Coevolutionary networks of reinforcement-learning agents

Ardeshir Kianercy and Aram Galstyan

Phys. Rev. E **88**, 012815 — Published 24 July 2013

DOI: 10.1103/PhysRevE.88.012815

Co-evolutionary Networks of Reinforcement Learning Agents

Ardeshir Kianercy and Aram Galstyan
Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA

This paper presents a model of network formation in repeated games where the players adapt their strategies and network ties simultaneously using a simple reinforcement learning scheme. It is demonstrated that the co-evolutionary dynamics of such systems can be described via coupled replicator equations. We provide a comprehensive analysis for three-player two-action games, which is the minimum system size with non-trivial structural dynamics. In particular, we characterize the Nash Equilibria (NE) in such game, and examine the local stability of the rest points corresponding to those equilibria. We also study general N-player networks via both simulations and analytical methods, and find that in the absence of exploration, the stable equilibria consist of *star* motifs as the main building blocks of the network. Furthermore, in all stable equilibria the agents play pure strategies, even when the game allows mixed NE. Finally, we study the impact of exploration on learning outcomes, and observe that there is a critical exploration rate above which the symmetric and uniformly connected network topology becomes stable.

PACS numbers: 02.50.Le,87.23.Cc,87.23.Ge,05.45.-a

I. INTRODUCTION

Networks depict complex systems where nodes correspond to entities and links encode interdependencies between them. Generally, dynamics in networks is introduced via two different approaches. In the first approach, the links are assumed to be static, while the nodes are endowed with internal dynamics (epidemic spreading, opinion formation, signaling, synchronizing and so on). And in the second approach, nodes are treated as passive elements, and the main focus is on the evolution of network topology.

More recently, it has been suggested that separating individual and network dynamics fails to capture realistic behavior of networks. Indeed, in most real–world networks both attributes of individuals (nodes) and topology of the network (links) evolve in tandem. Models of such adaptive co-evolving networks have attracted significant interest in recent years both in statistical physics [1–5] and game theory and behavioral economics communities [6–11].

To describe coupled dynamics of individual attributes and network topology, here we suggest a simple model of co–evolving network that is based on the notion of interacting adaptive agents. Specifically, we propose network–augmented multi–agent systems where the agents play repeated games with their neighbors, and adapt both their behaviors and the network ties depending on the outcome of their interactions. To adapt, the agents use a simple learning mechanism to reinforce (penalize) behaviors and network links that produce favorable (unfavorable) outcomes. Furthermore, the agents use an action selection mechanism that allows to control exploration/exploitation tradeoff via a temperature-like parameter. We have previously demonstrated [12] that the collective evolution of such a system can be described by appropriately defined replicator dynamics equations. Originally suggested in the context of evolutionary game theory (e.g., see [13, 14]), replicator equations have been used to model collective learning in systems of interacting self–interested agents [15]. Ref. [12] provided a generalization to the scenario where the agents adapt not only their strategies (probability of selecting a certain action) but also their network structure (the set of other agents that play against). This generalization results a system of coupled non-linear equations that describe the simultaneous evolution of agent strategies and network topology.

Here we use the framework suggested in [12] to examine the learning outcomes in networked games. We provide a comprehensive analysis of three-player two-action games, which are the simplest systems that exhibit non-trivial structural dynamics. We analytically characterize the rest-points and their stability properties in the absence of exploration. Our results indicate that in the absence of exploration, the agents always play pure strategies even when the game allows mixed Nash equilibria. For the general N-player case, we find that the stable outcomes correspond to star-like motifs, and demonstrate analytically the stability of a star motif. We also demonstrate the instability of the symmetric network configuration where all the pairs are connected to each other with uniform weights.

We also study the the impact of exploration on the co-evolutionary dynamics. In particular, our results indicate that there is a critical exploration rate above which the uniformly connected network is a globally stable outcome of the learning dynamics.

The rest of the paper is organized as follows: we next derive the replicator equations characterizing co-evolution of network structure and strategies of agents. In Section III we focus on learning without exploration, describe Nash equilibria of the game, and characterize the rest-points of learning dynamics according to their stability properties. We consider the the impact of exploration on learning in Section IV, and provide some concluding remarks in Section V.

II. CO-EVOLVING NETWORKS VIA REINFORCEMENT LEARNING

Let us consider a set of agents that play repeated games with each other. We differentiate agents by indices $x, y, z, \ldots$. At each round of the game, an agent has to choose another agent to play with, and an action from the pool of available actions. Thus, time–dependent mixed strategies of agents are characterized by a joint probability distribution over the choice of the neighbors and the actions.

We assume that the agents adapt to their environment through a simple reinforcement mechanism. Among different reinforcement schemes, here we focus on (stateless) Q-learning [16]. Within this scheme, the agents' strategies are parameterized through, so called Q–functions that characterize relative utility of a particular strategy. After each round of game, the Q functions are updated according to the following rule,

$$Q^i_{xy}(t+1) = Q^i_{xy}(t) + \alpha[R^i_{x,y}(t) - Q^i_{xy}(t)] \qquad (1)$$

where $R^i_{x,y}(Q^i_{x,y})$ is the expected reward (Q value) of agent x for playing action i with agent y, and α is a parameter that determines the learning rate (which can set to $\alpha = 1$ without a loss of generality).

Next, we have to specify how agents choose a neighbor and an action based on their Q-function. Here we use the Boltzmann exploration mechanism where the probability of a particular choice is given as [17]

$$p^i_{xy} = \frac{e^{\beta Q^i_{xy}}}{\sum_{\tilde{y},j} e^{\beta Q^j_{x\tilde{y}}}} \qquad (2)$$

Where, p^i_{xy} is the probability that the agent x will play with agent y and choose action i. Here the inverse *temperature* $\beta \equiv 1/T > 0$ controls the tradeoff between exploration and exploitation; for $T \to 0$ the agents always choose the action corresponding to the maximum Q–value, while for $T \to \infty$ the agents' choices are completely random.

We now assume that the agents interact with each other many times between two consecutive updates of their strategies. In this case, the reward of the i–th agent in Equation 1 should be understood in terms of the *average reward*, where the average is taken over the strategies of other agents, $R^i_{x,y} = \sum_j A^{ij}_{xy} p^j_{yx}$, where A^{ij}_{xy} is the reward (payoff) of agent x playing strategy i against the agent y who plays strategy j. Note that generally speaking, the payoff might be asymmetric.

We are interested in the continuous approximation to the learning dynamics. Thus, we replace $t+1 \to t + \delta t$, $\alpha \to \alpha \delta t$, and take the limit $\delta t \to 0$ in (1) to obtain

$$\dot{Q}^i_{xy} = \alpha[R^i_{x,y} - Q^i_{xy}(t)] \qquad (3)$$

Differentiating 2, using Eqs. 2, 3, and scaling the time $t \to \alpha\beta t$ we obtain the following replicator equation [15]:

$$\frac{\dot{p}^i_{xy}}{p^i_{xy}} = \sum_j A^{ij}_{xy} p^j_{yx} - \sum_{i,j,\tilde{y}} A^{ij}_{x\tilde{y}} p^i_{x\tilde{y}} p^j_{\tilde{y}x} + T \sum_{\tilde{y},j} p^j_{x\tilde{y}} \ln \frac{p^j_{x\tilde{y}}}{p^i_{xy}} \qquad (4)$$

Equations 4 describe the collective adaptation of the Q–learning agents through repeated game–dynamical interactions. The first two terms indicate that the probability of playing a particular pure strategy increases with a rate proportional to the overall goodness of that strategy, which mimics fitness-based selection mechanisms in population biology [13]. The second term, which has an entropic meaning, does not have a direct analogue in population biology [15]. This term is due to the Boltzmann selection mechanism that describes the agents' tendency to randomize over their strategies. Note that for $T = 0$ this term disappears, so the equations reduce to the conventional replicator system [13].

So far, we discussed learning dynamics over a general strategy space. We now make the assumption that the agents' strategies factorize as follows,

$$p^i_{xy} = c_{xy} p^i_x \ , \ \sum\nolimits_y c_{xy} = 1, \ \ \sum\nolimits_i p^i_x = 1. \qquad (5)$$

Here c_{xy} is the probability that the agent x will initiate a game with the agent y, whereas p^i_x is the probability that he will choose action i. Thus, the assumption behind this factorization is that the probability that the agent will perform action i is independent of whom the game is played against. Substituting 5 in 4 yields,

$$\dot{c}_{xy} p^i_x + c_{xy} \dot{p}^i_x = c_{xy} p^i_x \Bigg[\sum_j a^{ij}_{xy} c_{yx} p^j_y - \sum_{i,y,j} a^{ij}_{x,y} c_{xy} c_{yx} p^i_x p^j_y \\ - T \Bigg[\ln c_{xy} + \ln p^i_x - \sum_y c_{xy} \ln c_{xy} - \sum_j p^j_x \ln p^j_x \Bigg] \Bigg] \qquad (6)$$

Next, we sum both sides in Equation 6, once over y and then over i, and make use of the normalization conditions in Eq. 5 to obtain the following co-evolutioanry dynamics of actions and connections probabilities:

$$\begin{aligned} \frac{\dot{p}^i_x}{p^i_x} &= \sum_{\tilde{y},j} A^{ij}_{x\tilde{y}} c_{x\tilde{y}} c_{\tilde{y}x} p^j_{\tilde{y}} - \sum_{i,j,\tilde{y}} A^{ij}_{x\tilde{y}} c_{x\tilde{y}} c_{\tilde{y}x} p^i_x p^j_{\tilde{y}} \\ &+ T \sum_j p^j_x \ln(p^j_x / p^i_x) \qquad (7) \\ \frac{\dot{c}_{xy}}{c_{xy}} &= c_{yx} \sum_{i,j} A^{ij}_{xy} p^i_x p^j_y - \sum_{i,j,\tilde{y}} A^{ij}_{x\tilde{y}} c_{x\tilde{y}} c_{\tilde{y}x} p^i_x p^j_{\tilde{y}} \\ &+ T \sum_{\tilde{y}} c_{x\tilde{y}} \ln(c_{x\tilde{y}} / c_{xy}) \qquad (8) \end{aligned}$$

Equations 7 and 8 are the replicator equations that describe the collective evolution of both the agents' strategies and the network structure.

The following remark is due: Generally, the replicator dynamics in matrix games are invariant with respect to adding any column vector to the payoff matrix. However, this invariance does not hold in the present networked game. The reason for this is the following: if an agent does not have any incoming links (i.e., no other agent plays with him/her), then he always gets a zero reward. Thus, the zero reward of an isolated agent serves as a reference point. This poses a certain problem. For instance, consider a trivial game with a constant reward matrix $a_{ij} = P$. If $P > 0$, then the agents will tend to play with each other, whereas for $P < 0$ they will try to avoid the game by isolating themselves (i.e., linking to agents that do not reciprocate).

To address this issue, we introduce an *isolation payoff* C_I that an isolated agent receives at each round of the game. It can be shown that the introduction of this payoff merely subtracts C_I from the reward matrix in the replicator learning dynamics. Thus, we parameterize the game matrix as follows:

$$a_{ij} = b_{ij} + C_I \tag{9}$$

where matrix B defines a specific game.

Although it is beyond the scope of the present paper, an interesting question is what the reasonable values for the parameter C_I are. In fact, what is important is the value of C_I relative to the reward at the corresponding Nash equilibria, i.e., whether *not playing at all* is better than *playing and receiving a potentially negative reward.* Different values of C_I describe different situations. In particular, one can argue that certain social interactions are apparently characterized by large C_I, where not participating in a game is seen as a worse outcome than participating and getting negative rewards. In the following, we treat C_I as an additional parameter that changes in a certain range, and examine its impact on the learning dynamics.

A. Two-action games

We focus on symmetric games where the reward matrix is the same for all pairs (x, y), $A_{xy} = A$:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{10}$$

Let p_α, $\alpha \in \{x, y, \ldots, \}$, denote the probability for agent α to play action 1 and c_{xy} is the probability that the agent x will initiate a game with the agent y. For two action games, the learning dynamics Eqs. (7) , and (8) becomes:

$$\begin{aligned} \frac{\dot{p}_x}{p_x(1-p_x)} &= \sum_{\tilde{y}} (ap_{\tilde{y}} + b)c_{x\tilde{y}}c_{\tilde{y}x} + T\log\frac{1-p_x}{p_x} && (11) \\ \frac{\dot{c}_{xy}}{c_{xy}} &= r_{xy} - R_x + T\sum_{\tilde{y}} c_{x\tilde{y}} \ln\frac{c_{x\tilde{y}}}{c_{xy}} && (12) \end{aligned}$$

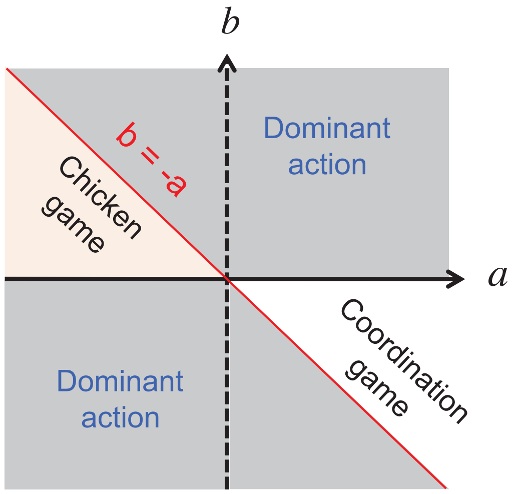


FIG. 1: (Color online) Categorization of 2-action games based on the reward matrix structure in the (a, b) plane.

where

$$\begin{aligned} r_{xy} &= c_{yx}(ap_xp_y + bp_x + dp_y + a_{22}) && (13) \\ R_x &= \sum_{\tilde{y}} (ap_xp_{\tilde{y}} + bp_x + dp_{\tilde{y}} + a_{22})c_{x\tilde{y}}c_{\tilde{y}x} && (14) \end{aligned}$$

Here we have defined the following parameters:

$$\begin{aligned} a &= a_{11} - a_{21} - a_{12} + a_{22} && (15) \\ b &= a_{12} - a_{22} && (16) \\ d &= a_{21} - a_{22} && (17) \end{aligned}$$

The parameters a and b allow a categorization of two action games as follows:

- *Dominant action games:* $-\frac{b}{a} > 1$ *or* $-\frac{b}{a} < 0$
- *Coordination game:* $a > 0, b < 0$ *and* $1 \geq -\frac{b}{a}$
- *Anti-Coordination (Chicken) game:* $a < 0, b > 0$ *and* $1 \geq -\frac{b}{a}$

Before proceeding further, we elaborate on the connection between the rest points of the replicator system for $T = 0$, and the game-theoretic notion of Nash Equilibrium (NE) [1]. For $T = 0$ (no exploration) in the conventional replicator equations, all NE are necessarily the rest points of the learning dynamic. The inverse is not true - not all rest points correspond to NE - and only the stable ones do. Note that in the present model the first statement does not necessarily hold. This is because we have assumed the strategy factorization Eq. 5, due to which

[1] Recall that a joint strategy profile is called Nash equilibrium if no agent can increase his expected reward by *unilaterally* deviating from the equilibrium.

Prisoner's Dilemma	C	D
C	3	0
D	4	2

Coordination Game	S	H
S	6	0
H	3	2

FIG. 2: Examples of reward matrices for typical two-action games.

equilibria where the agents adopt different strategy with different players is not allowed. Thus, any NE that do not have the factorized form simply cannot be described in this framework. The second statement, however, remains true, and stable rest points do correspond to NE.

III. LEARNING WITHOUT EXPLORATION

Fo $T = 0$, the learning dynamics Eqs. 11, 12 attain the following form,

$$\frac{\dot{p}_x}{p_x(1-p_x)} = \sum_{\tilde{y}} (ap_{\tilde{y}} + b)c_{x\tilde{y}}c_{\tilde{y}x} \qquad (18)$$

$$\frac{\dot{c}_{xy}}{c_{xy}} = r_{xy} - R_x \qquad (19)$$

Consider the dynamics of strategies given by Equation 18. Clearly, the vertices of the simplex, $p_x = \{0, 1\}$ are the rest points of the dynamics. Furthermore, in case the game allows a mixed NE, then the configuration where all the agents play the mixed NE $p_x = -b/a$ is also a rest point of the dynamics. As it will be shown below, however, this configuration is not stable, and for $T = 0$, the only stable configurations correspond to the agents playing pure strategies.

A. 3-player games

We now consider the case of three players in two-action game. This scenario is simple enough for studying it comprehensively, yet it still has non-trivial structural dynamics, as we will demonstrate below.

1. *Nash Equilibria*

We start by examining the Nash equilibria for two classes of two-action games, Prisoner Dilemma (PD) and a coordination game [2]. In PD, the players have to choose between *Cooperation* and *Defection*, and the payoff matrix elements satisfy $b_{21} > b_{11} > b_{22} > b_{12}$; see Fig. 2. In two-player PD game, defection is a *dominant* strategy – it always yields a better reward regardless of the other player choice – thus, the only Nash Equilibrium is a mutual defection. And in coordination game, the players have an incentive to select the same action. This game has two pure Nash equilibria, where the agents choose the same action, as well as a mixed Nash equilibrium. A general coordination game reward elements have the relationship $b_{11} > b_{21}, b_{22} > b_{12}$ (see Fig. 2).

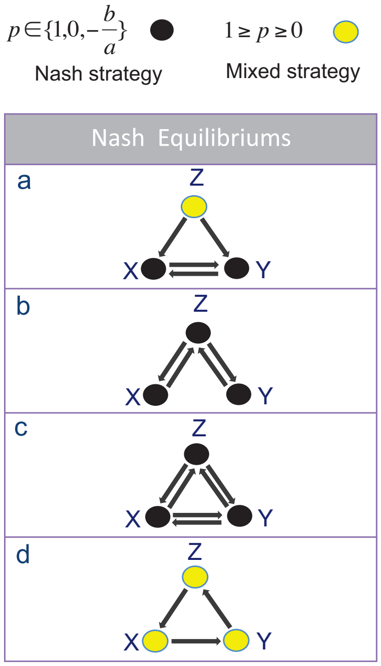


FIG. 3: (Color online) 3-player network Nash equilibria for Prisoner's Dilemma and Coordination game; see the text for more details.

In the 3-agent scenario, a simple analysis yields four possible network topologies corresponding to NE depicted in Fig. 3. In all of those configurations, the agents that are not isolated select strategies that correspond to two-agent NE. Thus, in the case of PD, non-isolated agents always defect, whereas for the coordination game, they can select one of three possible NE. We now examine those configurations in more details.

Configuration I In this configuration, the agents x and y play only with each other, whereas the agent z s isolated: $c_{xy} = c_{yx} = 1$. Note that for this to be a NE, the agents x and y should not be "tempted" to switch and play with the agent z. For instance, in the case of PD, this yields $p_z b_{21} < b_{22}$, otherwise players x, y will be better of linking with the isolated agent z and

[2] The behavior of the coordination and anti-coordination games are qualitatively similar in the context of the present work, so here we do not consider the latter.

exploiting his cooperative behavior [3].

Configuration II In the second configuration, there is a central agent (z) who plays with the other two: $c_{xz} = c_{yz} = 1, c_{zx} + c_{zy} = 1$. Note that this configuration is continuously degenerate as the central agent can distribute his link weight arbitrarily among the two players. Additionally, the isolation payoff should be smaller then than the reward at the equilibrium (e.g., $b_{22} > C_I$ for PD). Indeed, if the latter condition is reversed, then one of the agents, say x, is better off linking with y instead of z, thus "avoiding" the game altogether.

Configuration III: The third configuration corresponds to a uniformly connected networks where all the links have the same weight $c_{xy} = c_{yz} = c_{cx} = \frac{1}{2}$. It is easy to see that when all three agents play NE strategies, there is no incentive to deviate from the uniform network structure.

Configuration IV: Finally, in the last configuration none of the links are reciprocated so that the players do not play with each other: $c_{xy}c_{yx} = c_{xz}c_{zx} = c_{yz}c_{zy} = 0$. This cyclic network is a Nash equilibrium when the isolation payoff C_I is greater than the expected reward of playing NE in the respective game.

2. *Stable rest points of learning dynamics*

The factorized Nash equilibria discussed in the previous section are the rest points of the replicator dynamics. However, not all of those rest points are stable, so that not all the equilibria can be achieved via learning. We now discuss the stability property of the rest points.

One of the main outcomes of our stability analysis is that at $T = 0$ the symmetric network configuration is not stable. This is in fact a more general results that applies to N-agent networks, as is shown in the next section. As we will demonstrate later, the symmetric network can be stabilized when one allows exploration.

The second important observation is that even when the game allows mixed NE, such as in coordination game, any network configuration where the agents play mixed strategy is unstable for $T = 0$ (see Appendix A). Thus, the only outcome of the learning is a configuration where the agents play pure strategies.

The surviving (stable) configurations are listed in Fig. 4. Their stability can be establishes by analyzing the eigenvalues of the corresponding Jacobian. Consider, for instance, the configuration with one isolated player.

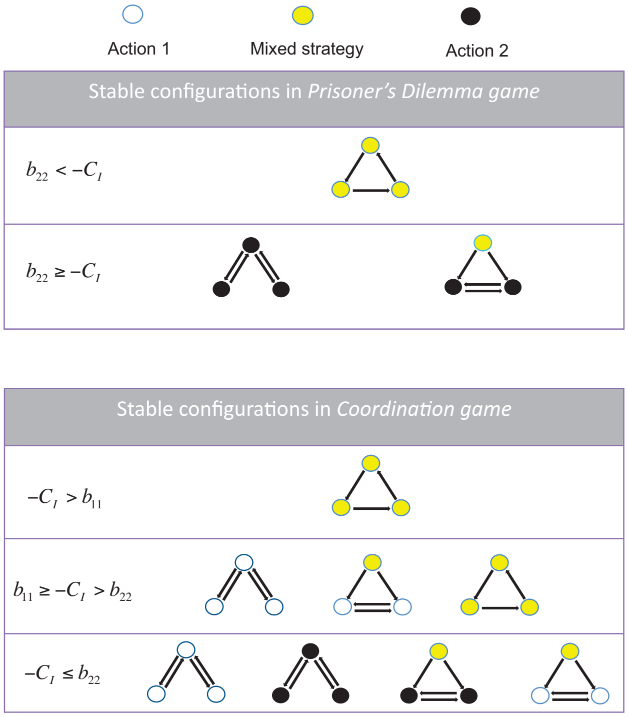

Action 1
Mixed strategy
Action 2
Stable configurations in *Prisoner's Dilemma game*
$b_{22} < -C_I$
$b_{22} \geq -C_I$
Stable configurations in *Coordination game*
$-C_I > b_{11}$
$b_{11} \geq -C_I > b_{22}$
$-C_I \leq b_{22}$

FIG. 4: (Color online) Stable rest points of the learning dynamics for Prisoner's Dilemma (upper panel) Coordination game (lower panel).

The corresponding eigenvalues are

$$\lambda_1 = r_{xz} - r_{xy} \ , \ \lambda_2 = r_{yz} - r_{yx} \ , \ \lambda_3 = 0$$
$$\lambda_4 = (1 - 2p_x)(r_x^1 - r_x^2) < 0 \ ,$$
$$\lambda_5 = (1 - 2p_y)(r_y^1 - r_y^2) < 0 \ , \ \lambda_6 = 0$$

For Prisoner's Dilemma this configuration is marginally stable when agents x, y play defect and $r_{xy} > 0$, $r_{yx} > 0$. It happens only when $b_{22} \geq -C_I$ which means that the isolation payoff should be less than the expected reward for defection. Furthermore, one should also have $r_{xz} < r_{xy}$, $r_{yz} < r_{yx}$, which indicate that the neither x nor y would get better expected reward by switching and playing with z (e.g., condition for NE). And for the coordination game , assuming that $b_{11} > b_{22}$ this configuration is stable when $b_{11} \geq -C_I > b_{22}$, $b_{22} \geq -C_I$.

Similar reasoning can be used for the other configurations shown in Fig. 4. Note, finally, that there is a coexistence of multiple equlibria for range of parameter, except when the isolation payoff is sufficiently large, for which the cyclic (non-reciprocated) network is the only stable configuration.

B. N-player games

In addition to the three agent scenario, we also examined the co-evolutionary dynamics of general N-agent

[3] Note that the dynamics will eventually lead to a different rest point where z is now plays defect with both x and y

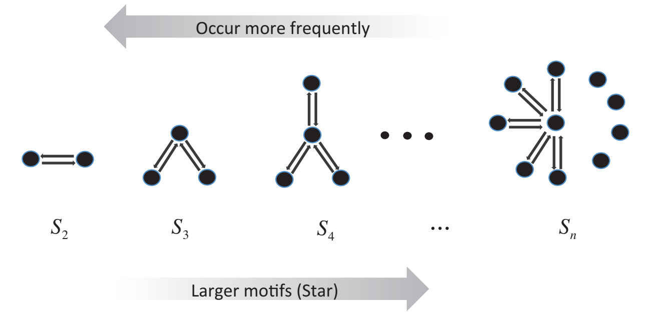


FIG. 5: Observed stable configurations of co-evolutionary dynamics for $T = 0$.

systems, using both simulations and analytical methods. We observed in our simulations that the stable outcomes of the learning dynamic consist of *star* motifs S_n, where a central node of degree $n-1$ connects to $n-1$ nodes of degree 1 [4]. Furthermore, we observed that the basin of attraction of motifs shrinks as motif size grows, so that smaller motifs are more frequent.

We now demonstrate the stability of the star motif S_n in n player two action games. Let player x be the central player, so that all other players are only connected to x, $c_{\alpha x} = 1$. Recall that the Jacobian of the system is a block diagonal matrix with blocks J_{11} with elements $\frac{\partial \dot{c}_{ij}}{\partial c_{mn}}$ and J_{22} with has elements as $\frac{\partial \dot{p}_m}{\partial p_n}$(see Appendix A). When all players play a pure strategy $p_i = 0, 1$ in a star shape motif, it can be shown that J_{22} is diagonal matrix with diagonal elements of form $(1-2p_x)\sum_{\tilde{y}}(ap_{\tilde{y}}+b)c_{x\tilde{y}}c_{\tilde{y}x}$, whereas J_{11} is an upper triangular matrix, and its diagonal elements are either zero or have the form $-(ap_xp_y + bp_x + dp_y + a_{22})c_{xy}$ where x is the central player.

For the Prisoner's Dilemma, the Nash Equilibrium corresponds to choosing the second action (defection) , i.e. $p_\alpha = 0$. Then the diagonal elements of J_{22}, and thus its eigenvalues, equal $bc_{x\tilde{y}}$. J_{11}, on the other hand, has n^2-2n eigenvalues , $(n-1)$ of them are zero and the rest have the form of $\lambda = -a_{22}c_{x\tilde{y}}$. Since for the Prisoner's Dilemma one has $b < 0$ then the start structure is stable as long as $b_{22} > C_I$.

A similar reasoning can be used for the Coordination game, for which one has $b < 0$ and $a + b > 0$. In this case, the star structure is stable when either $b_{11} > -C_I$ or $b_{22} > -C_I$, depending on whether the agents coordinate on the first or second actions, respectively.

We conclude this section by elaborating on the (in)stability of the N-agent symmetric network configuration, where each agent is connected to all the other agents with the same connectivity $\frac{1}{N-1}$. As shown in Appendix B, this configuration can be a rest point of the learning dynamics Eq. (18) only when all agents play the same strategy, which is either $0, 1$ or $-b/a$. Consider now the first block of the Jacobian in Eq. A1, i.e. J_{11}. It can be shown that the diagonal elements of J_{11} are identically zero, so that $Tr(J_{11}) = 0$. Thus, either all the eigenvalues of J_{11} are zero (in which case the configuration is marginally stable), or there is at least one eigenvalue that is positive, thus making the symmetric network configuration unstable at $T = 0$.

IV. LEARNING WITH EXPLORATION

In this section we consider the replicator dynamics for non-vanishing exploration rate $T > 0$. For two agent games, the effect of the exploration has been previously examined in Ref. [18], where it was established that for a class of games with multiple Nash equilibria the asymptotic behavior of learning dynamics undergoes a drastic changes at critical exploration rates and only one of those equilibria survives. Below, we study the impact of the exploration in the current networked version of the learning dynamics.

For 3-player, 2- action games we have six independent variables $p_x, p_y, p_z, c_{xy}, c_{yz}, c_{zx}$. The strategy variables evolve according to the following equations:

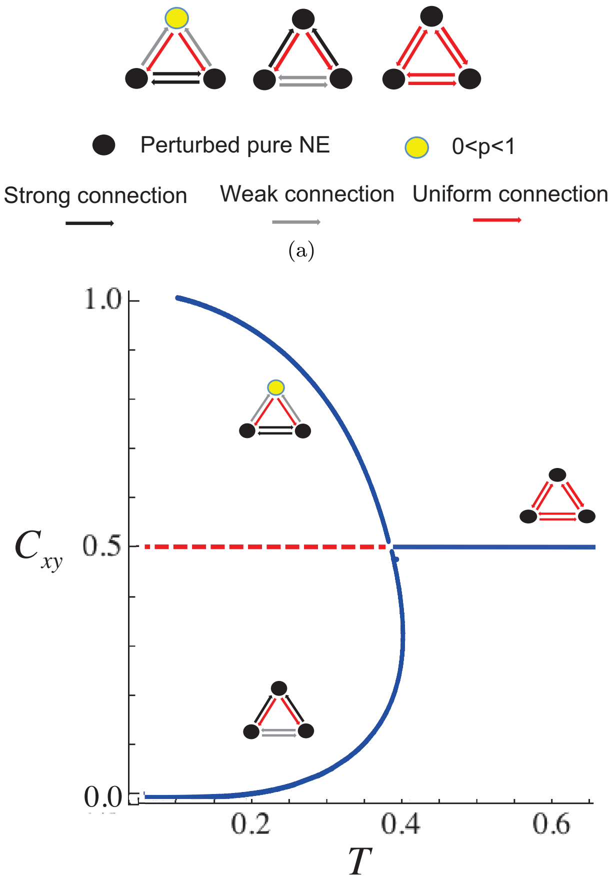
$$\begin{aligned}
\frac{\dot{p_x}}{(1-p_x)p_x} &= (ap_y+b)w_{xy} + (ap_z+b)w_{xz} + T\log\frac{1-p_x}{p_x} \\
\frac{\dot{p_y}}{(1-p_y)p_y} &= (ap_z+b)w_{yz} + (ap_x+b)w_{xy} + T\log\frac{1-p_y}{p_y} \\
\frac{\dot{p_z}}{(1-p_z)p_z} &= (ap_x+b)w_{xz} + (ap_y+b)w_{yz} + T\log\frac{1-p_z}{p_z} \\
\frac{\dot{c}_{xy}}{c_{xy}(1-c_{xy})} &= r_{xy} - r_{xz} + T\log\frac{1-c_{xy}}{c_{xy}} \\
\frac{\dot{c}_{yz}}{c_{yz}(1-c_{yz})} &= r_{yz} - r_{yx} + T\log\frac{1-c_{yz}}{c_{yz}} \\
\frac{\dot{c}_{zx}}{c_{zx}(1-c_{zx})} &= r_{zx} - r_{zy} + T\log\frac{1-c_{zx}}{c_{zx}}
\end{aligned}$$

Here we have defined $w_{xy} = c_{xy}(1-c_{yz})$, $w_{xz} = (1-c_{xy})c_{zx}$, and $w_{yz} = c_{yz}(1-c_{zx})$, and a, b, d are defined in Eqs. 15, 16 and 17.

Fig. 6(a) shows three possible network configurations that correspond to the fixed points of the above dynamics. The first two configurations are perturbed version of a star motif (stable solution for $T = 0$), whereas the third one corresponds to symmetric network where all players connect to the other players with equal link weights.

Furthermore, in Fig. 6(b) we show the behavior of the learning outcomes for a PD game, as one varies the temperature. For sufficiently small T, the only stable configurations are the perturbed star motifs, and the symmetric network is unstable. However, there is a critical value T_c above which the symmetric network becomes globally stable.

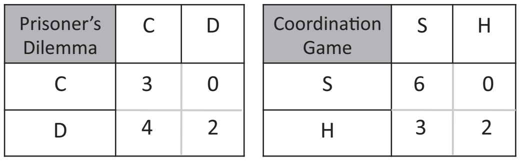
[4] This is true when the isolation payoff is smaller compared to the NE payoff. In the opposite case the dynamics settles into a configuration without reciprocated links.



(a)

(b)

FIG. 6: a) (Color online) Possible network configurations for three-player PD (Fig. 2); (b) Bifurcation diagram for varying temperature. Two blue solid lines correspond to the configurations with one isolated agent and one central agent. The symmetric network configuration is unstable at low temperature (red line), and becomes globally stable above a critical temperature.



Prisoner's Dilemma	C	D
C	3	0
D	4	2

Coordination Game	S	H
S	6	0
H	3	2

FIG. 7: (Color online) Impact of the exploration on the stable outcomes of a coordination game in Fig. (2). The top panel shows the bifurcation of strategy p versus T, whereas the bottom panel shows stability region of the symmetric network configuration in the $C_I - T$ plane. Here the critical temperature is $T_c = 0.36$.

Next, we consider the stability of the symmetric networks. As shown in Appendix B, the only possible solution in this configuration is when all the agents play the same strategy, which can be found from the following equation:

$$(ap + b) = 2T \log \frac{p}{1-p} \tag{20}$$

The behavior of this equation (without the factor 2 in the rhs) was analyzed in details in [18]. In particular, for games with a single NE, this equation allows a single solution that corresponds to the perturbed NE. For games with multiple equlibria, on the other hand, there is a critical exploration rate there is a temperature T_c: For $T < T_c$ there are two stable and one unstable solution, whereas for $T \geq T_c$ there is a single globally stable solution.

We use these insights to examine the stability of the symmetric network configuration for the coordination game, depending on the parameters T and C_I; see Appendix C. In this example $a = 5$, $b = -2$, $d = 1$ for all three agents. Figure 7 shows the bifurcation diagram of p (probability of choosing the first action) plotted versus T. Below the critical temperature, there are three three solutions, two of which (that correspond to the perturbed pure Nash equilibria) are stable. And Fig. (7) shows the domain of T and C_I for stable homogenous equilibrium. When $T \to 0$, the domain of C_I shrinks until it becomes a point at $T = 0$ where $-C_I$ is equal to the NE reward (Fig. (7)).

V. DISCUSSION

We have studied the co-evolutionary dynamics of strategies and link structure in a network of reinforcement learning agents. By assuming that the agents' strategies allow appropriate factorization, we derived a system of a coupled replicator equations that describe the mutual evolution of agent behavior and network topology. We used these equations to fully characterize the stable learning outcomes in the case of three agents and two action games. We also established some analytical results for the more general case of N-player two-action games.

We demonstrated that in the absence of any strategy exploration (zero temperature limit) learning leads to network composed of star-like motifs. Furthermore, the agents on those networks play only pure NE, even when the game allows a mixed NE. Also, even though the learning dynamics allows rest points with a uniform network (e.g., an agent plays with all the other agents with the same probability) , those equilibria are not stable at $T = 0$. The situation changes when the agents explore their strategy space. In this case, the stable network structures undergo bifurcation as one changes the exploration rate. In particular, there is a critical exploration rate above which the uniform network becomes a globally stable outcome of the learning dynamics.

We note that the main premise behind the strategy factorization use here is that the agents use the same strategy profile irrespective of whom they play against. While this assumption is perhaps valid under certain circumstances, it certainly has its limitations that need to be studied further through analytical results and empirical data. Furthermore, the other extreme where the agent employs unique strategy profiles for each of his partners does not seem very realistic either, as it would impose considerable cognitive load on the agent. A more realistic assumption is that the agents have a few strategy profile that roughly correspond to the type of the agent he is interacting with. The approach presented here can be in principle generalized to the latter scenario.

VI. ACKNOWLEDGMENTS

We thank Armen Allahverdyan for his comments and contributions during the initial phase of this work. This research was supported in part by the National Science Foundation under grant No. 0916534 and the US AFOSR MURI grant No. FA9550-10-1-0569.

Appendix A: Local Stability Analysis of the Rest Points

To study the local stability properties of the rest points in the system given by Eqs.18 and 19 , we need to analyze the eigenvalues of the corresponding Jacobian matrix. For n-player two-action game, we have n action variables and $l = n(n-2)$ link variables, so that the total number of independent dynamical variables is $n + l = n(n-1)$. We can represent the Jacobian as follows,

$$J = \begin{pmatrix} \frac{\partial \dot{c}_{ij}}{\partial c_{mn}} & \frac{\partial \dot{c}_{ij}}{\partial p_m} \\ \\ \frac{\partial \dot{p}_m}{\partial c_{ij}} & \frac{\partial \dot{p}_m}{\partial p_n} \end{pmatrix} = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix} \tag{A1}$$

Here the diagonal blocks J_{11} and J_{22} are $l \times l$ and $n \times n$ square matrices, respectively. Similarly, J_{12} and J_{21} are $l \times n$ and $n \times l$ matrices, respectively.

In the most general case, the full analysis of the Jacobian is intractable. However, the problem can be simplified for $T = 0$. Indeed, consider the lower off-diagonal block of the Jacobian, J_{21}, the elements of which have the form

$$\frac{\partial \dot{p}_i}{\partial c_{ij}} = p_i(1-p_i)c_{ji}(ap_i + b) \tag{A2}$$

Consulting the rest point condition given by Eqs. 18, one can see that J_{21} is identically zero. By using the *block matrix determinant* identity, the characteristic polynomial of the Jacobian assumes the following factorized form

$$p(\lambda) = \mathbf{det(J_{11} - \lambda I) det(J_{22} - \lambda I) = 0} \tag{A3}$$

The above factorization facilitates the stability analysis for certain cases that we focus now:

a. (In)Stability of mixed strategies for $T = 0$ Let us show that the configurations where the agents mix either on their actions or links cannot be stable at $T = 0$. Here we just need to consider the submatrix J_{22}. We now show that this matrix always has at least one positive eigenvalues when players adopt the mixed Nash equilibrium $p = -b/a$. Indeed, it can be shown that J_{22} is a non-zero matrix with zero diagonal elements. Recall that for any square matrix A the $Tr(A) = \sum \lambda_i$ then $Tr(J_{11}) = 0$ means at least one of its eigenvalues is always positive, so that the mixed Nash configuration is unstable. The same line of reasoning can be applied to the configuration where the agents mix over the links.

Appendix B: Agent Strategies in Symmetric Networks

Let us consider a 2-action n players game. Each player i chooses action one with probability p_i. Here we prove that player n and player $n-1$ in homogenous network have the same strategy, i.e., $p_n = p_{n-1}$. Consider the Eq. 11 for players n, $n-1$ and $n-2$,

$$p_1 + p_2 + \cdots + p_{n-2} + p_{n-1} = k \log \frac{p_n}{1-p_n} - c \tag{B1}$$

$$p_1 + p_2 + \cdots + p_{n-2} + p_n = k \log \frac{p_{n-1}}{1-p_{n-1}} - c \tag{B2}$$

where

$$K = -\frac{T(n-1)^2}{a} \ , \ c = \frac{b(n-1)}{a} \tag{B3}$$

Also, let us define a function g as

$$g(p_n) = x_n + k \log \frac{p_n}{(1-p_n)} \tag{B4}$$

Now , by subtracting the two Eq. B1 and B2, we have $g(p_n) = g(p_{n-1})$. Since $0 < p_i < 1$, then function g is a monotonic function, so $g(p_n) = g(p_{n-1}) \leftrightarrow pn = p_{n-1}$. By repeating the same reasoning for the remaining p_i one can prove that $p_1 = p_2 = \cdots = p_n$.

Appendix C: Stability of Symmetric 3-player network

For 3-player 2-action games, the Jacobian corresponding to the symmetric network configuration consists of the following blocks:

$$J_{11} = \begin{pmatrix} -T & -v & -v \\ -v & -T & -v \\ -v & -v & -T \end{pmatrix} \tag{C1}$$

$$J_{12} = \begin{pmatrix} 0 & m & -m \\ -m & 0 & m \\ m & -m & 0 \end{pmatrix} \tag{C2}$$

$$J_{21} = \begin{pmatrix} 0 & -g & g \\ g & 0 & -g \\ -g & g & 0 \end{pmatrix} \tag{C3}$$

$$J_{22} = \begin{pmatrix} -T & k & k \\ k & -T & k \\ k & k & -T \end{pmatrix}. \tag{C4}$$

where we have defined

$$v = \frac{ap^2 + bp + dp + b_{22} + C_I}{4} \tag{C5}$$

$$m = \frac{ap + d}{8} \tag{C6}$$

$$g = \frac{p(1-p)(ap+b)}{2} \tag{C7}$$

$$k = \frac{ap(1-p)}{4}. \tag{C8}$$

and p is the probability of selecting the first action, which is the same for all the agents in the symmetric network configuration. The six eigenvalues that determine the stability of the configuration can be calculated analytically and are as follows:

$$\begin{aligned} \lambda_1 &= 2k - T \\ \lambda_2 &= -T - 2v \\ \lambda_{3,4} &= \frac{1}{2}(-k - 2T + v - \sqrt{12gm + (k+v)^2}) \\ \lambda_{5,6} &= \frac{1}{2}(-k - 2T + v + \sqrt{12gm + (k+v)^2}). \end{aligned}$$

These expressions can be used to (numerically) identify the stability region of the configuration in the parameter space (T, C_I), as shown in Fig. 7.

[1] J. M. Pacheco, A. Traulsen, and M. A. Nowak, Physical Review Letters **97**, 258103 (2006).

[2] T. Gross and B. Blasius, Journal of the Royal Society Interface **5**, 259 (2008).

[3] S. Castellano, C.;Fortunato and V. Loreto, Reviews of Modern Physics **81**, 591 (2009).

[4] M. Perc and A. Szolnoki, BioSystems **99**, 109 (2010).

[5] G. Zschaler, The European Physical Journal-Special Topics **211**, 1 (2012).

[6] D. Lazer, Journal of Mathematical Sociology **25**, 69 (2001).

[7] M. Jackson and A. Watts, Games and Economic Behavior **41**, 265 (2002).

[8] G. Demange, *Group formation in economics: networks, clubs and coalitions* (Cambridge Univ. Press, 2005).

[9] S. Goyal and F. Vega-Redondo, Games and Economic Behavior **50**, 178 (2005).

[10] S. Goyal, *Connections: an introduction to the economics of networks* (Princeton University Press, 2009).

[11] M. Staudigl, International Journal of Game Theory **42**, 179 (2013).

[12] A. Kianercy, A. Galstyan, and A. Allahverdyan, in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (2012).

[13] J. Hofbauer and K. Sigmund, *Evolutionary games and Population dynamics* (Cambridge University Press, 1998).

[14] J. Hofbauer and K. Sigmund, Bulletin of the American Mathematical Society **40**, 479 (2003).

[15] Y. Sato and J. Crutchfield, Physical Review E **67** (2003).

[16] C. Watkins and P. Dayan, Machine learning **8**, 279 (1992).

[17] R. Sutton and A. Barto, *Reinforcement learning: An introduction* (The MIT press, 2000).

[18] A. Kianercy and A. Galstyan, Physical Review E **85**, 041145 (2012).