Coupled Replicator Equations for the Dynamics of Learning in Multiagent Systems

Yuzuru Sato[1, 2, *] and James P. Crutchfield[2, †]

[1]Brain Science Institute, Institute of Physical and Chemical Research (RIKEN), 2-1 Hirosawa, Saitama 351-0198, Japan
[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501

(Dated: January 7, 2003)

Starting with a group of reinforcement-learning agents we derive coupled replicator equations that describe the dynamics of collective learning in multiagent systems. We show that, although agents model their environment in a self-interested way without sharing knowledge, a game dynamics emerges naturally through environment-mediated interactions. An application to rock-scissors-paper game interactions shows that the collective learning dynamics exhibits a diversity of competitive and cooperative behaviors. These include quasiperiodicity, stable limit cycles, intermittency, and deterministic chaos—behaviors that should be expected in heterogeneous multiagent systems described by the general replicator equations we derive.

PACS numbers: 05.45.-a, 02.50.Le, 87.23.-n Santa Fe Institute Working Paper 02-04-017

Adaptive behavior in multiagent systems is an important interdisciplinary topic that appears in various guises in many fields, including biology [1], computer science [2], economics [3], and cognitive science [4]. One of the key common questions is how and whether a group of intelligent agents truly engages in collective behaviors that are more functional than individuals acting alone.

Suppose that many agents interact with an environment and each independently builds a model from its sensory stimuli. In this simple type of coupled multiagent system, collective learning (if it occurs) is a dynamical behavior driven by agents' environment-mediated interaction [5, 6]. Here we show that the collective dynamics in multiagent systems, in which agents use reinforcement learning [7], can be modeled using a generalized form of coupled replicator equations.

While replicator dynamics were introduced originally for evolutionary game theory [8], the relationship between reinforcement learning and replicator equations has been developed only recently [9]. Here, we extend these considerations to multiagent systems, introducing the theory behind a previously reported game-theoretic model [10]. We show that replicator dynamics emerges as a special case of the continuous-time limit for multiagent reinforcement-learning systems. The overall approach, though, establishes a general framework for dynamical-systems analyses of adaptive behavior in collectives.

Notably, in learning with perfect memory, our model reduces to the form of a multipopulation replicator equation introduced in Ref. [11]. For two agents with perfect memory interacting via a zero-sum rock-scissors-paper game the dynamics exhibits Hamiltonian chaos [10]. In contrast, as we show here, with memory decay multiagent systems generally become dissipative and display the full range of nonlinear dynamical behaviors, including limit cycles, intermittency, and deterministic chaos.

Our multiagent model begins with simple reinforcement-learning agents. To clarify the development, we assume that there are two such agents X and Y that at each time step take one of N actions: $i = 1, \ldots, N$. Let the probability for X to chose action i be $x_i(n)$ and $y_i(n)$ for Y, where n is the number of the learning iterations from the initial state at $n = 0$. The agents' choice distributions at time n are $\mathbf{x}(n) = (x_1(n), \ldots, x_N(n))$ and $\mathbf{y}(n) = (y_1(n), \ldots, y_N(n))$, with $\Sigma_i x_i(n) = \Sigma_i y_i(n) = 1$.

Let R^X_{ij} and R^Y_{ij} denote the rewards for X taking action i and Y action j at step n, respectively. Given these actions, X's and Y's memories, $Q^X_i(n)$ and $Q^Y_i(n)$, of the past benefits from their actions are governed by

$$\begin{aligned} Q^X_i(n+1) - Q^X_i(n) &= R^X_{ij} - \alpha_X Q^X_i(n) \text{ and} \\ Q^Y_i(n+1) - Q^Y_i(n) &= R^Y_{ij} - \alpha_Y Q^Y_i(n) , \end{aligned} \tag{1}$$

where $\alpha_X, \alpha_Y \in [0, 1)$ control each agent's memory decay rate and $Q^X_i(0) = Q^Y_i(0) = 0$. The agents choose their next actions according to the Q's, updating their choice distributions as follows:

$$x_i(n) = \frac{e^{\beta_X Q^X_i(n)}}{\Sigma_j e^{\beta_X Q^X_j(n)}} \text{ and } y_i(n) = \frac{e^{\beta_Y Q^Y_i(n)}}{\Sigma_j e^{\beta_Y Q^Y_j(n)}}, \tag{2}$$

where $\beta_X, \beta_Y \in [0, \infty]$ control the learning sensitivity: how much the current choice distributions are affected by past rewards. Using Eq. (2), the dynamic governing the change in agent state is given by:

$$x_i(n+1) = \frac{x_i(n) e^{\beta_X (Q^X_i(n+1) - Q^X_i(n))}}{\Sigma_j x_j(n) e^{\beta_X (Q^X_j(n+1) - Q^X_j(n))}} , \tag{3}$$

and similarly for $y_i(n+1)$.

Consider the continuous-time limit corresponding to agents performing a large number of actions (iterates of Eqs. (1)) for each choice-distribution update (iterates of Eq. (3)). In this case, we have two different time scales—that for agent-agent interactions and that for learning. We assume that the learning dynamics is very slow compared to interactions and so $\mathbf{x}$ and $\mathbf{y}$ are essentially constant during the latter. Then, based on Eq. (3), continuous-time learning for agent X is governed by

$$\dot{x}_i = \beta_X x_i (\dot{Q}_i^X - \Sigma_j \dot{Q}_j^X x_j) \ , \tag{4}$$

and for the dynamic governing memory updates we have

$$\dot{Q}_i^X = R_i^X - \alpha_X Q_i^X \ , \tag{5}$$

where R_i^X is the reward for X choosing action i, averaged over Y's actions during the time interval between learning updates. Putting together Eqs. (2), (4), and (5) one finds

$$\frac{\dot{x}_i}{x_i} = \beta_X [R_i^X - \Sigma_j x_j R_j^X] + \alpha_X I_i^X \ , \tag{6}$$

where $I_i^X \equiv \Sigma_j x_j \log(x_j/x_i)$ represents the effect of memory with decay parameter α_X. (The continuous-time dynamic of Y follows in a similar manner.) Eq. (6), extended to account for any number of agents and actions, constitutes our general model for reinforcement-learning multiagent systems.

Simplifying again, assume a fixed relationship between pairs (i, j) of X's and Y's actions and between rewards for both agents: $R_{ij}^X = a_{ij}$ and $R_{ij}^Y = b_{ij}$. Assume further that $\mathbf{x}$ and $\mathbf{y}$ are independently distributed, then the time-average rewards for X and Y become

$$R_i^X = \Sigma_j a_{ij} y_j \text{ and } R_i^Y = \Sigma_j b_{ij} x_j \ , \tag{7}$$

In this restricted case, the continuous-time dynamic is:

$$\begin{aligned} \frac{\dot{x}_i}{x_i} &= \beta_X [(A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}] + \alpha_X I_i^X \ , \\ \frac{\dot{y}_i}{y_i} &= \beta_Y [(B\mathbf{x})_i - \mathbf{y} \cdot B\mathbf{x}] + \alpha_Y I_i^Y \ , \end{aligned} \tag{8}$$

where $(A)_{ij} = a_{ij}$ and $(B)_{ij} = b_{ij}$, $(A\mathbf{x})_i$ is the ith element of the vector $A\mathbf{x}$, and β_X and β_Y control the time scale of each agent's learning.

We can regard A and B as X's and Y's game-theoretic payoff matrices for action i against opponent's action j [18]. In contrast with game theory, which assumes agents have exact knowledge of the game structure and of other agent's strategies, reinforcement-learning agents have no knowledge of a "game" in which they are playing, only a myopic model of the environment—other agent(s)—given implicitly via the rewards they receive. Nonetheless, a game dynamics emerges—via R^X and R^Y in Eq. (6)—as a description of the collective's *global behavior*.

Given the basic equations of motion for the reinforcement-learning multiagent system (Eq. (8)), one becomes interested in, on the one hand, the time evolution of each agent's state vector in the simplices $\mathbf{x} \in \Delta_X$ and $\mathbf{y} \in \Delta_Y$ and, on the other, the dynamics in the higher-dimensional *collective* simplex $(\mathbf{x}, \mathbf{y}) \in \Delta_X \times \Delta_Y$. Following Ref. [12], we transform from $(\mathbf{x}, \mathbf{y}) \in \Delta_X \times \Delta_Y$ to $\mathbf{U} = (\mathbf{u}, \mathbf{v}) \in \mathbf{R}^{2(N-1)}$ with $\mathbf{u} = (u_1, \ldots, u_{N-1})$ and $\mathbf{v} = (v_1, \ldots, v_{N-1})$, where $u_i = \log(x_{i+1}/x_1)$ and $v_i = \log(y_{i+1}/y_1)$, $(i = 1, \ldots, N-1)$. The result is a new version of our simplified model (Eqs. (8)), useful both for numerical stability during simulation and also for analysis in certain limits:

$$\begin{aligned} \dot{u}_i &= \beta_X \frac{\sum_{j=1}^{N-1} \tilde{a}_{i,j} e^{v_j} + \tilde{a}_{i0}}{1 + \sum_{j=1}^{N-1} e^{v_j}} - \alpha_X u_i \\ \dot{v}_i &= \beta_Y \frac{\sum_{j=1}^{N-1} \tilde{b}_{i,j} e^{u_j} + \tilde{b}_{i0}}{1 + \sum_{j=1}^{N-1} e^{u_j}} - \alpha_Y v_i \ , \end{aligned} \tag{9}$$

where $\tilde{a}_{ij} = a_{i+1,j+1} - a_{1,j+1}$ and $\tilde{b}_{ij} = b_{i+1,j+1} - b_{1,j+1}$. Since the dissipation rate γ in $\mathbf{U}$ is

$$\gamma = \Sigma_i \frac{\partial \dot{u}_i}{\partial u_i} + \Sigma_j \frac{\partial \dot{v}_j}{\partial v_j} = -(N-1)(\alpha_X + \alpha_Y), \tag{10}$$

Eqs. (8) are conservative when $\alpha_X = \alpha_Y = 0$ and the time average of a trajectory is the Nash equilibrium of the game specified by A and B, if a limit set exists in the interior of $\Delta_X \times \Delta_Y$[19]. Moreover, if the game is zero-sum, the dynamics are Hamiltonian in $\mathbf{U}$ with

$$\begin{aligned} H = &- \beta_Y [\Sigma_j x_j^* u_j - \log(1 + \Sigma_j e^{u_j})] \\ &- \beta_X [\Sigma_j y_j^* v_j - \log(1 + \Sigma_j e^{v_j})] \ , \end{aligned} \tag{11}$$

where $(\mathbf{x}^*, \mathbf{y}^*)$ is an interior Nash equilibrium [12].

To illustrate the dynamical-systems analysis of learning in multiagent systems using the above framework, we now analyze the behavior of the two-person rock-scissors-paper interaction [20]. This familiar game describes a nontransitive three-sided competition: rock beats scissors, scissors beats paper, and paper beats rock. The reward structure (environment) is given by:

$$A = \begin{bmatrix} \epsilon_X & 1 & -1 \\ -1 & \epsilon_X & 1 \\ 1 & -1 & \epsilon_X \end{bmatrix} \text{ and } B = \begin{bmatrix} \epsilon_Y & 1 & -1 \\ -1 & \epsilon_Y & 1 \\ 1 & -1 & \epsilon_Y \end{bmatrix} \ , \tag{12}$$

where $\epsilon_X, \epsilon_Y \in [-1.0, 1.0]$ are the rewards for ties. The mixed Nash equilibrium is $x_i^* = y_i^* = 1/3, (i = 1, 2, 3)$—the centers of Δ_X and Δ_Y. If $\epsilon_X = -\epsilon_Y$, the game is zero-sum.

In the special case of perfect memory ($\alpha_X = \alpha_Y = 0$) and with equal learning sensitivity ($\beta_X = \beta_Y$), the linear version (Eqs. (8)) of our model (Eq. (6)) reduces to multipopulation replicator equations [11]:

$$\frac{\dot{x}_i}{x_i} = [(A\mathbf{y})_i - \mathbf{x} \cdot A\mathbf{y}] \text{ and } \frac{\dot{y}_i}{y_i} = [(B\mathbf{x})_i - \mathbf{y} \cdot B\mathbf{x}] \ . \tag{13}$$

The game-theoretic behavior in this case with rock-scissors-paper interactions (Eqs. (12)) was investigated

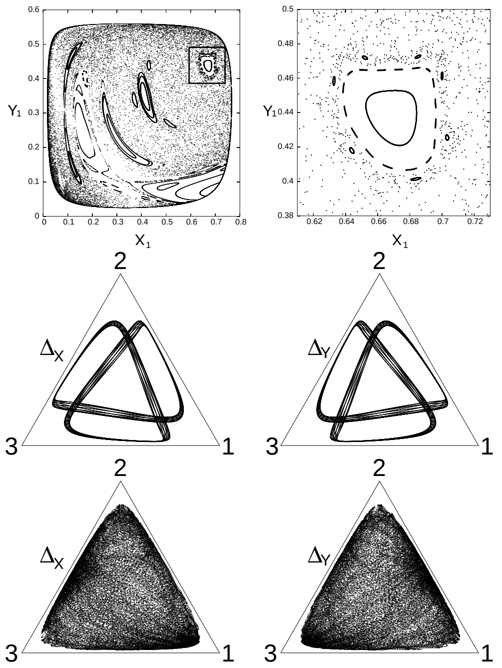

FIG. 1: Quasiperiodic tori and chaos: $\epsilon_X = -\epsilon_Y = 0.5$, $\alpha_X = \alpha_Y = 0$, and $\beta_X = \beta_Y$. We give a Poincaré section (top) on the hyperplane defined by $\dot{u}_1 = 0$ and $\dot{v}_1 > 0$; that is, in the $(\mathbf{x}, \mathbf{y})$ space: $(3 + \epsilon_X)y_1 + (3 - \epsilon_X)y_2 - 2 = 0$ and $(3 + \epsilon_Y)x_1 + (3 - \epsilon_Y)x_2 - 2 < 0$. There are 23 randomly selected initial conditions with energies $H = -1/3(u_1 + u_2 + v_1 + v_2) + \log(1 + e^{u_1} + e^{u_2}) + \log(1 + e^{v_1} + e^{v_2}) = 2.941693$, which surface forms the outer border of $H \leq 2.941693$. Two rows (bottom): Representative trajectories, simulated with a 4th-order symplectic integrator [13], starting from initial conditions within the Poincaré section. The upper simplices show a torus in the section's upper right corner; see the enlarged section at the upper right. The initial condition is $(\mathbf{x}, \mathbf{y}) = (0.3, 0.054196, 0.645804, 0.1, 0.2, 0.7)$. The lower simplices are an example of a chaotic trajectory passing through the regions in the section that are a scatter of dots; the initial condition is $(\mathbf{x}, \mathbf{y}) = (0.05, 0.35, 0.6, 0.1, 0.2, 0.7)$.

in [10]. Here, before contrasting our more general setting, we briefly recall the behavior in these special cases, noting several additional results.

Figure 1 shows Poincaré sections of Eqs. (13)'s trajectories on the hyperplane ($\dot{u}_1 = 0, \dot{v}_1 > 0$) and representative trajectories in the individual agent simplices Δ_X and Δ_Y. When $\epsilon_X = -\epsilon_Y = 0.0$, we expect the system to be integrable and only quasiperiodic tori should exist. Otherwise, $\epsilon_X = -\epsilon_Y > 0.0$, Hamiltonian chaos can occur with positive-negative pairs of Lyapunov exponents [10]. The dynamics is very rich, there are infinitely many distinct behaviors near the unstable fixed point at the center—the classical Nash equilibrium—and a periodic orbit arbitrarily close to any chaotic one. Moreover, when the game is not zero-sum ($\epsilon_X \neq \epsilon_Y$), transients to heteroclinic cycles are observed [10]: On the one hand, there

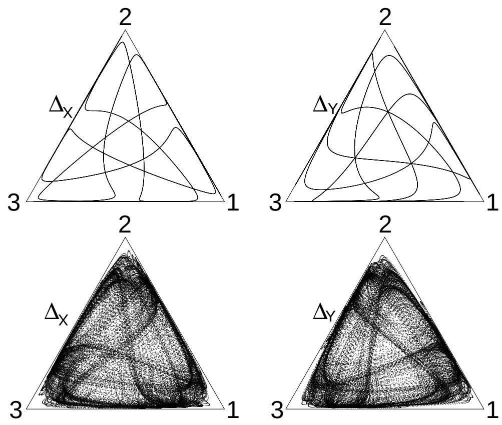

FIG. 2: Limit cycle (top: $\epsilon_Y = 0.025$) and chaotic attractors (bottom: $\epsilon_Y = -0.365$), with $\epsilon_X = 0.5$, $\alpha_X = \alpha_Y = 0.01$, and $\beta_X = \beta_Y$.

are intermittent behaviors in which the time spent near pure strategies (the simplicial vertices) increases subexponentially with $\epsilon_X + \epsilon_Y < 0$ and, on the other hand, with $\epsilon_X + \epsilon_Y > 0$, chaotic transients persist; cf. [14].

Our framework goes beyond these special cases and, generally, beyond the standard multipopulation replicator equations (Eqs. (13)) due to its accounting for the effects of individual and collective learning and since the reward structure and the learning rules need not lead to linear interactions. For example, if the memory decay rates (α_X and α_Y) are positive, the system becomes dissipative and exhibits limit cycles and chaotic attractors; see Fig. 2. Figure 3 (top) shows a diverse range of bifurcations as a function of ϵ_Y: dynamics on the hyperplane ($\dot{u}_1 = 0$, $\dot{v}_1 > 0$) projected onto y_1. When the game is nearly zero-sum, agents can reach the stable Nash equilibrium, but chaos can also occur, when $\epsilon_X + \epsilon_Y > 0$. Figure 3 (bottom) shows that the largest Lyapunov exponent is positive across a significant fraction of parameter space; indicating that chaos is common. The dual aspects of chaos, irregularity and coherence, imply that agents may behave cooperatively or competitively (or switch between both) in the collective dynamics. Such global behaviors ultimately derive from self-interested, myopic learning.

Within this framework a number of extensions suggest themselves as ways to investigate the emergence of collective behaviors. The most obvious is the generalization to an arbitrary number of agents with an arbitrary number of strategies and the analysis of behaviors in thermodynamic limit; see, e.g., [15] as an alternative approach. It is relatively straightforward to develop an extension to the linear-reward version (Eqs. (8)) of our model. For three agents X, Y, and Z, one obtains:

$$\frac{\dot{x}_i}{x_i} = \beta_X [\Sigma_{j,k} a_{ijk} y_j z_k - \Sigma_{j,k,l} a_{jkl} x_j y_k z_l] + \alpha_X I_i^X \ , \ (14)$$

with tensor $(A)_{ijk} = a_{ijk}$, and similarly for Y and Z.

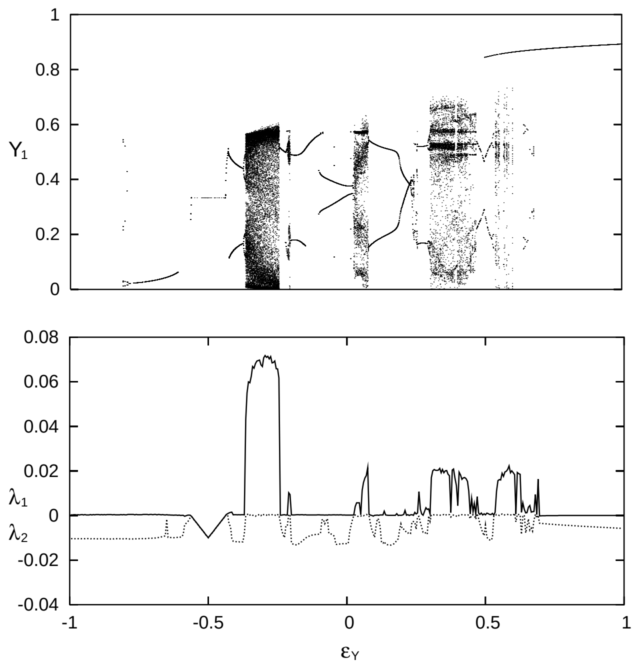

FIG. 3: Bifurcation diagram (top) of dissipative (learning with memory loss) dynamics projected onto coordinate y_1 from the Poincaré section hyperplane ($\dot{u}_1 = 0$, $\dot{v}_1 > 0$) and the largest two Lyapunov exponents λ_1 and λ_2 (bottom) as a function of $\epsilon_Y \in [-1, 1]$. Here with $\epsilon_X = 0.5$, $\alpha_X = \alpha_Y = 0.01$, and $\beta_X = \beta_Y$. Simulations show that λ_3 and λ_4 are always negative.

Not surprisingly, this is also a conservative system when the α's vanish. However, extending the general collective learning equations (Eq. (6)) to multiple agents is challenging and so will be reported elsewhere.

To be relevant to applications, one also needs to develop a statistical dynamics generalization [16] of the deterministic equations of motion to account for finite and fluctuating numbers of agents and also finite histories used in learning. Finally, another direction, especially useful if one attempts to quantify collective function in large multiagent systems, will be structural and information-theoretic analyses [17] of local and global learning behaviors and, importantly, their differences. Analyzing the stored information in each agent versus that in the collective, the causal architecture of information flow between an individual agent and the group, and how individual and global memories are processed to sustain collective function are projects now made possible using this framework.

We presented a dynamical-systems model of collective learning in multiagent systems, which starts with reinforcement-learning agents and reduces to coupled replicator equations, demonstrated that individual-agent learning induces a global game dynamics, and investigated some of the resulting periodic, intermittent, and chaotic behaviors with simple (linear) rock-scissors-papers game interactions. Our model gives a macroscopic description of a network of learning agents that can be straightforwardly extended to model a large number of heterogeneous agents in fluctuating environments. Since deterministic chaos occurs even in this simple setting, one expects that in high-dimensional and heterogeneous populations typical of multiagent systems intrinsic unpredictability will become a dominant collective behavior. Sustaining useful collective function in multiagent systems becomes an even more compelling question in light of these results.

The authors thank D. Albers, J. D. Farmer, E. Akiyama, P. Patelli, and C. Shalizi. This work was supported at the Santa Fe Institute under the Network Dynamics Program by Intel Corporation and under DARPA agreement F30602-00-2-0583. YS's participation was supported by the Postdoctoral Researchers Program at RIKEN.

* Electronic address: ysato@bdc.brain.riken.go.jp
† Electronic address: chaos@santafe.edu

[1] S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, and E. Bonabeau, eds., *Self-Organization in Biological Systems* (Princeton University Press, Princeton, 2001).
[2] H. A. Simon, *The Sciences of the Artificial*, Karl Taylor Compton Lectures (MIT Press, Cambridge, 1996), first edition 1969.
[3] H. P. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton University Press, Princeton, 1998).
[4] E. Hutchins, *Cognition in the Wild* (MIT Press, Cambridge, 1996).
[5] O. E. Rossler, Ann. NY Acad, Sci. **504**, 229 (1987).
[6] M. Taiji and T. Ikegami, Physica **D134**, 253 (1999).
[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT Press, 1998).
[8] P. Taylor and L. Jonker, Math. Bio. **40**, 145 (1978).
[9] T. Borgers and R. Sarin, J. Econ. Th. **77**, 1 (1997).
[10] Y. Sato, E. Akiyama, and J. D. Farmer, Proc. Natl. Acad. Sci. USA **99**, 4748 (2002).
[11] P. Taylor, J. Appl. Prob. **16**, 76 (1979).
[12] J. Hofbauer, J. Math. Biol. **34**, 675 (1996).
[13] H. Yoshida, Phys. Lett. **A150**, 262 (1990).
[14] T. Chawanya, Prog. Theo. Phys. **94**, 163 (1995).
[15] M. Marsili, D. Challet, and R. Zecchina, Physica **A 280**, 522 (2000).
[16] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell, Theoret. Comp. Sci. **229**, 41 (1999).
[17] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989), see also, J. P. Crutchfield and D. P. Feldman, CHAOS (March 2003) in press.
[18] Eqs. (7) specify the von Neumann-Morgenstern utility (J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, (Princeton University Press, 1944)).
[19] Cf. P. Schuster et al, Biol. Cybern. **40**, 1 (1981).
[20] Such interactions are observed in natural social and bi-

ological communities; cf. B. Kerr et al, Nature **418**, 171 (2002).