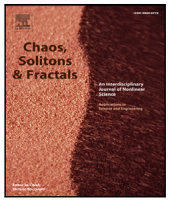
Chaos, Solitons and Fractals 187 (2024) 115377

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

journal homepage: www.elsevier.com/locate/chaos



Learning and propagation: Evolutionary dynamics in spatial public goods games through combined Q-learning and Fermi rule

Yong Shen, Yujie Ma, Hongwei Kang *, Xingping Sun, Qingyi Chen

School of Software, Yunnan University, Kunming 650000, China

ARTICLE INFO

Keywords:
Public goods game
Learning dynamics
Imitation dynamics
Agent cooperation

ABSTRACT

Propagation is crucial for acquiring information, and learning involves deep information processing. Imitation dynamics, commonly used in spatial public goods games, represents strategy propagation in society, while learning dynamics enables agents to self-learn through environmental interactions. In this paper, Q-learning and the Fermi update rule are used to compare differences between learning dynamics and imitation dynamics in simulation experiments. Our study finds that Q-learning is continuously updating the Q-table during its evolution, forming a heterogeneous gain matrix. However, Q-learning's interaction with the environment is indirect, with almost no network reciprocity, and cannot form clusters in contrast to the Fermi update rule. The Fermi update rule focus only on immediate payoffs. Furthermore, we combine imitation dynamics and learning dynamics, integrating the advantages of both. Defectors and cooperators form a special semi-stable checkerboard-like state. The update rules, the number of states, and the configuration of conversion actions as strategies have led to this special structure. Moreover, we found that parameter choices in the new evolutionary dynamics affect the percentage of cooperation and evolutionary convergence. Overall, this paper offers new insights into learning dynamics and imitation dynamics and provides a foundation for integrating complex mechanisms of evolution.

1. Introduction

Cooperation has always been a powerful force driving human society forward, playing a crucial role in the course of civilization [1–3]. Whether it is the division of labor in ancient hunter-gatherer societies and farming societies or the cooperation between modern enterprises or countries, cooperation has always been the pillar of human survival and development. In the face of the economic, political, and scientific challenges of the 21st century, it is particularly important to understand and analyze cooperation [4,5]. Evolutionary game theory provides a crucial theoretical framework for studying cooperation and behavioral evolution [6–9]. Public goods game have become a widely studied and important area of evolutionary games, offering various options for facilitating cooperation.

Public goods game are a multi-individual cooperation problem. It requires individuals to contribute to a shared resource but does not guarantee individual benefits. In spatial public goods game, individuals can choose not to share resources, they can profit from others' contributions without bearing the cost of cooperation. The cooperation dilemma lies in the tendency of individuals to maximize their short-term payoffs without considering the overall payoffs and long-term payoffs, and thus unwilling to contribute to fall into total betrayal [10–13]. Therefore, researchers have proposed a series of mechanisms such as rewards [14–19], punishments [20–25], and reputation [26–31], taxation mechanism [32–34], Heterogeneous Investments [35], and Exclusionary Mechanisms [36,37], to address the challenges of cooperation and enhance the level of cooperation. These mechanisms are consistent with the laws of social operation, and promote the evolution of cooperation in multiple dimensions. However, the above studies usually play the evolutionary games under the framework of the Fermi update rule [38] or replication dynamics [39]. The Fermi update rule randomly selects a neighbor as an object to learn from its Strategy by comparing with it the payoff of the previous step. This process emphasizes the influence of immediate payoffs and social network structure on Strategy selection. However, the game is a rational choice made by individuals to pursue and realize their payoffs. Selfish and short-sighted Strategies often lead to the breakdown of cooperative relationships.

Recently, learning dynamics, exemplified by reinforcement learning, has opened new avenues for analyzing cooperation and games. Reinforcement learning allows agents to learn actions through environmental interaction to maximize expected cumulative payoffs. Learning

* Corresponding author.
E-mail address: hwkang@ynu.edu.cn (H. Kang).

https://doi.org/10.1016/j.chaos.2024.115377
Received 1 July 2024; Received in revised form 26 July 2024; Accepted 4 August 2024
Available online 13 August 2024

dynamics involves the processes of learning and memory within an evolutionary framework. These processes enable individuals to adopt a more long-term perspective. Many researchers have explored cooperation using reinforcement learning [40–44]. Q-learning is a classic algorithm in reinforcement learning [45–48]. The algorithm involves agents interacting with the environment and constructing a Q-table. The Q-table estimates the state–action values for the agent's behaviors in different states. The value estimation is determined by memory and the current environment. This reflects the influence of learning and experience on decision-making.

The rest of the paper is structured as follows: In Section 2, the Strategy update rule of this paper is described in detail. In Section 3, simulation experiments on the Strategy update rule are carried out and the experimental results are analyzed in depth. Finally, In Section 4, the paper is summarized.

2. Model

We consider a two-strategy Spatial Public Goods Game (SPGG) on an $N = L \times L$ square lattice with Von Neumann neighborhoods and periodic boundaries, where each vertex represents agents. Agents play in groups with their $k = 4$ nearest neighbors. They belong to $g = 1, \ldots, G(G = 5)$ overlapping groups, each containing $k + 1$ members. These five groups are centered on agents x and its four neighbors. Initially, agents are randomly assigned to one of two strategies: betrayal ($s_x = D$) or cooperation ($s_x = C$). The cooperative strategy ($s_x = C$) contributes $c = 1$ unit to the public pool, while the betrayal strategy ($s_x = D$) does not contribute. The total investment in each group is the number of cooperative strategies multiplied by the contribution value. This investment is then multiplied by a synergy factor r and equally distributed among agents. To compare evolutionary dynamics without model influence, we use the most basic model of the public goods game. The model is as follows:

$$\begin{aligned} \Pi_C^g &= r\frac{N_C^g}{G} - 1, \\ \Pi_D^g &= r\frac{N_C^g}{G}, \end{aligned} \tag{1}$$

where N_C and N_D represent the number of cooperators and defectors in the group, respectively, and r is the synergy factor.

Agents in SPGG all play games with four neighbors centered on themselves. Thus, each agent x participates in five games with a cumulative payoff of Π_x, defined as:

$$\Pi_x = \sum_{g=1}^{G} \Pi_x^g. \tag{2}$$

We employ three different algorithms for evolutionary games, namely Q-learning, the Fermi update rule, and evolutionary dynamics combining propagation and learning. In order to correspond the figures to evolutionary dynamics, the following illustrations use $CaseI$, $CaseII$, $CaseIII$ instead of the Q-learning, the Fermi update rule, and evolutionary dynamics combining propagation and learning.

2.1. Q-learning

In Q-learning, we define the s of the original public goods game model as the state of the current agent. Then, a is defined as the decision made in the current state s. The set S of states is equal to the set A of action as $\{D, C\}$. In particular, the current agent's state is also the action of the previous time step agent, and the action a made by the agent in the current state is also the state of the agent in the next time step. The immediate reward r is defined as the cumulative payoff $\Pi(t)$ of the current step. Q-table is a two-dimensional array where rows denote different states and columns denote different actions. Q-table stores the expected cumulative value of the action a in state s. Q-table is represented as follows:

$$Q(t) = \begin{bmatrix} Q_{D,D}(t) & Q_{D,C}(t) \\ Q_{C,D}(t) & Q_{C,C}(t) \end{bmatrix}. \tag{3}$$

In Q-learning, agents use a $\epsilon - greedy$ algorithm to make decisions. Each agent has a $1 - \epsilon$ probability of choosing the action with the highest action value in the current state. Otherwise, a random action is selected. This allows agents to select the optimal action based on the state action value, and also to explore alternative actions in case of environmental changes. Typically, ϵ is set 0.02. When the maximum Q-value corresponds to more than one action, agents randomly select one of the actions. Q-learning uses an updating formula based on temporal difference to estimate the state action value. The updating formula is as follows:

$$Q_{(s,a)}(t+1) = Q_{(s,a)}(t) + \alpha[\Pi(t) + \gamma Q^{\max}_{(s',a')}(t) - Q_{(s,a)}(t)], \tag{4}$$

Where s' denotes the state at the next time step, and a' denotes the action at the next time step. $Q^{\max}_{(s',a')}(t)$ denotes the maximum Q-value corresponding to all possible actions in the next state s'. It represents the expected value of the best action agents can take. $\alpha \in [0, 1]$ denotes the learning rate, which controls the speed of updating the Q-value. $\gamma \in [0, 1]$ denotes the discount factor, which determines the weight of future expectations in Q-learning.

2.2. The Fermi update rule

In imitation dynamics, agent x chooses a neighbor y at random, compares its payoff Π_x with this neighbor's payoff Π_y, and learns y's strategy with probability W, defined as:

$$W(S_x \leftarrow S_y) = \frac{1}{1 + \exp\left(\frac{\Pi_x - \Pi_y}{K}\right)}, \tag{5}$$

Where K represents the noise in the propagation process, controlling the extent to which differences in payoffs affect the probability of strategy learning. In the Fermi rule, individuals tend to favor strategies with higher payoffs. Previous work using Fermi functions to mimic dynamics has made important contributions to spatial public goods games on square lattice topologies [49,50]. These works investigated how noise levels, critical mass thresholds, and group sizes affect the stability and evolutionary paths of cooperation. They revealed important findings such as the noise level is topologically independent of the topology of cooperation evolution, group size determines the noise dependent of cooperation in spatial public goods games, and critical mass can effectively promote cooperation in spatial public goods games. In order to better compare the different evolutionary dynamics independent of other factors, some of the most representative parameters and rules are selected in this paper.

2.3. Learning and propagation

In the evolutionary dynamics that combines learning and propagation, agents will also have a Q-table to evaluate state–action values. The state set S, the action set A and the Q-table definition are consistent with Q-learning. The way Q-table is updated and agents select actions is different from Q-learning. The Q-table is updated as follows:

$$Q_{(s,a)}(t+1) = (1-\eta)Q_{(s,a)}(t) + \eta[\Pi(t) + \gamma Q^{\max}_{(s',a')}(t)], \tag{6}$$

where $\eta \in [0, 1]$ denotes the learning rate and the decay of past experiences. $\gamma \in\in [0, 1]$ is the expected yield discount factor for the next state s'. $Q^{\max}_{(s',a')}(t)$ denotes the maximum Q-value corresponding to all possible actions in the next state s'.

The Q-value in the Q-table does not directly guide agents to make a decision. Agents will randomly choose a neighbor, compare its $Q^x_{(s,a)}(t)$

with this neighbor's $Q^{y}_{(s,a)}(t)$ and learn the Strategy for y with a probability W, defined as:

$$W(S_x \leftarrow S_y) = \frac{1}{1+\exp\left(\frac{Q^{x}_{(s,a)}(t)-Q^{y}_{(s,a)}(t)}{K}\right)}. \tag{7}$$

The specific evolutionary process is shown in Algorithm 1.

All three evolutionary dynamics implement the evolutionary process using asynchronous Monte Carlo simulations (MCs) on a square lattice of $L = 200$. Initially, each agent is randomly assigned to be either a cooperator (C) or a defector (D) with equal probability, unless otherwise specified. To minimize the effect of uncertainty, the average of 10 independent experiments was used as the final result.

Algorithm 1: Evolutionary games with $CaseIII$ algorithm

```
Initialize the Q-table to zero for each i ∈ L × L square lattice;
Initial each agent i's strategy C,D randomly;
for each time step t ∈ [1, n] do
    for each node i ∈ L × L do
        select an neighbor y randomly;
        calculate W according to Eq. (7);
        Generate a random probability ϵ;
        if w > ϵ then
            Learn the strategy of the agent's neighbors
        end
    end
    for each node i ∈ L × L do
        calculate i's payoff π according to Eq. (1);
    end
    for each node i ∈ L × L do
        update Q-table according to Eq. (6);
    end
end
```

3. Results

3.1. Q-learning forms a unique matrix of heterogeneous perceived benefits

First, we set all initial strategies as defectors to observe the evolution of Q-learning, as shown in Fig. 1. This initialization implies that all agents will not gain, reflecting a poor Nash equilibrium state. Occasional cooperators in this state will have low or negative payoffs due to being surrounded by defectors. When the Q-table is initialized to 0, it represents that all agents lack prior knowledge and will favor exploration at the game's start. Thus, the initial strategy does not significantly affect the convergence result. Over time, agents explore and estimate the values of actions, recorded in the Q-table, until the values eventually converge. The Q-table is similar to the payoff matrix in a two-player game. However, in the special matrix, the two sides are not two agents, but the current state and the next action. This means players focus on their current state and expected next move, choosing strategies without considering their neighbors' information. Agents indirectly perceive environmental information through each step of the game. In SPGG, each agent forms a heterogeneous perceived benefit matrix (Q-table) through games and perception. This matrix is used to evaluate the value of the current state action. In traditional public goods game, agents adjust their strategy only by the payoff from the previous step. In contrast, in Q-learning, agents iterate their perceptual payoff matrix during the game, guiding their actions through this matrix and forming a unique mechanism for learning and memorization.

Observing the updating formula Eq. (4) for the Q-table in Q-learning, agents perceive environmental information indirectly through each step of the game. This information influences the Q-table iteration only through the agents' payoffs and future expectations. The snapshots show that Q-learning evolves based solely on the agent's experience without direct neighbor interactions, leading to almost no network reciprocity. There is no strong connection or interaction between cooperators and defectors.

3.2. Comparison of Q-learning and the Fermi update rule

To compare the effects of Q-learning and the Fermi update rule on the evolution of cooperation, we examine the differences in experimental results when the synergy factor r varies. As shown in Fig. 2, when $r = 3.6$, the proportion of Q-learning cooperators is the first to start climbing. When $r = 3.7$, the Fermi update rule cooperators start to appear. As r increases, the proportion of cooperators rises. This indicates that cooperators are more likely to persist and survive when r is small in Q-learning.

However, Q-learning shows a relatively flat increase. In contrast, under Fermi's rule, the proportion of cooperators rapidly climbs from 0 to about 0.4 once r reaches the threshold for cooperator survival. Additionally, the ϵ-greedy exploration mechanism in Q-learning prevents cooperators from completely disappearing even when r is less than 3.6, maintaining the proportion of agents exploring Strategy C at 0.01. However, this does not mean cooperators survive. In each step, 0.01 proportion of cooperators are not the same agents. These agents are selected to explore strategies. After this step, they will revert to being defectors. We consider cooperators to survive only when the proportion of cooperators exceeds 0.01. Moreover, at $r = 5.0$, a high value of the synergy factor, the proportion of the Fermi update rule cooperators is close to 1, while Q-learning stabilizes around 0.65. This indicates that Q-learning is a stability-seeking evolutionary dynamics that can negatively affect the growth of cooperators beyond a certain point.

Figs. 3 and 4 show the time evolution curves and snapshots of Q-learning and the Fermi update rule updates at $r = 3.6, 3.8$ and 5.0, respectively. The time evolution curves show defectors in red and cooperators in blue. Snapshots show defectors in black and cooperators in white. A comparison of the time evolution curves reveals that Q-learning evolves more gently than the Fermi update rule. Q-learning remains relatively stable around $t < 100$, during which time the Q-table continues to be updated until it converges, and then the cooperators and defectors begin to diverge in large numbers.

Looking at the snapshots, the Fermi update rule starts with the defectors dominating and then the cooperators forming clusters to protect themselves. Then the cooperators have higher payoffs when r is larger and start expanding to occupy the defectors' territories in all directions. Cooperators' territories are gradually swallowed up by the defectors when r is smaller. The clusters are shown in Fig. 5. Observing the clusters, the center of the cluster of cooperators has high payoffs. Cooperators at the edge of the clusters have similar payoffs as defectors. Defectors who are not at the edge of the cooperator cluster have zero payoff. It can be seen that the clusters effectively protect the cooperators. Q-learning struggles to form clusters of cooperators and thus fails to effectively shield them, resulting in a slow increase in the proportion of cooperators as the value of r grows.

Q-learning's memory mechanisms are stable, hindering any strategy that tries to change the current stable state and are not adapted to rapidly changing environments. The Fermi update rule allows for the rapid propagation of strategies that can quickly adapt to environmental changes. Q-learning uses only its payoff and experience to update and construct its strategy, whereas the Fermi update rule uses its own and others' payoffs to construct its strategy. The key difference is that Q-learning and the Fermi update rule use different information and apply it to different scenarios. The Fermi update rule requires information about its neighbors to disseminate strategies. Therefore, using Q-learning as an evolutionary dynamics requires it to undergo environmental changes and gradually update the Q-table to adapt. This hinders the propagation of strategies, forming clusters, and protecting cooperators.

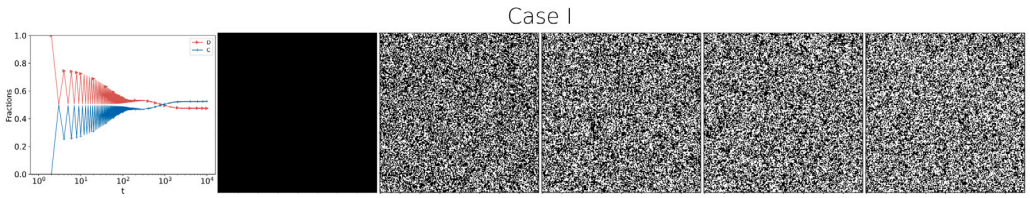


Fig. 1. Time evolution curves of cooperators (blue) and defectors (red) over time and snapshots of cooperators (white) and defectors (black). Notably, all initial strategies are set as defectors. The snapshots are plotted over time (from left to right, $t = 1, 10, 100, 1000$ and 10000). The illustration shows the evolution from all defectors to an even mix of defectors and cooperators. Results are shown for $r = 4.7$, $\gamma = 0.8$, $\alpha = 0.8$ and $L = 200$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

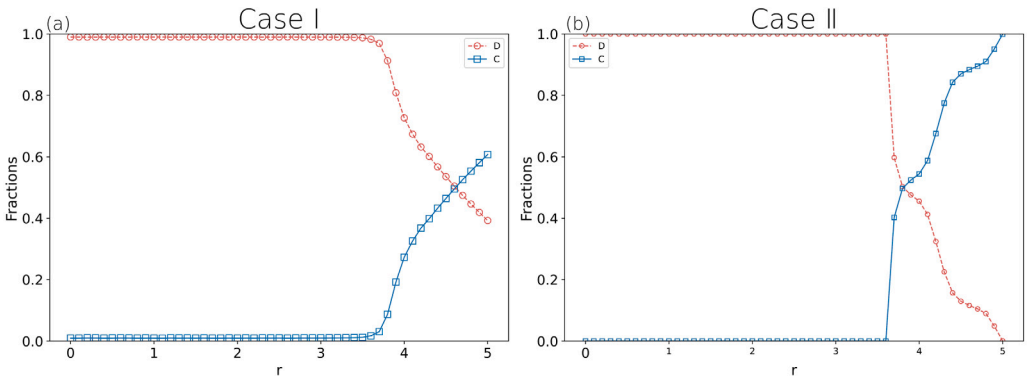


Fig. 2. For different values of r, the proportion of cooperators (blue) and the fraction of defectors (red). The illustration shows cooperators surviving earlier in Q-learning ($CaseI$) and reaching full cooperators sooner in the fermi update rule ($CaseII$). Results (a) are shown for $K = 0.1$ and $L = 200$. Results (b) are shown for $\gamma = 0.8$, $\alpha = 0.8$ and $L = 200$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

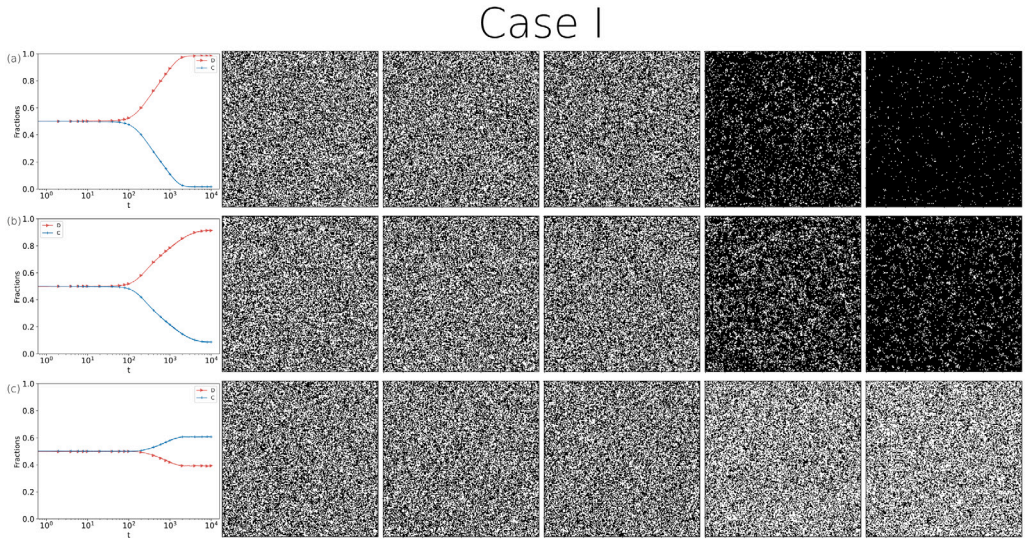


Fig. 3. Time evolution curves of cooperators (blue), defectors (red) over time and snapshots of cooperators (white) and defectors (black). Agents are randomly assigned to cooperators or defectors. Snapshots are plotted over time (from left to right, $t = 1, 10, 100, 1000$ and 10000). The evolution of Q-learning in the illustration is very stable and evolves over time before it starts to diverge. The snapshots exhibit no network reciprocity. Results are shown for $r = 3.6, 3.8, 5.0$, $\gamma = 0.8$, $\alpha = 0.8$ and $L = 200$, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To compare the evolutionary invasion process between the two, we examined the evolution of cooperators and defectors in a pairwise semi-distribution, as shown in Fig. 6. We placed all defectors at the top of the grid and all cooperators at the bottom. With this setup, the invasion dynamics of cooperators and defectors under Q-learning and the Fermi updating rule can be observed more clearly. In Q-learning, cooperators situated within clusters may try out the betrayal strategy due to the exploratory mechanism. This is reflected not only in individual behavioral transitions but also in the gradual change of the overall strategy distribution. This exploration enriches the strategy space and reduces homogeneity, making the boundaries between cooperators and defectors less distinct. Each agent bases its behavior more on environmental feedback than direct neighbor influence, reducing direct spatial effects. This results in interspersed cooperators and defectors, lacking obvious aggression and clustering. In contrast, the Fermi update rule shows more clearly aggression and confrontation between cooperators and defectors, with cooperators forming clusters to protect themselves from defectors. This mutual aggression and protection is visible in the

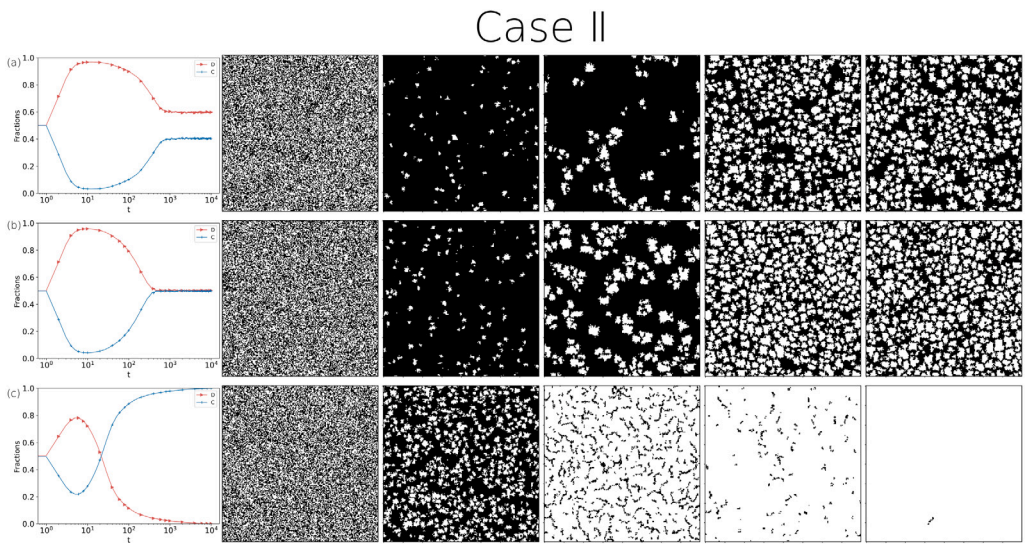


Fig. 4. Time evolution curves of cooperators (blue) and defectors (red) over time and snapshots of cooperators (white) and defectors (black). Agents are randomly assigned to cooperators or defectors. Snapshots are plotted over time (from left to right, $t = 1, 10, 100, 1000$ and 10000). The evolution of the Fermi update rule in the illustration usually starts with the defector's dominance and evolves to the gradual dominance of the cooperator at higher r. The snapshots exhibit strong network reciprocity. Results are shown for $r = 3.6, 3.8, 5.0$, $L = 200$, $K = 0.1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

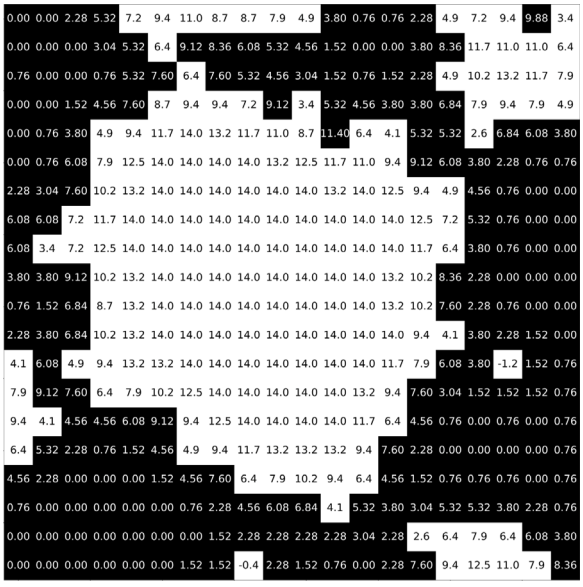


Fig. 5. The figure shows clusters drawn from a snapshot of cooperators (white) and defectors (black) when $r = 3.8$, $\gamma = 0.8$, $\alpha = 0.8$, $L = 200$, $K = 0.1$. The numbers therein represent payoffs. The core members of the cooperative cluster accrue significant payoffs. At the margins, cooperators' returns mirror those of defectors. Defectors who are not at the edge of the cooperator cluster have zero payoff.

figure, reflecting the essential difference under the two evolutionary dynamics.

Comparing Q-learning and the Fermi update rule reveals that the Fermi update rule allows cooperators to form clusters, offering better protection. The Fermi update rule requires less learning time, converges faster, and adapts more quickly to environmental changes. However, agents in the Fermi update rule focuses only on the benefits of the moment in decision making without consideration of experience and future. Q-learning provides better stability, allowing for the earlier emergence and survival of cooperators, and avoiding worse Nash equilibrium.

3.3. Evolutionary games combining learning dynamics and imitation dynamics

The Fermi update rule represents societal propagation, while Q-learning represents learning ability. In this paper, we combine learning dynamics and imitation dynamics to investigate SPGG under the synergy of these two evolutionary dynamics. Fig. 7 illustrates the cooperative evolution of the new evolutionary dynamics as the synergy factor r varies. The new evolutionary dynamics replaces one's own payoff and the neighbor's payoff in the Fermi rule with the Q-value and the neighbor's Q-value. In the Fermi update rule, only the last step's payoff is compared with the neighbor's payoff. In the new dynamics, this payoff is replaced with the Q-table, representing the agent's decision based on both its own and the neighbor's state–action values. The Q-table update formula employs η to discount past experiences, where $1 - \eta$ representing the learning rate. This allows the historical Q-value to decay and gives more influence to the previous step's payoff. Observation of Fig. 7 shows that cooperators can survive with $r = 2.5$ and $r = 3.4$ results in almost full cooperation. The values of r at which cooperators emerge and achieve full cooperation are much lower than those in Q-learning and the Fermi update rule. The rate of increase in the proportion of cooperators as r increases is also much faster. Since the exploration mechanism is eliminated, there is no 0.01 proportion of cooperators retained at smaller r.

To illustrate the initial survival of cooperators and a high percentage of cooperators without full cooperation, we chose time-evolution plots and snapshots for $r = 2.5$, $\frac{25}{9}$ and 3.3, as shown in Fig. 8. Observing The time evolution curves, compared to Q-learning, the new evolutionary dynamics emerge faster during early differentiation, rather than after a long period of learning. Compared to the Fermi update rule, it emerges slower during early differentiation. However, it shows the same initial defector dominance as the Fermi update rule and maintains defector dominance until convergence when r is small. At the right r, there is a shift from initial defector dominance to cooperator dominance. Observing the snapshots, the new evolutionary dynamics could enable cooperators to form clusters that protect them. However, at $r = 2.5$, the clusters originally full of cooperators are gradually invaded, but not completely, preserving some cooperators. At $r = \frac{25}{9}$, clusters of cooperators, initially massive and dominant, are gradually but not completely invaded by defectors from $t = 1000$ onwards. This process creates a homogeneous mixture of cooperators and defectors, resembling a checkerboard. Then, we extract a small portion of this uniform mixture to observe and analyze the reasons for coexistence.

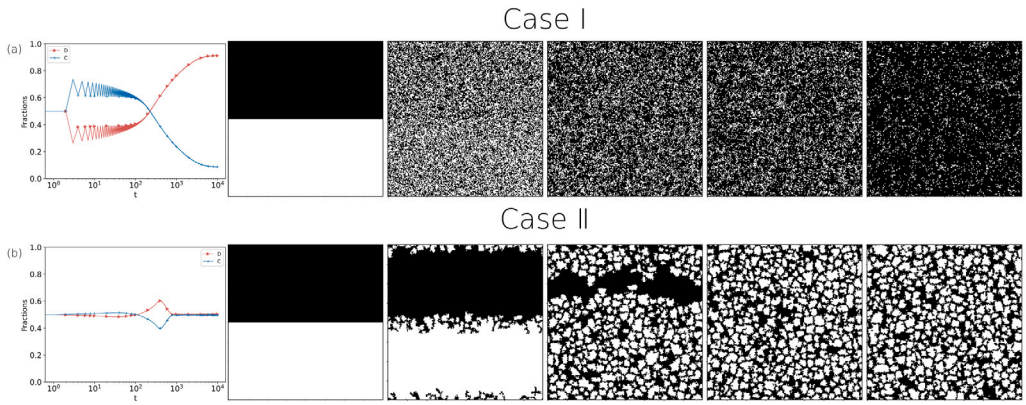


Fig. 6. Time evolution curves of cooperators (blue), defectors (red) over time and snapshots of cooperators (white) and defectors (black). The initial Strategy is the top half for defectors and the bottom half for cooperators. The snapshots are for change over time (from left to right, $t = 1, 10, 100, 1000$ and 10000). The differences between Q-learning and the Fermi update rule in terms of invasion and evolution are shown in the illustration. Results are shown for $r = 3.8$, $L = 200$, $K = 0.1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

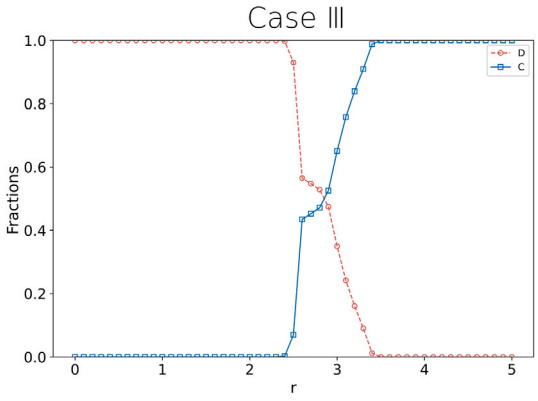


Fig. 7. The figure shows the proportion of cooperators (blue) and the proportion of defectors (red) for different values of r. The illustration shows that cooperators survive earlier and reach full cooperation earlier in the new evolutionary dynamics compared to Q-learning and the Fermi update rule. Results are shown for $\eta = 0.8$, $\gamma = 0.8$, $L = 200$ and $K = 0.5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Stable analysis

$$P_C = P_D = \sum_{g=1}^{G} \Pi_C^g = \sum_{g=1}^{G} \Pi_D^g \tag{8}$$

In a checkerboard-like state, the required r is $\frac{25}{9}$ according to Eq. (8) when cooperators and defectors have the same payoff, with a payoff of approximately 4.44. As shown in Fig. 9, (a) is the snapshots at $r = \frac{25}{9}$ and $T = 100000$, (b) is the uniformly mixed portion centered on defectors and (c) is the uniformly mixed portion centered on cooperators. The checkerboard-like state is a well-known consequence of role-separation. This same role-separation is present in previous work [51–54], and some of it appears to be checkerboard-like. These studies reveal that selfishness and fraternity, extortion strategies, anti-coordination mechanisms, and spatial structures promote cooperation while resisting the spread of betrayal and to some extent enable co-existence between roles. These works provide some explanation for this situation. In this paper, this distribution allows the two parties to maintain some degree of spatial separation, with cooperators willing to make sacrifices to maintain the cooperation, and the defectors taking advantage of the cooperators' sacrifices without completely destroying the structure of cooperation. When the distribution of strategies reaches equilibrium, both cooperators and defectors tend to remain in positions that maximize their gains. At the same time, agents achieve equilibrium in gains through interactions. Even this distribution achieves the highest population gains. We calculate the payoffs of the second-order neighbors of the central agent for one step of the game. These payoffs are correlated with the number of cooperators in the second-order neighbors. Defectors have four second-order cooperative neighbors. However, cooperators have nine second-order cooperator neighbors, which compensates for their disadvantage in payoffs. The spatial distribution pattern allows for relatively stable and similar payoffs for cooperators and defectors making the mixed interior relatively stable. Defectors at the edge of cooperator clusters have more second-order neighbor cooperators, resulting in higher payoff, allowing them to invade cooperator clusters and dominate the homogeneous mixture. The difference from previous work is that this paper uses a combination of Learning Dynamics and Imitation Dynamics updates without changing the public goods game model. The superposition of the indirect influence of neighbors on agents in Learning Dynamics and the direct influence of neighbors on agents in Imitation Dynamics enhances the influence of the environment on the agent's strategy and takes the benefits of surrounding neighbors into account in one's decision.

However, this board state is semi-stable, with agents frequently switching between cooperation and betrayal, as seen in Fig. 9(b) and Fig. 9(c). This behavior has been referred to as CDC in previous studies with a high percentage in the Q-learning. As shown in Fig. 10, CDC behavior constitutes a large percentage in Q-learning. In Fig. 9, because the two strategies have different ranges of r from surviving to full cooperation, we chose the most representative synergy factor rather take than the same. In the new evolutionary dynamics, behavior is more prevalently observed under appropriate r. There are three reasons for this in terms of strategy and structure: First, Q-learning suffers from the problem of CDC, and treating the conversion strategy as a separate action exacerbates the agent's tendency to switch strategies. This phenomenon is inherited in the new evolutionary dynamics. Second, the board structure, surrounded by different strategies, aggravates the conversion of strategies. Third, having only two strategies promotes frequent strategy switching. More strategies can provide agents with more choices, reducing the frequency of switching between just two options. Szabó G et al. [52] considered that the system tends to transform into a reversed board-like arrangement when an agent modifies its strategy. If all agents can change strategies, the iteration of individual strategy changes create a composite structure, exhibiting one of two symmetrical ordered designs. This is consistent with the phenomena observed with the synchronized update strategy in this paper.

In the Q-learning, an agent chooses behaviors or strategies in various states by constructing a Q-table. Therefore, we use the Q-table to

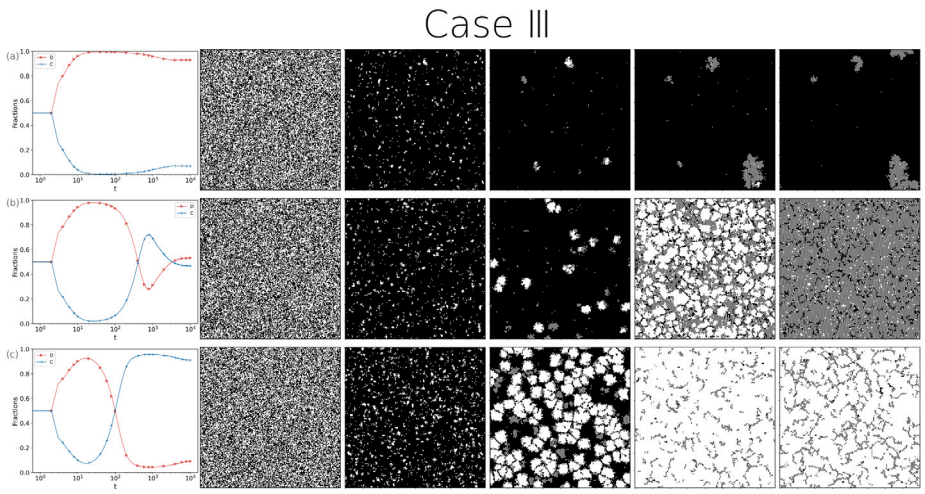


Fig. 8. Time evolution curves of cooperators (blue), defectors (red) over time and snapshots of cooperators (white) and defectors (black). Agents are randomly assigned to cooperators or defectors. Snapshots are plotted as changes over time (from left to right, $t = 1, 10, 100, 1000$ and 10000). The illustration shows the clustering of Q-learning in comparison to the new evolutionary dynamics. In particular, checkerboard-like states appear in the snapshots. Results are shown for $r = 2.5$, $\frac{25}{9}$ and 3.3, $\eta = 0.8$, $\gamma = 0.8$, $L = 200$ and $K = 0.5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

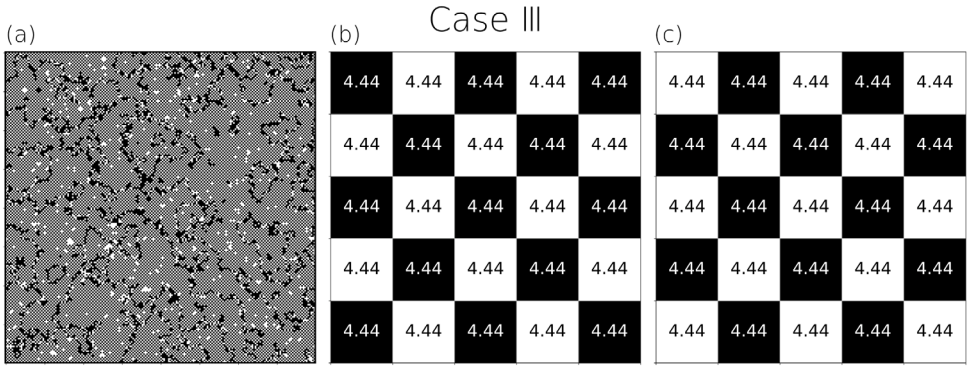


Fig. 9. Illustrated are snapshots of cooperators (white) and defectors (black) (a), and two 5*5 snapshots taken from the 'checkerboard' mounted portion from which the betrayers and cooperators are evenly mixed. (b) centered on defectors and (c) centered on cooperators. The numbers therein represent payoffs. Results are shown for $r = \frac{25}{9}$, $\eta = 0.8$, $\gamma = 0.8$, $L = 200$ and $K = 0.5$.

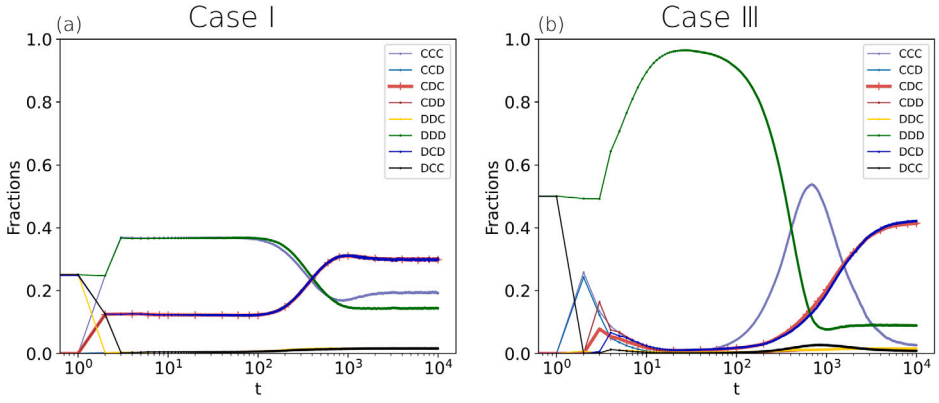


Fig. 10. The figure shows the fractions of strategy for three consecutive steps of the agent as t varies. Fig. 10(a) is the result for Q-learning at $r = 4.7$. Fig. 10(b) is the result for the new evolutionary dynamics at $r = \frac{25}{9}$. A large number of CDC strategies (both CDC and DCD) have emerged during the evolutionary process. This represents that agents are constantly in the process of strategy switching. Note that the CDC and DCD curves in Fig. 10(a) almost overlap, and we distinguish them by thickening the CDC curve and setting the point to '+'.

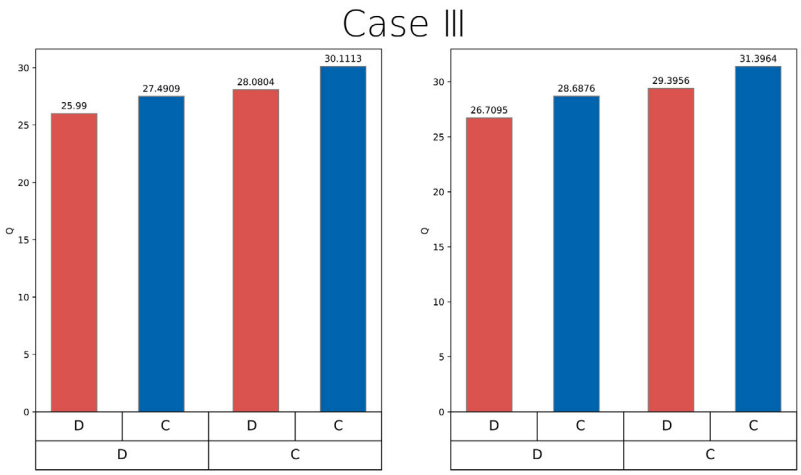


Fig. 11. The bar graphs show the average Q-values for D and C in the converged state of agents. The four bars correspond to the four Q-table values. From left to right, they are $Q_{D,D}$, $Q_{D,C}$, $Q_{C,D}$ and $Q_{C,C}$. The red bar represents the action as Defector, and the blue bar represents the action as Cooperation. Results are shown for $r = \frac{25}{9}$, $\eta = 0.8$, $\gamma = 0.8$, $L = 200$ and $K = 0.5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

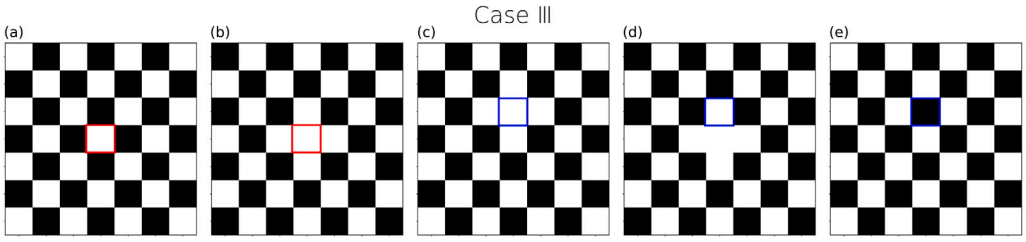


Fig. 12. The figure shows some of the states of 7×7 extracted from the snapshots that vary with t. Fig. 12(a) through Fig. 12(e) show a round of cycling from a 'checkerboard' state with a uniform mix of defectors, and cooperators, to a return to a 'checkerboard' state. The numbers therein represent payoffs. Red indicates agnet1, blue indicates agent2. Results are shown for $r = \frac{25}{9}$, $\eta = 0.8$, $\gamma = 0.8$, $L = 200$ and $K = 0.5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

understand the value of the agent's state action and the effect of the state action value on the agent's decision making. In Fig. 11, we plot the mean Q-values of defectors and cooperators in agents' steady state for $r = \frac{25}{9}$. Observe that $Q_{C,C} > Q_{C,D} > Q_{D,C} > Q_{D,D}$ for the overall Q-values of both cooperators and defectors. Under the new evolutionary dynamics, agents' strategy does not simply select the largest Q-value in the current state. The strategy selection is also influenced by the neighbors, so the Q-value size does not directly determine the strategy choice. At $r = \frac{25}{9}$, most agents adopt the CDC strategy, and the Q-table reveals the values of $Q_{C,D}$ and $Q_{D,C}$ are similar. In dynamic equilibrium and inter-switching cases, similar Q-values fluctuate, leading to instability. This is the reason for the frequent switching of strategies.

To explain the small cooperators' clusters in the snapshot graphs, we extract a series of consecutive localized snapshots from the evolutionary game. Fig. 12(a) through Fig. 12(e) illustrate a cycle from a 'checkerboard' state, where defectors and cooperators are evenly mixed, back to a 'checkerboard' state. We call the agent in the fourth row and fourth column agent 1, and the agent in the third row and fourth column agent 2. In Fig. 12(a), agent 1 is in the state of CC, and all other agents are in the state of CDC. In Fig. 12(b), all except agent 1 have undergone a strategy transition. A small cluster of cooperators forms around agent 1. In Fig. 12(b), agent 2 learns the strategy of agent 1 and transitions to the state of CC. In Fig. 12(c), agent 2 maintains the strategy of cooperators, forming a small cluster centered on agent 2. In Fig. 12(c), agent 2 adopts the CDC behavior of the neighbors and transitions back to the state of CDC. Because the neighbors are surrounded by cooperators in this step, they remain cooperators in Fig. 12(d). By Fig. 12(e), due to the change in CDC, it reverts to the initial 'checkerboard' state with a homogeneous mix of defectors and cooperators. This series of processes makes the whole structure relatively stable.

3.5. Sensitivity analysis

Finally, we explore the effect of variation in the parameters (η, γ) of the new evolutionary dynamics rule on cooperative behavior, as shown in Fig. 13(a). When $r = 2.9$, the weight of $Q_{s,a}(t)$ in the Q-value is large when the parameter η is small. This indicates that agents prefer past experiences. They are more likely to stick to the current strategy and is not influenced by current payoffs. This is shown in simulation experiments where cooperators who form clusters are more likely to resist defectors' invasions. As η increases, the proportion of cooperators decreases, and agents focus more on immediate benefits and the expected benefits of the next action, making them more susceptible to defector invasions. When γ is small, agents focus more on immediate payoffs, making them more susceptible to defector invasions too. As γ increases, the proportion of cooperators rises, and the weight of the expected payoff from the next move increases, encouraging agents to stick to their strategy. Additionally, there are four points to note. First, when $\eta = 0$, the Q-value will always be 0, meaning the probability of agents learning the neighbor's strategy is always $\frac{1}{2}$. Second, a significant difference occurs when $\gamma = 1.0$ compared to $\gamma = 0.99$. This is because $Q^{\max}_{(s',a')}(t)$ weights are increased to the limit value, making the updated value almost entirely determined by the maximum future payoff, with no discount for future payoffs. Consequently, the accumulated future payoff may become infinite, resulting in a continuous increase in the Q-value, which will not converge. When $\gamma = 1$, over many iterations, its decay is determined solely by η, and a change of 0.01 makes a significant difference during long-term accumulation. Third, when $\eta = 1$ and $\gamma = 0$, the Q-value is determined only by the payoff from the previous step, causing the algorithm to degenerate into the Fermi update rule. Finally, the proportion of cooperators gradual

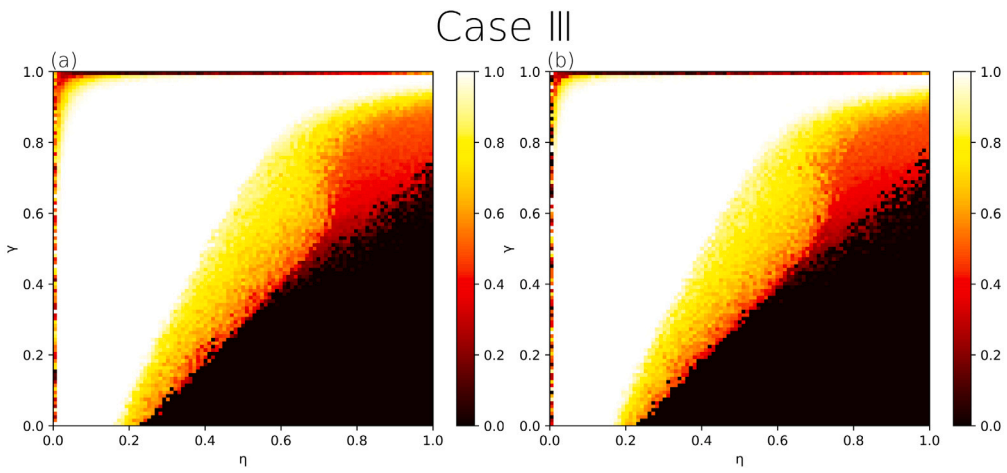


Fig. 13. Fig. 13(a) shows the heat map obtained by changing η and γ. Fig. 13(b) shows the heat map for η and γ obtained by varying the conditions for convergence and ending the game instead of simply fixing the number of steps of the game. Results are shown for $r = \frac{25}{9}$, $L = 200$ and $K = 0.5$.

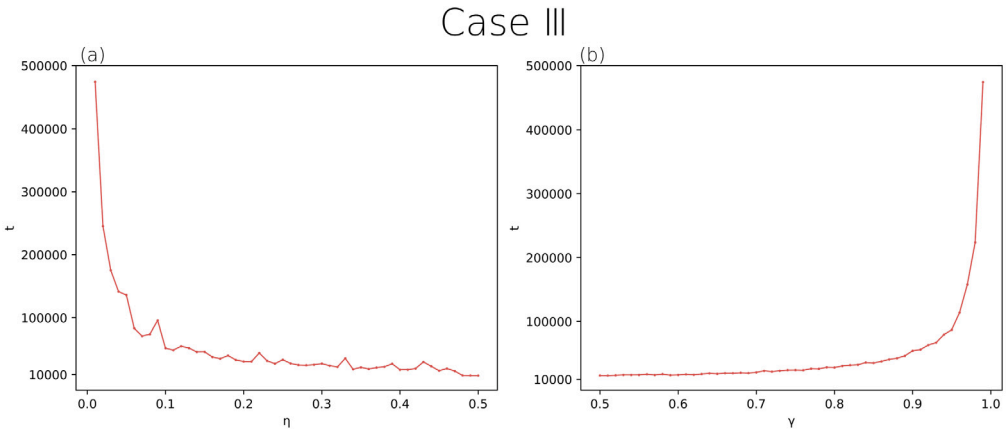


Fig. 14. The figures show the number of steps required for game convergence with η and γ, using stricter convergence conditions and terminating the game under extreme parameter conditions. The figure shows that under extreme parameter conditions, too small a η and too large a γ result in the number of steps required for game convergence becoming too large. Results are shown for $r = \frac{25}{9}$, $L = 200$ and $K = 0.5$.

decrease in the upper left corner of the heat map with η decreases and γ increases. This is due to the gradual slowdown in the convergence rate of the evolutionary dynamics.

We define a strict convergence condition: the variance of the proportion of cooperators from $t - 1000$ to step t must be less than 0.005. Observation of Fig. 13(a) and (b) shows little change in regions except for the upper left corner. In Fig. 13(b), the upper left region is significantly smaller, and the proportion of cooperators is higher. This indicates that the phenomenon in the upper left corner of (a) is due to insufficient game steps. A smaller η and a larger γ increase the number of game steps required for convergence.

Further, we analyze the effect of η and γ values on the number of convergence steps in more extreme parameter cases, as shown in Fig. 14. We define an extremely stringent convergence condition: $\rho_C > 0.99$, and the variance of the proportion of cooperators from $t - 1000$ to t steps is less than 0.005. The simulation experiment stops when this convergence condition is reached. Fig. 14(a) shows the number of steps needed to converge as η varies from 0.01 to 0.5 for $\gamma = 0.99$. Fig. 14(b) shows the number of steps needed to converge as γ varies from 0.5 to 0.99 for $\eta = 0.01$. When $\gamma = 0.99$, the number of iteration steps required for convergence gradually decreases as η increases. When η is small, each update has less effect on the Q-value, requiring more iterations to significantly change the Q-value, resulting in slower convergence. When $\eta = 0.01$, as γ increases, the number of iteration steps required gradually increases. An increase in the value of γ results in a greater emphasis on future expectations. This neglects the value of current payoffs and slows down the convergence rate of the evolutionary dynamics.

4. Conclusion

The research goal of this paper is to investigate learning dynamics and imitation dynamics in SPGG. Firstly, this paper analyzes the performance of agents with learning dynamics represented by Q-learning. Agents iterate the Q-table to form a heterogeneous perceptual payoff matrix during the evolutionary process. Through learning and memorization, it incorporates long-term considerations into agents' decision-making. However, Q-learning indirectly perceives the environment and interacts with information. This results in almost no network reciprocity.

Then, this paper analyzes the differences between learning dynamics and imitation dynamics in SPGG. In the comparison, we have identified two phenomena. The first phenomenon is that under Learning and Memory Dynamics, cooperators are more likely to persist and cooperate in more demanding survival situations. The minimum synergy factors required for Q-learning and the Fermi update rule cooperators to survive are 3.6 and 3.7, respectively. Q-learning requires slightly less. The second phenomenon is that learning does not adapt as quickly as propagation. The memory mechanisms of Q-learning are relatively stable. Indirect information interactions do not adapt well

to rapidly changing environments. As the synergy factor increases, the proportion of the Fermi update rule cooperators rises much faster than Q-learning. Q-learning starts to show significant strategy divergence only around $t = 100$ steps. We further analyze the difference between Q-learning and the Fermi update rule regarding defector invasion. The Fermi update rule shows strong network reciprocity, gradually invading from the edge of the cooperators, while cooperators form clusters to protect themselves. Q-learning has almost no network reciprocity and shows no obvious interaction phenomena, resulting in cooperators and defectors being interspersed. In Q-learning, each agent decides its behavior based on learning from environmental feedback rather than being directly influenced by the neighbors, thus reducing the direct spatial effect.

Further, we combine the learning dynamics, represented by Q-learning, and imitation dynamics, represented by Q-learning, to study new agent behaviors in SPGG. The new evolutionary dynamics enable agents to learn, memorize, and form clusters. Experiments demonstrate that the new evolutionary dynamics can achieve strategy differentiation in the early stage of evolution. Under specific synergy factors, it can facilitate a shift from defector dominance to cooperator prevalence, showing good adaptability and stability of the cooperation ratio. Notably, at $r = \frac{25}{9}$, it forms a uniformly mixed 'checkerboard' structure of cooperators and defectors". This phenomenon reveals a complex interplay between cooperative and competitive strategies. And CDC behavior is widespread in Q-learning. We find that this behavior is inherited in the new evolutionary dynamics. Taking $Q_{D,C}$ and $Q_{C,D}$ as strategies influences agents' strategy switching. Additionally, the existence of only two strategies and the specific network structure aggravate strategy switching in the new evolutionary dynamics.

Sensitivity analyses reveal the significant influence of parameters η and γ in the new evolutionary dynamics. Small η values strengthen agents' reliance on experience and help promote the formation of clusters to defend against defectors. In contrast, large η values make agents more concerned with immediate and expected payoffs, reducing cooperation. Adjusting γ affects the weight of future payoffs. When γ is large, the weight of expected payoffs rises, favoring strategy continuity and increasing cooperation. These parameters significantly affect the speed of convergence and the stability of cooperative behavior. Changes in these parameters affect the balance of short-term benefits, experience, and future expectations in strategy design.

In summary, this study deeply analyzes the different characteristics of learning dynamics and imitation dynamics across multiple dimensions. It also demonstrates how to combine the two to study agents' adaptation and cooperation in complex social interactions through new evolutionary dynamics. The new evolutionary dynamics form a new pattern of coexistence between the two strategies, providing a new perspective for understanding the strategy evolution of agents in complex networks. Although this study has yielded some conclusions, some limitations should be noted in future research. First, this study primarily focuses on the evolution of two basic strategies, whereas decision-making often involves more behavioral options. Second, the network structure used in this study is relatively simplified. The updating formula based on temporal difference in Q-learning provides only a one-dimensional evaluation of the Q-value. Given these limitations, future work can consider more complex models, more diverse networks, and multi-dimensional Q-value evaluations. It is hoped that this paper provides a theoretical basis for future research to explore the combination of multiple evolutionary dynamics and the integration of more complex social mechanisms.

Code availability

The source code is available at https://github.com/Tychema/Learning-And-Propagation.

CRediT authorship contribution statement

Yong Shen: Writing – review & editing, Validation, Funding acquisition. **Yujie Ma:** Writing – original draft, Formal analysis. **Hongwei Kang:** Methodology, Funding acquisition, Conceptualization. **Xingping Sun:** Writing – review & editing, Investigation. **Qingyi Chen:** Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research was supplied by the Open Foundation of Key Laboratory of Software Engineering of Yunnan Province (Grant Nos. 2020SE 308 and 2020SE309) and the new round of "Double First-class" Project of Yunnan University (Grant Nos. CY22624103).

References

[1] Dawes RM, Thaler RH. Anomalies: cooperation. J Econ Perspect 1988;2(3):187–97. http://dx.doi.org/10.1257/jep.2.3.187.
[2] Perc M. Phase transitions in models of human cooperation. Phys Lett A 2016;380(36):2803–8. http://dx.doi.org/10.1016/j.physleta.2016.06.017.
[3] Perc M, Jordan JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. Statistical physics of human cooperation. Phys Rep 2017;687:1–51. http://dx.doi.org/10.1016/j.physrep.2017.05.004.
[4] Pennisi E. How did cooperative behavior evolve. Science 2005;309(5731):93. http://dx.doi.org/10.1126/science.309.5731.93.
[5] Kennedy D, Norman C. What don't we know? Introduction. Science 2005;309(5731):75. http://dx.doi.org/10.1126/science.309.5731.75.
[6] Nowak MA, May RM. Evolutionary games and spatial chaos. Nature 1992;359(6398):826–9. http://dx.doi.org/10.1038/359826a0.
[7] Weibull JW. Evolutionary game theory. Cambridge, Massachusetts, USA: MIT Press; 1997.
[8] Hauert C, Szabó G. Game theory and physics. Am J Phys 2005;73(5):405–14. http://dx.doi.org/10.1119/1.1848514.
[9] Szabó G, Fath G. Evolutionary games on graphs. Phys Rep 2007;446(4–6):97–216. http://dx.doi.org/10.1016/j.physrep.2007.04.004.
[10] Komorita SS. Social dilemmas. 1st ed.. New York: Routledge; 2019.
[11] Nowak MA, May RM. The spatial dilemmas of evolution. Int J Bifurcation Chaos 1993;3(01):35–78. http://dx.doi.org/10.1142/S0218127493000040.
[12] Macy MW, Flache A. Learning dynamics in social dilemmas. Proc Natl Acad Sci USA 2002;99(suppl_3):7229–36. http://dx.doi.org/10.1073/pnas.092080099.
[13] Wang Z, Kokubo S, Jusup M, Tanimoto J. Universal scaling for the dilemma strength in evolutionary games. Phys Life Rev 2015;14:1–30. http://dx.doi.org/10.1016/j.plrev.2015.04.033.
[14] Chen X, Sasaki T, Brännström Å, Dieckmann U. First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. J R Soc Interface 2015;12(102):20140935. http://dx.doi.org/10.1098/rsif.2014.0935.
[15] Santos M. The evolution of anti-social rewarding and its countermeasures in public goods games. Proc Royal Soc B 2015;282(1798):20141994. http://dx.doi.org/10.1098/rspb.2014.1994.
[16] Okada I, Yamamoto H, Toriumi F, Sasaki T. The effect of incentives and meta-incentives on the evolution of cooperation. PLoS Comput Biol 2015;11(5):1–17. http://dx.doi.org/10.1371/journal.pcbi.1004232.
[17] Wu Y, Chang S, Zhang Z, Deng Z. Impact of social reward on the evolution of the cooperation behavior in complex networks. Sci Rep 2017;7:41076. http://dx.doi.org/10.1038/srep41076.
[18] Du C, Jia D, Jin L, Shi L. The impact of neutral reward on cooperation in public good game. Eur Phys J B 2018;91(10):234. http://dx.doi.org/10.1140/epjb/e2018-90052-6.
[19] Szolnoki A, Perc M. Reward and cooperation in the spatial public goods game. Europhys Lett 2010;92(3):38003. http://dx.doi.org/10.1209/0295-5075/92/38003.
[20] Helbing D, Szolnoki A, Perc M, Szabó G. Punish, but not too hard: how costly punishment spreads in the spatial public goods game. New J Phys 2010;12(8):083005. http://dx.doi.org/10.1088/1367-2630/12/8/083005.

[21] Szolnoki A, Szabó G, Perc M. Phase diagrams for the spatial public goods game with pool punishment. Phys Rev E 2011;83(3):036101. http://dx.doi.org/10.1103/PhysRevE.83.036101.
[22] Chen X, Szolnoki A, Perc M. Probabilistic sharing solves the problem of costly punishment. New J Phys 2014;16(8):083016. http://dx.doi.org/10.1088/1367-2630/16/8/083016.
[23] Chen X, Szolnoki A, Perc M. Competition and cooperation among different punishing strategies in the spatial public goods game. Phys Rev E 2015;92(1):012819. http://dx.doi.org/10.1103/PhysRevE.92.012819.
[24] Oya G, Ohtsuki H. Stable polymorphism of cooperators and punishers in a public goods game. J Theoret Biol 2017;419:243–53. http://dx.doi.org/10.1016/j.jtbi.2016.11.012.
[25] Liu J, Meng H, Wang W, Li T, Yu Y. Synergy punishment promotes cooperation in spatial public good game. Chaos Solitons Fractals 2018;109:214–8. http://dx.doi.org/10.1016/j.chaos.2018.01.019.
[26] Dong Y, Hao G, Wang J, Liu C, Xia C. Cooperation in the spatial public goods game with the second-order reputation evaluation. Phys Lett A 2019;383(11):1157–66. http://dx.doi.org/10.1016/j.physleta.2019.01.021.
[27] Milinski M, Semmann D, Krambeck H-J. Reputation helps solve the 'tragedy of the commons'. Nature 2002;415(6870):424–6. http://dx.doi.org/10.1038/415424a.
[28] Fu F, Hauert C, Nowak MA, Wang L. Reputation-based partner choice promotes cooperation in social networks. Phys Rev E 2008;78(2):026117. http://dx.doi.org/10.1103/PhysRevE.78.026117.
[29] Chen M-h, Wang L, Sun S-w, Wang J, Xia C-y. Evolution of cooperation in the spatial public goods game with adaptive reputation assortment. Phys Lett A 2016;380(1–2):40–7. http://dx.doi.org/10.1016/j.physleta.2015.09.047.
[30] Quan J, Zhou Y, Wang X, Yang J-B. Information fusion based on reputation and payoff promotes cooperation in spatial public goods game. Appl Math Comput 2020;368:124805. http://dx.doi.org/10.1016/j.amc.2019.124805.
[31] Shen Y, Yin W, Kang H, Zhang H, Wang M. High-reputation individuals exert greater influence on cooperation in spatial public goods game. Phys Lett A 2022;428:127935. http://dx.doi.org/10.1016/j.physleta.2022.127935.
[32] Griffin C, Belmonte A. Cyclic public goods games: Compensated coexistence among mutual cheaters stabilized by optimized penalty taxation. Phys Rev E 2017;95(5):052309. http://dx.doi.org/10.1103/PhysRevE.95.052309.
[33] Wang S, Liu L, Chen X. Tax-based pure punishment and reward in the public goods game. Phys Lett A 2021;386:126965. http://dx.doi.org/10.1016/j.physleta.2020.126965.
[34] Lee H-W, Cleveland C, Szolnoki A. Supporting punishment via taxation in a structured population. Chaos Solitons Fractals 2024;178:114385. http://dx.doi.org/10.1016/j.chaos.2023.114385.
[35] Cao X-B, Du W-B, Rong Z-H. The evolutionary public goods game on scale-free networks with heterogeneous investment. Phys A 2010;389(6):1273–80. http://dx.doi.org/10.1016/j.physa.2009.11.044.
[36] Liu L, Chen X, Szolnoki A. Competitions between prosocial exclusions and punishments in finite populations. Sci Rep 2017;7:46634. http://dx.doi.org/10.1038/srep46634.
[37] Szolnoki A, Chen X. Alliance formation with exclusion in the spatial public goods game. Phys Rev E 2017;95:052316. http://dx.doi.org/10.1103/PhysRevE.95.052316.
[38] Szabó G, Tőke C. Evolutionary prisoner's dilemma game on a square lattice. Phys Rev E 1998;58(1):69–73. http://dx.doi.org/10.1103/PhysRevE.58.69.
[39] Schuster P, Sigmund K. Replicator dynamics. J Theoret Biol 1983;100(3):533–8. http://dx.doi.org/10.1016/0022-5193(83)90445-9.
[40] Izquierdo LR, Izquierdo SS, Gotts NM, Polhill JG. Transient and asymptotic dynamics of reinforcement learning in games. Games Econom Behav 2007;61(2):259–76. http://dx.doi.org/10.1016/j.geb.2007.01.005.
[41] Lipowski A, Gontarek K, Ausloos M. Statistical mechanics approach to a reinforcement learning model with memory. Phys A 2009;388(9):1849–56. http://dx.doi.org/10.1016/j.physa.2009.01.028.
[42] Jia D, Guo H, Song Z, Shi L, Deng X, Perc M, et al. Local and global stimuli in reinforcement learning. New J Phys 2021;23(8):083020. http://dx.doi.org/10.1088/1367-2630/ac170a.
[43] Wang L, Jia D, Zhang L, et al. Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning. Nonlinear Dynam 2022;108:1837–45. http://dx.doi.org/10.1007/s11071-022-07289-7.
[44] Song Z, Guo H, Jia D, Perc M, Li X, Wang Z. Reinforcement learning facilitates an optimal interaction intensity for cooperation. Neurocomputing 2022;513:104–13. http://dx.doi.org/10.1016/j.neucom.2022.09.109.
[45] Watkins CJ, Dayan P. Q-learning. Mach Learn 1992;8:279–92. http://dx.doi.org/10.1007/BF00992698.
[46] Hasselt H. Double Q-learning. Adv Neural Inf Process Syst 2010;23:2613–21, https://papers.nips.cc/paper/2010/hash/091d584fced301b442654dd8c23b3fc9-Abstract.html.
[47] Han O, Ding T, Bai L, He Y, Li F, Shahidehpour M. Evolutionary game based demand response bidding strategy for end-users using Q-learning and compound differential evolution. IEEE Trans Cloud Comput 2021;10(1):97–110. http://dx.doi.org/10.1109/TCC.2021.3117956.
[48] Shi Y, Rong Z. Analysis of Q-learning like algorithms through evolutionary game dynamics. IEEE Trans Circuits Syst II Express Briefs 2022;69(5):2463–7. http://dx.doi.org/10.1109/TCSII.2022.3161655.
[49] Szolnoki A, Perc M, Szabó G. Topology-independent impact of noise on cooperation in spatial public goods games. Phys Rev E 2009;80:056109. http://dx.doi.org/10.1103/PhysRevE.80.056109.
[50] Szolnoki A, Perc M. Impact of critical mass on the evolution of cooperation in spatial public goods games. Phys Rev E 2010;81:057101. http://dx.doi.org/10.1103/PhysRevE.81.057101.
[51] Szabó G, Szolnoki A. Selfishness, fraternity, and other-regarding preference in spatial evolutionary games. J Theoret Biol 2012;299:81–7. http://dx.doi.org/10.1016/j.jtbi.2011.03.015.
[52] Szabó G, Szolnoki A, Czakó L. Coexistence of fraternity and egoism for spatial social dilemmas. J Theoret Biol 2013;317:126–32. http://dx.doi.org/10.1016/j.jtbi.2012.10.014.
[53] Szolnoki A, Perc M. Evolution of extortion in structured populations. Phys Rev E 2014;89:022804. http://dx.doi.org/10.1103/PhysRevE.89.022804.
[54] Amaral MA, Perc M, Wardil L, Szolnoki A, da Silva Jr EJ, da Silva JKL. Role-separating ordering in social dilemmas controlled by topological frustration. Phys Rev E 2017;95:032307. http://dx.doi.org/10.1103/PhysRevE.95.032307.