Co-Learning of Strategy and Structure Achieves Full Cooperation in Complex Networks with Dynamical Linking

Xiaoqing Fan[1], **Chin-wing Leung**[2] and **Paolo Turrini**[2]
[1]Mathematics of Real World System CDT, University of Warwick
[2]Department of Computer Science, University of Warwick
{xiaoqing.fan, chin-wing.leung, p.turrini}@warwick.ac.uk

Abstract

Social dilemmas are an important benchmark to study the emergence of cooperation among autonomous learning agents and impressive results were recently achieved in two-player games by reinforcement learning agents equipped with a partner selection module. However, the same cannot be said for games on networks. When surrounded by many other defectors, cooperators suffer harsher punishments and find it hard to replicate, making mass defection quickly take over. The frameworks studied so far for the emergence of cooperation in social dilemmas on networks have shown the key role of dynamical linking, the capacity of agents to select their own neighbours, but they have also relied on hard-wired heuristics, such as imitation dynamics, designed to favour cooperation. In this paper, we remove this constraint and study a population of agents that can autonomously learn whether to cooperate or defect with any of their neighbours in a social dilemma, as well as whether to form or sever social ties with others. Building on a seminal framework for the emergence of cooperation in complex social networks with dynamical linking, we implement our agents as Sarsa learners with Boltzmann exploration and equipped with partner selection actions. We show, for the first time, that these agents can reach a fully cooperative society without requiring ad-hoc heuristics. In doing so, we confirm the fundamental role of timescales, the relative speed at which strategy and structure updates occur, for the emergence of cooperation, highlighting the intricate interplay between network dynamics and decision-making in agent societies.

1 Introduction

Social dilemmas are situations where the social benefits of contributing to a common good are overshadowed by the individual incentives of taking advantage of others' contributions. They arise in many real-world domains, such as paying taxes, preserving common resources or investing in green energy and their resolution is essential for tackling many of the global challenges we face today. Designing artificial agents that autonomously learn to promote cooperation in social dilemmas is one of the biggest challenges Cooperative AI [Fatima *et al.*, 2024], with benchmark competitions being held at top AI venues seeking to "evaluate how agents can adapt their cooperative skills to interact with novel partners in unforeseen situations" [Trivedi *et al.*, 2023]. Impressive results were recently obtained with Reinforcement Learning (RL) agents in two-player social dilemmas equipped with a unilateral and single-partner selection module [Anastassacos *et al.*, 2020], which were able to autonomously learn reciprocity through the Tit-for-Tat strategy [Axelrod, 1984]. RL agents were also able to show reciprocity through partner selection strategies, such as the Out-for-Tat rule [Zheng *et al.*, 2017], outcasting defectors and keeping ties with cooperators, which was co-learnt in combination with Tit-for-Tat to sustain a cooperative society [Leung and Turrini, 2024].

While breakthroughs were obtained for two-player games, the same cannot be said for games on networks, where agents interact with potentially many neighbours. When surrounded by many other defectors, cooperators suffer harsher punishments and find it hard to replicate, making mass defection quickly take over. The frameworks studied so far have emphasised once again the key role of partner selection [Pacheco *et al.*, 2006] and the corresponding timescales, the ratio at which partner selection and in-game strategies co-occur [Pacheco *et al.*, 2006; Santos *et al.*, 2006a]. However, achieving cooperative societies in such frameworks generally relied on hard-wired heuristics, such as imitation dynamics and rule-based partner selection [Santos *et al.*, 2006a], inherently designed to favour cooperation. All in all, the work on Reinforcement Learning in games on networks has not yet been able to show that cooperation can be obtained with learning alone [Fulker *et al.*, 2021; Leung *et al.*, 2024].

Contribution In this paper we show, for the first time, that a fully cooperative society can be sustained in games on networks by RL agents only, without relying on hard-wired decision-making heuristics or imitation dynamics. To do so, we build on the seminal model by Pacheco, Traulsen and Nowak [Pacheco *et al.*, 2006] for social dilemmas on complex networks with dynamical linking, where agents are equipped with fixed probabilities to form and sever links with others, and choose whether to cooperate or defect by imitating their

*Code available at https://github.com/Xiaoq-Fan/DynamicLink.

most successful neighbour. We replace such heuristics with the Sarsa algorithm under Boltzmann action selection. We show that our agents are able to reach a fully cooperative society and identify the in-game and partner selection strategies that are simultaneously co-learnt to support it, showcasing the evolution of the underlying network structure. By examining the dominant policies at different learning phases, we show how agents are able to come up with cooperation-inducing strategies such as the Out-for-Tat (OFT), which outcasts defectors and keeps ties with cooperators, and the Tit-for-Tat (TFT) strategy, which copies the opponent's last strategy, in the Prisoner's Dilemma game - the hardest of the social dilemmas. We also confirm the fundamental role of timescales, the relative speed at which strategic and structural updates occur, highlighting the intricate interplay between network dynamics and decision-making in structured agent societies.

2 Related Work

Understanding what makes self-interested individuals cooperate is a key question for many fields of science, such as economics, evolutionary and social psychology and biology [Nowak, 2006]. Various studies, empirical [Rand *et al.*, 2011; Wang *et al.*, 2012; Zhang *et al.*, 2016] and theoretical [Segbroeck *et al.*, 2009; Zheng *et al.*, 2017; Bara *et al.*, 2022; Santos *et al.*, 2006a], have shown how the capacity of individuals to choose reliable partners is key for this to happen.

The interplay between strategic and structural updates naturally induces a two-dimensional timescale, the ratio of which will determine the resulting cooperation rates. This was shown in models with interaction propensity [Santos *et al.*, 2006a; Santos *et al.*, 2006b], optional social dilemmas [Zhang *et al.*, 2016; Zheng *et al.*, 2017], group selection [Santos *et al.*, 2006a], unilateral [Bara *et al.*, 2022] and bilateral [Wang *et al.*, 2012] attachment.

Agent-based simulation models abound for the emergence of cooperation in social dilemmas, e.g.,[Gilbert, 1995; Salazar *et al.*, 2011; Santos and Pacheco, 2005], as well as coordination games on networks, e.g., [Segbroeck *et al.*, 2010], with agents generally driven by heuristic rules. Mechanisms such as reputation were also studied, as a way to isolate and punish defectors [Sabater and Sierra, 2002; Pujol *et al.*, 2002; Perreau de Pinninck *et al.*, 2010; Santos *et al.*, 2018].

Reinforcement learning has recently emerged as the main framework to model and analyse the equilibrium behaviour of self-interested agents and the study of cooperation in social dilemmas is no exception. Breakthroughs were made in the analysis of common pool resources [Pérolat *et al.*, 2017] and partner selection was shown to be a key mechanism for the emergence of cooperation in two-player social dilemmas [Anastassacos *et al.*, 2020] and in two-player optional social dilemmas [Leung *et al.*, 2024]. While the work by [Anastassacos *et al.*, 2020] focuses on unilateral single-partner selection with full knowledge of everyone's last action, our approach studies ties obtained by mutual consent with potentially multiple partners, under more realistic observability constraints.

In games on networks, [Fulker *et al.*, 2021] modelled the co-evolution of network weights, representing individuals' openness to interact, and in-game strategies through imitation, while in [Foley *et al.*, 2018] a co-evolutionary model is presented where strategy and structure evolve by reinforcement learning, but only able to account for the emergence of conventions, while social dilemmas require more complex learning approaches. A recent attempt using RL for partner selection [Leung *et al.*, 2024] was able to retrieve a fully cooperative society building on [Santos *et al.*, 2006a] but still relying on imitation dynamics at the game level.

Our paper stems from a seminal contribution by Pacheco, Traulsen and Nowak, developed in the field of computational physics [Pacheco *et al.*, 2006]. In their framework agents have a propensity to form new links with others which are severed through a set death rate, while imitation dynamics drive the evolution of the in-game strategy. When these parameters are properly set, the network structure will evolve to make cooperators prevail gain higher fitness and take over the entire population. The authors also demonstrated the key role of timescales: only when the rate of dynamical linking is fast enough, the population achieve full cooperation.

To show the emergence of cooperation, the above contributions have all relied on hard-wired heuristics or imitation dynamics in either partner or in-game action selection, which we entirely replace using Reinforcement Learning.

Paper Structure Section 3 reviews social dilemmas and the Sarsa algorithm with Boltzmann exploration, while Section 4 describes our co-learning algorithm. The analysis of the emergence of cooperation, the learnt strategies, timescales and network evolution are presented in Section 5. Section 6 discusses potential future directions.

	C	D
C	R, R	S, T
D	T, S	P, P

	C	D
C	$3, 3$	$-1, 5$
D	$5, -1$	$0, 0$

Table 1: Payoff Matrix for the Prisoner's Dilemma. The game is structured such that the inequalities $T > R > P > S$ and $2R > T+S$ hold true. The bi-matrix on the left presents the general payoff matrix, which we instantiate with the one on the right.

3 Preliminaries

3.1 Social Dilemmas on Networks

Social dilemmas are a benchmark model for the emergence of cooperation in a population of autonomous agents, with the Prisoner's Dilemma (PD) serving as the main example and the hardest to deal with. The PD is a 2-player symmetric game where both players can choose to Cooperate (C) or Defect (D). Players receive a payoff based on the game outcome: R (mutual cooperation reward), P (mutual defection punishment), S (sucker's payoff for cooperating with a defector), or T (temptation payoff for defecting against a cooperator). Its payoff matrix is presented in Table 1. While mutual cooperation provides the best collective outcome, defection is the dominant strategy, leading to a Nash equilibrium of mutual defection, encoding the tension between societal and individual interest that is typical of social dilemmas.

When playing on networks, following [Pacheco *et al.*, 2006], we have that at each round of a game a pair of neighbouring agents are randomly selected to play the PD game. Upon receiving the respective rewards as a function of their individual decisions, they will revise their strategy accordingly. We implement this process, as well as the protocols for forming and severing new links, using Sarsa with Boltzmann exploration.

3.2 Sarsa and Boltzmann Exploration

We train our agents using Sarsa [Sutton and Barto, 2018], a model-free, on-policy reinforcement learning algorithm. Each agent learns its policy independently. At each time step, the agent observes the current state of the environment, denoted as $s_t \in S$, where S represents the set of all possible states. The agent then chooses an available action $a_t \in A$ from a set A. We denote the corresponding Q-value as $Q^i(s_t, a_t)$, estimating the expected accumulated discounted reward of choosing action a_t at state s_t. Based on its action and the current state, the agent receives a reward $r_{t+1} \in R$ and transitions to a new state $s_{t+1} \in S$, following the state-transition probability $\mathcal{P}_{ss'} = \mathrm{P}(s_{t+1} = s' | s_t = s, a_t = a)$.

Let $G_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+1+k}$, where $\gamma \in [0, 1]$ and $t \in \{1, ..., T\}$, be the accumulated discounted rewards from an episode of the game. Then the Q-values of the agent are updated as follows:

$$Q^i(s_t, a_t) \leftarrow Q^i(s_t, a_t) + \alpha[G_t - Q^i(s_t, a_t)], \quad (1)$$

where $\alpha \in [0, 1]$ is the learning rate. The policy of agent i is defined using a Boltzmann exploration strategy, as follows:

$$\mathcal{P}^i(a_n) = \frac{e^{MQ^i(s,a_n)}}{\sum_k e^{MQ^i(s,a_k)}}, \quad (2)$$

where $\mathcal{P}(a_n)$ denotes the probability of selecting action a_n, and M is the inverse temperature parameter that controls the exploration-exploitation trade-off. A high temperature drives the agent towards exploration, whereas a low temperature towards exploitation, favouring actions with higher Q-values.

4 Co-Learning of Strategy and Structure

Our model, which we present next, directly extends the one in [Pacheco *et al.*, 2006] to RL-based decision-makers, allowing for the co-learning of strategy and structure in complex networks with dynamical linking. Starting with a complete network, agents update their social ties by forming new links or severing existing ones, while learning what to do in the PD at the same time. The relative frequency at which agents update their social ties versus playing a PD game depends on a timescale ratio (T_s/T_a), where T_a is the timescale for link update and T_s is the timescale for strategy update. When $T_a \ll T_s$, for example, link update is much faster than strategy update, and vice versa. Without loss of generality, we fix $T_s = 1$ and vary T_a to simplify the process.

Algorithm 1 describes our interaction protocol. In each episode, there will be H [1] rounds of selection, a pair of agents are selected to conduct a PD game play with probability $p = \frac{T_a}{1+T_a}$ or perform a dynamical linking otherwise (lines 6-7).

In the case of dynamical linking, a random pair of agents i and j is selected. If they are neighbours in the current network, the agents will decide whether to maintain or sever the link (lines 9-16). Both agents will decide based on the opponent's last action, for (s^i, s^j) representing states agents are in, where $s \in \{DB_C, DB_D\}$, and (a^i, a^j) the actions chosen, where $a \in \{DB_Y, DB_N\}$. Here, DB_Y represents "yes" (break the link), and DB_N indicates "no" (keep the

Algorithm 1 Social Dilemmas with Dynamical Linking

Input: N, H, α, τ, β, p, T

1: Initialize $Agent$ with N, α, τ
2: Initialize $LastActions$ randomly
3: Initialize $Network$ as a complete graph
4: **for** $episode = 1$ to T **do**
5: **for** $round = 1$ to H **do**
6: Draw $x \in [0, 1)$ randomly
7: **if** $x < p = \frac{T_a}{1+T_a}$ **then**
8: Draw i, j from $Network$ randomly
9: **if** $i \in N(j)$ **then**
10: $s^i_{DB} \leftarrow LastActions[j]$
11: $s^j_{DB} \leftarrow LastActions[i]$
12: $a^i_{DB} \leftarrow Agents[i].getAction(s^i_{DB})$
13: $a^j_{DB} \leftarrow Agents[j].getAction(s^j_{DB})$
14: **if** a^i_{DB} **and** $a^j_{DB} ==$ "Y" **then**
15: $Network.addEdge(i, j)$
16: **end if**
17: **else**
18: $s^i_{DL} \leftarrow LastActions[i]$
19: $s^j_{DL} \leftarrow LastActions[j]$
20: $a^i_{DL} \leftarrow Agents[i].getAction(s^i_{DL})$
21: $a^j_{DL} \leftarrow Agents[j].getAction(s^j_{DL})$
22: **if** a^i_{DL} **or** $a^j_{DL} ==$ "Y" **then**
23: $Network.removeEdge(i, j)$
24: **end if**
25: **end if**
26: **else**
27: $s^i_{PD} \leftarrow LastActions[j]$
28: $s^j_{PD} \leftarrow LastActions[i]$
29: $a^i_{PD} \leftarrow Agents[i].getAction(s^i_{PD})$
30: $a^j_{PD} \leftarrow Agents[j].getAction(s^j_{PD})$
31: $r^i_{PD}, r^j_{PD} = playgame(a^i_{PD}, a^j_{PD})$
32: $LastActions[i] \leftarrow a^i_{PD}$
33: $LastActions[j] \leftarrow a^j_{PD}$
34: $Agent[i].updateReward(r^i_{PD})$
35: $Agent[j].updateReward(r^j_{PD})$
36: **end if**
37: **end for**
38: **for each** agent in Agents parallel do **do**
39: agent.train() with equation (1)
40: **end for**
41: **end for**

[1] We control this number such that at different timescale ratios the number of PD games played is constant at 1500, so that the experience agents gain from game play is equivalent.

link). If both agents choose DB_N, the link is preserved; if either chooses DB_Y, the link is severed. If selected agents i, j are not already linked, they need to decide whether to form a link with one another (lines 17-25). Notably, unlike e.g., [Anastassacos *et al.*, 2020], we let both agents decide based on their own last action, as they lack knowledge of each other, thus having $s^i, s^j \in \{DL_C, DL_D\}$ and $a^i, a^j \in \{DL_Y, DL_N\}$. If both agents choose DL_Y, a link is formed between them. Conversely, if either chooses DL_N, they remain unconnected. To prevent agents from becoming completely isolated, the link cannot be severed if it is their last remaining connection. Note that agents do not earn any immediate reward from either forming or severing a link.

When in the PD game play phase (lines 26-36), an agent i is randomly selected among all agents, and then one of its neighbours j is randomly drawn. They will then play a PD game, informed of the opponent's last action, (s^j, s^i), where $s \in \{PD_C, PD_D\}$. Each agent needs to choose between cooperate and defect, (a^i, a^j), where $a \in \{C, D\}$, and receives the corresponding reward, (r^i, r^j).

At the end of each episode, the agents will update their Q-values (lines 38-40). To do so, they employ Sarsa with Boltzmann exploration, with all Q-values initialised to zero. The learning rate is set to $\alpha = 0.05$. The temperature parameter is fixed at $M = 1$, and the discount rate is set to $\gamma = 1$. These parameters guide the learning process, balancing exploration and exploitation to optimise agent performance.

5 Cooperation Through Dynamical Linking

When agents play a social dilemma on networks and are allowed to update their ties based on the dynamical linking mechanism, full cooperation emerges. Figure 1 presents the percentage of outcomes for the PD game and the population rewards across episodes averaged over 200 simulations for a population of 100 agents, which we maintain throughout our experiments. Figure 1a shows that when agents are allowed to perform dynamical linking ($T_s/T_a = 1$), full cooperation emerges with maximised total rewards. Meanwhile, when the agents are not given a chance to adjust their social ties, i.e. under a random matching mechanism, mutual defection quickly dominates the entire population, effectively suppressing cooperative behaviour, as shown in Figure 1b.

5.1 The Four Phases of Learning

With dynamical linking, agents adjust their connections by forming or severing links in response to rewards, and cooperative behaviour emerges as the dominant strategy. This process results in a gradual and sustained increase in the population's overall rewards, as depicted in Figure 1a. Over time, cooperative agents prevail, eventually dominating. Looking into the learning process, we observe four phases:

- Phase 1 (episodes 0 to 500): Agents learn to defect.
- Phase 2 (episodes 500 to 7000): Agents learn that breaking ties with defectors and keeping them with cooperators (Out-for-Tat, OFT) and always forming a new link with strangers (Commit) is a good strategy.
- Phase 3 (episodes 7000 to 10000): OFT strategies become dominant in the population. Meanwhile, agents learn that cooperating with cooperators and defecting with defectors (Tit-for-Tat, TFT) is a good strategy. The number of agents adopting TFT grows sharply.
- Phase 4 (episodes 10000 onwards): We observe the emergence of a stable cooperative society, where all agents employ the Always Cooperate (All-C) or the Tit-for-Tat (TFT) strategies.

5.2 Emergent Strategic Types

By analysing the Q-values for different states, we can categorise the agent's policy into distinct strategic types. For instance, when creating a new link, if the Q-value for action N consistently exceeds that of action Y, regardless of their previous action ($Q(N|DL_C) > Q(Y|DL_C), Q(N|DL_D) > Q(Y|DL_D)$), the agent is classified as following the Isolate strategy. If the Q-value for action N is greater than that for action Y when their previous action was cooperation, but the opposite is true when the opponent defected, i.e., $(Q(N|DL_C) > Q(Y|DL_C), Q(N|DL_D) < Q(Y|DL_D))$, the agent is classified as employing the D-link strategy. Similar classifications can be applied to other states.

Thus, we can categorise agents' strategies into four distinct types for the link-formation decision stage, four different types for the link-breaking decision stage and another four types for the Prisoner's Dilemma (PD) game stage.

For the link-formation decision stage, they are (1) Always-Link (Commit), (2) C-Link where a cooperator agent always establishes links with strangers, whereas a defector agent avoids forming such links, (3) D-Link where the agent who is a defector always form a link with strangers and not form link if the agent is a cooperator, (4) Always-not-link (Isolate).

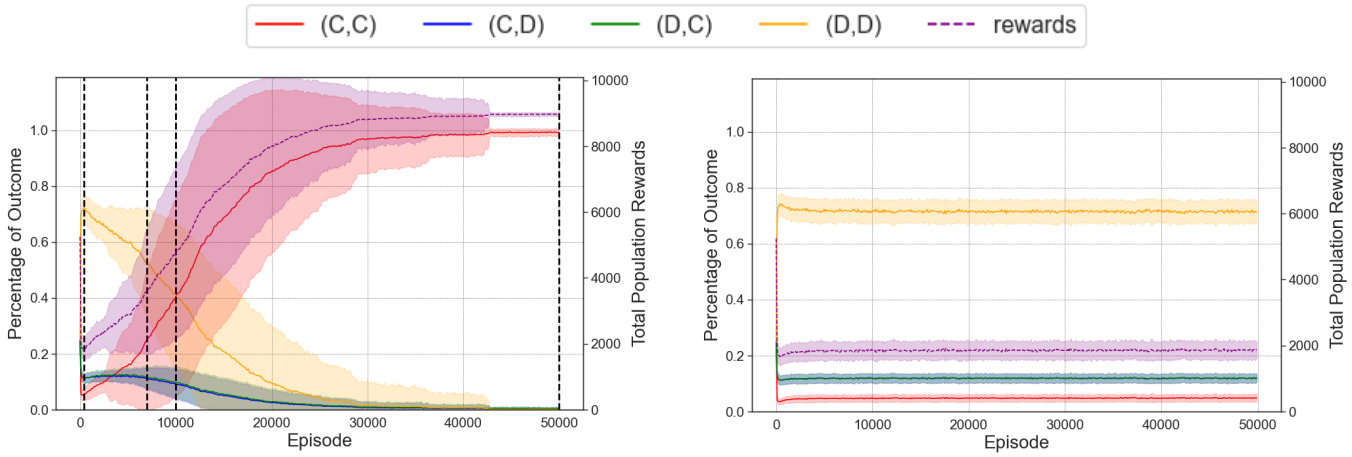
For the link-breaking decision stage, the strategy types are (1) Always-Stay (Stay), (2) Out-for-Tat (OFT), where agents cut the link with defectors and keep the link with cooperators, (3) Reverse-Out-for-Tat (R-OFT), where agents keep the link with defectors and cut the link with cooperators and (4) Always-Leave (Leave).

For the PD game stage, we have (1) Always-Cooperate (All-C), (2) Tit-for-Tat (TFT), where the agent plays cooperate when the opponent plays cooperate and plays defect when the opponent plays defect, (3) Reverse-Tit-for-Tat (R-TFT), where the agent plays cooperate when the opponent plays defect and plays defect when the opponent plays cooperate and (4) Always-Defect (All-D).

To analyse how strategies develop and stabilise within the population, we have created a box plot to show the number of agents that adopted each strategy type during the training process across various learning phases (Figure 2).

In the first phase, agents learn to defect quickly. As for linking, more agents learn to connect with strangers. Since the majority of agents are defectors, the Commit strategy offers a higher likelihood of successful exploitation of others. For the breaking strategy, some agents learn to OFT which leaves their neighbours who are defecting to avoid being further exploited. At this stage, agents generally have not come up with a clear strategy to choose their neighbour, therefore, defection quickly spreads and dominates the population.

In the second phase, as agents continue to adapt, more realise TFT is rewarding. All-D agents are down to 50. For the



(a) Matching in PDs with Dynamical Linking

(b) Random Matching

Figure 1: The mean and standard deviation of the percentage of outcomes in PD games and the population rewards across episodes. The population size is 100, results are averaged over 200 simulations. (a) When agents are allowed to perform dynamical linking ($T_s/T_a = 1$), full cooperation emerges. The vertical dashed lines indicate the end of each learning phase. (b) In random matching, defectors quickly take over the population, effectively suppressing cooperative behaviour.

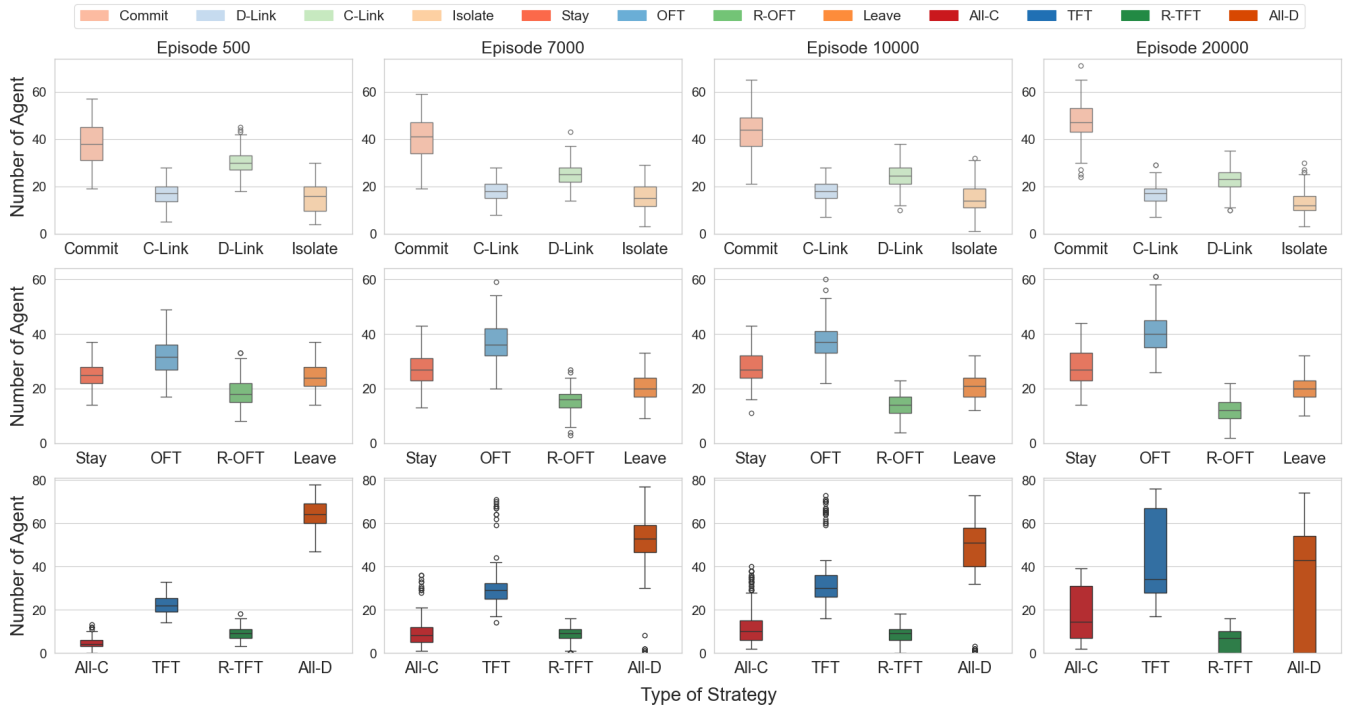


Figure 2: The box plots illustrate the number of agents that adopted each strategy type during the training process across various learning phases. Agents quickly learnt the Commit and OFT strategy which resulted in the development of All-C and TFT. This leads to a 100% cooperation rate in the population.

linking strategy, while defectors continue to learn to actively pursue links with strangers, the number of agents adopting Commit continues to grow. Regarding the breaking strategy, OFT becomes dominant in the population, which keeps the cooperators from being exploited by the defectors.

For the third phase, the number of TFT agents starts to grow significantly. For the linking strategy, more and more agents are willing to link with others. Agents adopting the TFT strategy are able to cooperate with cooperators to get higher rewards and defect with defectors to avoid being exploited. Thus, they are willing to form new connections to increase their chances of being chosen. The rise of TFT agents has significantly impacted defecting agents, encouraging them to resort to cooperative behaviour over time. There-

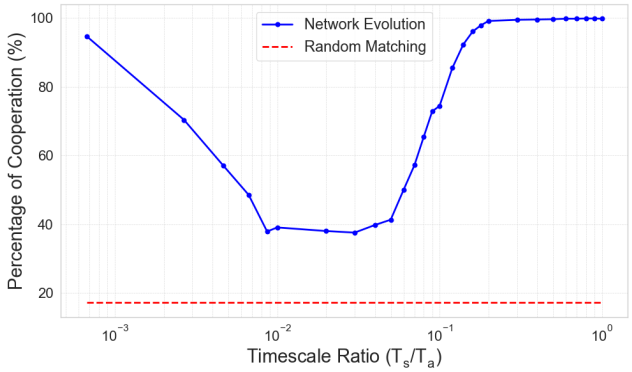


Figure 3: Cooperation rates result at different timescales. For each T_s/T_a ratio, we ran 200 simulations starting with a complete graph of $N = 100$ agents. When $T_s/T_a \geq 0.2$, full cooperation is achieved. When $0 \leq T_s/T_a < 0.2$, a V-shaped pattern is observed.

fore, the number of All-D agents experiences a steep decline.

For the last phase, Always-Cooperate and TFT agents gradually grow and eventually take over the entire population at the end of the simulations. This has led to a steady increase in cooperative outcomes (C, C) in the PD game. By episode $20,000$, more than 80% of agents are cooperating. For the linking strategy, the Commit policy continues to grow. In a cooperative society seeking to link with others benefits everyone, as this increases the chances of being selected for the PD game and generates higher rewards. In the end, we achieved full cooperation in the PD and the population comprised a stable combination of All-C agents and TFT agents.

The influence of TFT agents is evident, as they greatly reduce the effectiveness of the defecting strategy. When agents rely on the ALL D strategy, adopting TFT becomes the optimal response. Yet, TFT requires an agent to learn from long-term rewards, which is hard, as we have demonstrated in the case of random matching. The OFT strategy is far easier to learn as it reflects on the agent's short-term reward. This alleviates the loss of being exploited and gives agents time to discover TFT before becoming defectors. As the number of TFT agents increases, the immediate rewards for defectors drop sharply, further encouraging them to adopt this strategy. Therefore, when agents can choose their neighbours, the TFT strategy establishes itself as a norm, motivating others to follow suit and resulting in a stable cooperative society.

5.3 On the Effect of Timescales

From the previous experiments, we observed the emergence of full cooperation when the structure update and the PD game play operate at the same rate ($T_s/T_a = 1$). In this section, we try and look at the limit of the linking dynamics by reducing the ratio to perform a structure update. By setting the timescale ratio (T_s/T_a) from 0 to 1, we conduct the same experiment using algorithm 1 up to $200,000$ episodes. Figure 3 shows the percentage of cooperation among the population at the end of the simulation across different timescale ratios from 10^{-3} to 10^{0}. Results are averaged over 200 simulations, and the x-axis is presented in log scale.

As shown in Figure 3, the trend in emerging cooperation rates can be distinctly divided into two parts. When the timescale ratio is larger than 10^{-2}, the cooperation percentage increases monotonically from around 40% to 100% as the timescale increases. At the range of 10^{-2} to 6×10^{-2}, the cooperation level remains at 40%, as agents have fewer opportunities to modify their network connections. They are not able to cut ties with defectors fast enough, therefore the majority of agents learn to defect. As the timescale ratio grows, the cooperation level increases sharply. Full cooperation emerges when the timescale ratio reaches 10^{-1}. This indicates that when agents can rapidly adapt their network connections, they are more likely to form a network structure that reinforces mutual cooperation and suppresses defectors. This is fully in line with [Pacheco *et al.*, 2006].

Surprisingly though, when the timescale ratio is less than 10^{-2}, the percentage of cooperation grows gradually from 40% to around 95% as the timescale ratio declines to 10^{-3}. This is an interesting finding, which was not observed in [Pacheco *et al.*, 2006] and related contributions, which, on the other hand, never employed learning agents across the board. The reason for this is that when the timescale ratio becomes very low, the network structure will evolve into a graph with a low average degree. In the case of a timescale ratio of 10^{-3}, for example, only one link is updated per episode. Interestingly, networks with low average degrees favour the emergence of cooperation, as demonstrated in previous research [Santos and Pacheco, 2005]. We will expand on this point in the next section, where we look at network evolution closely.

To conclude, we note that without dynamical linking, i.e., in random matching, the cooperation rate is kept below 20%. The above results highlight the critical role of timescale ratios in shaping the dynamics of cooperation.

5.4 Charting the Network Evolution

In this section, we analyse how the network evolves when agents can perform dynamical linking. Figure 4 presents the evolution of the network structure of a simulation at timescale ratio $T_s/T_a = 1$ for episodes $500, 7000, 10000$ and 20000, which corresponds to the four learning phases. The vertices are coloured with the strategy type identified for the PD game. We can see the population is dominated by defectors in the earlier phase. As learning proceeds, cooperators learn OFT and cut links so that they will not be exploited by defectors. Thus, towards the middle of the simulation, cooperators become more isolated. As more and more agents learned the TFT strategy and interacted with defectors, the expected rewards for defectors dropped, forcing them to switch to consider other strategies. In the end almost all agents have learned the All-C and TFT strategies at episode $20,000$.

In Figure 5, we show the change of average degree for different game strategies across episodes with various timescales. We can observe that the overall mean degree drops dramatically for all timescales during the learning process. In the cases of higher timescales $(0.01, 0.1, 1)$, the average degree for All-C agents visibly drops below other types. This is consistent with the observation in Figure 4 that cooperators tend to isolate themselves. At timescales 1, since all defectors switch to adopt All-C or TFT at the end of simulations, we can see the average degree for All-C rises and

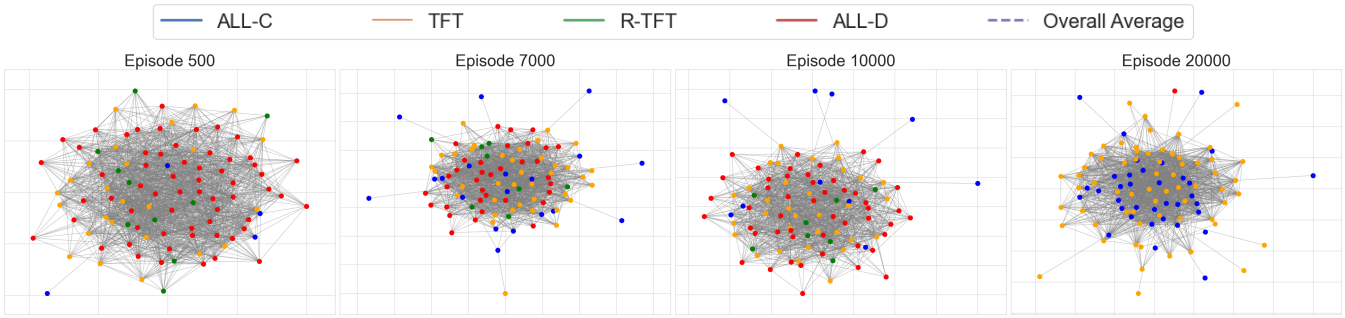


Figure 4: Network structure evolution at timescale ratio $T_s/T_a = 1$ during the learning process. Each node represents the agent and is coloured by the strategy type in PD the agent adopts. And the edges represent the neighbourhood relationships. The population is dominated by defectors at first, causing cooperators to cut their links with them. In the later stage, TFT agents are able to force the defectors to switch to other strategies.

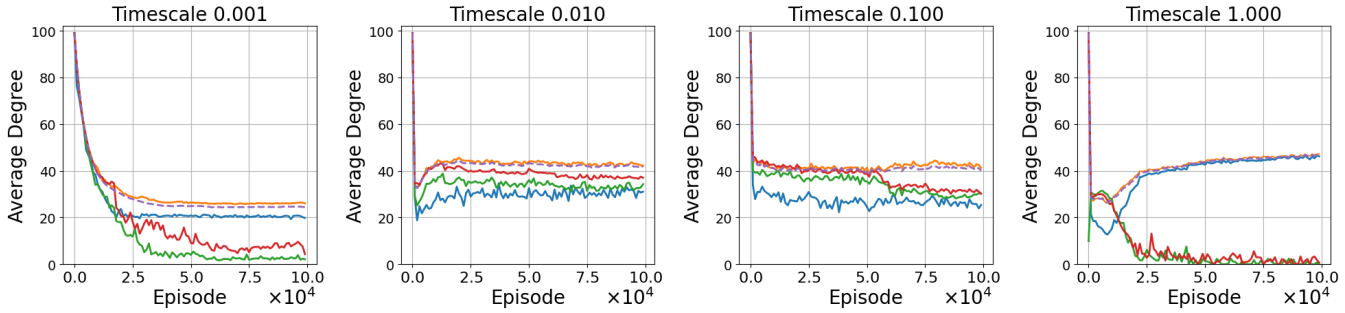


Figure 5: Average degree corresponding to PD strategy types between different timescales across episodes. Displayed results are for timescales 0.001, 0.01, 0.1, and 1 as representative examples. In all situations, the average degree of the network drops sharply in the earlier phase of learning. Later on, the average degree stabilises.

matches that of TFT. For timescales 0.01 and 0.1, where full cooperation does not emerge, the average degree of All-C is kept at a lower value compared to other types. We observe that defectors have fewer and fewer neighbours during the learning, and at a lower timescale (0.01), agents seem not to be learning to get rid of defectors. The persistence of the relatively high average degree of All-D seems to suppress the development of cooperation and result in a low percentage of cooperation at the end of training.

At a very low timescale (0.001), the average degree for the population is significantly lower than at a higher timescale. Agents generally learn to cut the links and be isolated, with TFT agents keeping relatively more connections. The low average degree for All-C and All-D agents separates their connections and this causes the defector to switch their strategy in the end, leading to high cooperation rates.

With random matching (timescale 0), the average degree for any type of agent is kept at 99 throughout the whole simulation, as no one can adjust their ties. This explains why enabling even a faint possibility of dynamical linking is bound to promote (significantly) higher levels of cooperation.

6 Discussion

We studied networks of agents playing a social dilemma with their neighbours, while being able to perform dynamical linking and modify the network structure itself. Building on a seminal model showing that cooperation can emerge when agents are endowed with cooperation-promoting heuristics [Pacheco *et al.*, 2006], we were able to obtain a fully cooperative society relying on RL agents only. We have also identified the co-evolutionary strategies that emerge to support such outcomes, as well as the evolution of the network structure, confirming the fundamental role of timescales in the emergence of cooperation [Santos *et al.*, 2006a; Pacheco *et al.*, 2006; Leung *et al.*, 2024].

Our findings pave the way for numerous opportunities for future research. The next natural step is to study the robustness of the hyperparameters, such as the learning rate, the temperature, and the discount factor, among others. Another important direction is the extension of the study to an n-player Prisoner's Dilemma on social networks, as well as working with games where rewards are based on collective behaviour. The theoretical understanding of the dynamics of two-dimensional timescales remains a big challenge, as spatio-temporal aspects are not considered in the classic replicator equation [Roca *et al.*, 2009], which is tightly linked with the policy gradient dynamics of RL algorithms [Börgers and Sarin, 1997; Tuyls *et al.*, 2003; Bloembergen *et al.*, 2015]. Yet, discovering the analytical solution of restricted instances might shed the needed light on the robustness of cooperation emergence in complex networks with dynamical linking.

Acknowledgments

XF acknowledges the support of the Engineering and Physical Sciences Research Council through the Mathematics of Systems II Centre for Doctoral Training at the University of Warwick [EP/S022244/1]. CL and PT acknowledge the support of the Leverhulme Trust for the Research Grant RPG-2023-050.

References

[Anastassacos *et al.*, 2020] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7047–7054. AAAI Press, 2020.

[Axelrod, 1984] Robert Axelrod. *The Evolution of Cooperation*. Basic, New York, 1984.

[Bara *et al.*, 2022] Jacques Bara, Paolo Turrini, and Giulia Andrighetto. Enabling imitation-based cooperation in dynamic social networks. *Auton. Agents Multi Agent Syst.*, 36(2):34, 2022.

[Bloembergen *et al.*, 2015] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res.*, 53:659–697, 2015.

[Börgers and Sarin, 1997] Tilman Börgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of economic theory*, 77(1):1–14, 1997.

[Fatima *et al.*, 2024] Shaheen Fatima, Nicholas R. Jennings, and Michael J. Wooldridge. Learning to resolve social dilemmas: A survey. *J. Artif. Intell. Res.*, 79:895–969, 2024.

[Foley *et al.*, 2018] Michael Foley, Patrick Forber, Rory Smead, and Christoph Riedl. Conflict and convention in dynamic networks. *Journal of The Royal Society Interface*, 15(140):20170835, 2018.

[Fulker *et al.*, 2021] Zachary Fulker, Patrick Forber, Rory Smead, and Christoph Riedl. Spite is contagious in dynamic networks. *Nature Communications*, 12(1):260, 2021.

[Gilbert, 1995] Nigel Gilbert. Emergence in social simulation. In Nigel Gilbert and Rosaria Conte, editors, *Artificial Societies: The Computer Simulation Of Social Life*. Routledge, 1995.

[Leung and Turrini, 2024] Chin-wing Leung and Paolo Turrini. Learning partner selection rules that sustain cooperation in social dilemmas with the option of opting out. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 1110–1118. ACM, 2024.

[Leung *et al.*, 2024] Chin-wing Leung, Tom Lenaerts, and Paolo Turrini. To promote full cooperation in social dilemmas, agents need to unlearn loyalty. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 111–119. ijcai.org, 2024.

[Nowak, 2006] Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

[Pacheco *et al.*, 2006] Jorge M Pacheco, Arne Traulsen, and Martin A Nowak. Coevolution of strategy and structure in complex networks with dynamical linking. *Physical review letters*, 97(25):258103, 2006.

[Pérolat *et al.*, 2017] Julien Pérolat, Joel Z. Leibo, Vinícius Flores Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3643–3652, 2017.

[Perreau de Pinninck *et al.*, 2010] Adrian Perreau de Pinninck, Carles Sierra, and Marco Schorlemmer. A multi-agent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems*, 21(3):397–424, Nov 2010.

[Pujol *et al.*, 2002] Josep M. Pujol, Ramon Sangüesa, and Jordi Delgado. Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, AAMAS '02, page 467–474, New York, NY, USA, 2002. Association for Computing Machinery.

[Rand *et al.*, 2011] David G. Rand, Samuel Arbesman, and Nicholas A. Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48):19193–19198, 2011.

[Roca *et al.*, 2009] Carlos P Roca, José A Cuesta, and Angel Sánchez. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. *Physics of life reviews*, 6(4):208–249, 2009.

[Sabater and Sierra, 2002] Jordi Sabater and Carles Sierra. Reputation and social network analysis in multi-agent systems. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 1 - AAMAS '02*, New York, New York, USA, 2002. ACM Press.

[Salazar *et al.*, 2011] Norman Salazar, Juan A. Rodriguez-Aguilar, Josep Ll. Arcos, Ana Peleteiro, and Juan C. Burguillo-Rial. Emerging cooperation on complex networks. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '11, page 669–676, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems.

[Santos and Pacheco, 2005] Francisco C Santos and Jorge M Pacheco. Scale-free networks provide a unifying frame-

work for the emergence of cooperation. *Physical review letters*, 95(9):098104, 2005.

[Santos *et al.*, 2006a] Francisco C. Santos, Jorge M. Pacheco, and Tom Lenaerts. Cooperation prevails when individuals adjust their social ties. *PLoS Comput. Biol.*, 2(10), 2006.

[Santos *et al.*, 2006b] Francisco C. Santos, Jorge M. Pacheco, and Tom Lenaerts. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences*, 103(9):3490–3494, 2006.

[Santos *et al.*, 2018] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. Social norms of cooperation with costly reputation building. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4727–4734. AAAI Press, 2018.

[Segbroeck *et al.*, 2009] Sven Van Segbroeck, Francisco C. Santos, Ann Nowé, Jorge M. Pacheco, and Tom Lenaerts. The coevolution of loyalty and cooperation. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009*, pages 500–505. IEEE, 2009.

[Segbroeck *et al.*, 2010] Sven Van Segbroeck, Steven de Jong, Ann Nowé, Francisco C. Santos, and Tom Lenaerts. Learning to coordinate in complex networks. *Adapt. Behav.*, 18(5):416–427, 2010.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Trivedi *et al.*, 2023] Rakshit Trivedi, Akbir Khan, Jesse Clifton, Lewis Hammond, Edgar A Duéñez-Guzmán, Dipam Chakraborty, John P Agapiou, Jayd Matyas, Sasha Vezhnevets, Barna Pásztor, et al. Melting pot contest: Charting the future of generalized cooperative intelligence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[Tuyls *et al.*, 2003] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for q-learning in multi-agent systems. In *The Second International Joint Conference on Autonomous Agents & Multiagent Systems, AAMAS 2003, July 14-18, 2003, Melbourne, Victoria, Australia, Proceedings*, pages 693–700. ACM, 2003.

[Wang *et al.*, 2012] Jing Wang, Siddharth Suri, and Duncan J. Watts. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36):14363–14368, 2012.

[Zhang *et al.*, 2016] Bo-Yu Zhang, Song-Jia Fan, Cong Li, Xiu-Deng Zheng, Jian-Zhang Bao, Ross Cressman, and Yi Tao. Opting out against defection leads to stable coexistence with cooperation. *Scientific reports*, 6:35902, October 2016.

[Zheng *et al.*, 2017] Xiu-Deng Zheng, Cong Li, Jie-Ru Yu, Shi-Chang Wang, Song-Jia Fan, Bo-Yu Zhang, and Yi Tao. A simple rule of direct reciprocity leads to the stable coexistence of cooperation and defection in the prisoner's dilemma game. *Journal of Theoretical Biology*, 420:12–17, 2017.