, Tuyls K. Frequency adjusted multi-agent Q-learning. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, 2010. 309–316

31 Leonardos S, Piliouras G. Exploration-exploitation in multi-agent learning: catastrophe theory meets game theory. Artif Intell, 2022, 304: 103653

32 Hu S, Leung C, Leung H. Modelling the dynamics of multiagent Q-learning in repeated symmetric games: a mean field theoretic approach. In: Proceedings of Advances in Neural Information Processing Systems, 2019

33 Axtell R. Effects of interaction topology and activation regime in several multi-agent systems. In: Proceedings of International Workshop on Multi-Agent Systems and Agent-Based Simulation, 2000. 33–48

34 Franks H, Griffiths N, Anand S S. Learning agent influence in MAS with complex social networks. Auton Agent Multi-Agent Syst, 2014, 28: 836–866

35 Klabunde A, Willekens F. Decision-making in agent-based models of migration: state of the art and challenges. Eur J Population, 2016, 32: 73–97

36 Will M, Groeneveld J, Frank K, et al. Combining social network analysis and agent-based modelling to explore dynamics of human interaction: a review. Socio-Environ Syst Model, 2020, 2: 16325

37 Wang S, Chen X, Xiao Z, et al. Optimization of institutional incentives for cooperation in structured populations. J R Soc Interface, 2023, 20: 20220653

38 Leung C W, Hu S, Leung H F. Formal modeling of reinforcement learning with many agents through repeated local interactions.

In: Proceedings of the 33rd International Conference on Tools with Artificial Intelligence (ICTAI), 2021. 714–718

39 Gong X, Xu J. Research on delay characteristics of information in scale-free networks based on multi-agent simulation. Procedia Comput Sci, 2013, 17: 989–1002

40 Liu Z, Nojavanzadeh D, Saberi A, et al. Scale-free collaborative protocol design for output synchronization of heterogeneous multi-agent systems with nonuniform communication delays. IEEE Trans Netw Sci Eng, 2022, 9: 2882–2894

41 Kaisers M, Bloembergen D, Tuyls K. A common gradient in multi-agent reinforcement learning. In: Proceedings of International Conference on Autonomous Agents and Multiagent Systems, 2012. 1393–1394

42 Panozzo F, Gatti N, Restelli M. Evolutionary dynamics of Q-learning over the sequence form. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2014

43 Kianercy A, Galstyan A. Dynamics of softmax Q-learning in two-player two-action games. 2011. ArXiv:1109.1528

44 Klos T, van Ahee G J, Tuyls K. Evolutionary dynamics of regret minimization. In: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2010. 82–96

45 Hu S, Leung C W, Leung H, et al. The dynamics of Q-learning in population games: a physics-inspired continuity equation model. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, 2022. 615–623

46 Luo Q, Liu L, Chen X. Evolutionary dynamics of cooperation in the N-person stag hunt game. Phys D-Nonlinear Phenom, 2021, 424: 132943

47 Villatoro D, Sen S, Sabater-Mir J. Topology and memory effect on convention emergence. In: Proceedings of IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009. 233–240

48 Villatoro D, Sabater-Mir J, Sen S. Social instruments for robust convention emergence. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), 2011. 420–425

49 Sen O, Sen S. Effects of social network topology and options on norm emergence. In: Proceedings of International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, 2009. 211–222

50 Yu C, Zhang M J, Ren F H. Collective learning for the emergence of social norms in networked multiagent systems. IEEE Trans Cybern, 2014, 44: 2342–2355

51 Nowé A, Vrancx P, De Hauwere Y M. Game theory and multi-agent reinforcement learning. In: Adaptation, Learning, and Optimization. Belin: Springer, 2012. 441–470

52 Cesa-Bianchi N, Gentile C, Lugosi G, et al. Boltzmann exploration done right. In: Proceedings of Advances in Neural Information Processing Systems, 2017

53 Crandall J W, Goodrich M A. Learning to compete, compromise, and cooperate in repeated general-sum games. In: Proceedings of the 22nd International Conference on Machine Learning, 2005. 161–168

54 Barabási A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286: 509–512

55 Erdős P, Rényi A. On the evolution of random graphs. Publ Math Inst Hung Acad Sci, 1960, 5: 17–60

56 Boccaletti S, Latora V, Moreno Y, et al. Complex networks: structure and dynamics. Phys Rep, 2006, 424: 175–308

57 Delgado J. Emergence of social conventions in complex networks. Artif Intell, 2002, 141: 171–185

58 Sen O, Sen S. Effects of social network topology and options on norm emergence. In: Proceedings of International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, 2009. 211–222