Second-Order Free-Riding on Antisocial Punishment Restores the Effectiveness of Prosocial Punishment

Attila Szolnoki[1,*] and Matjaž Perc[2,†]

[1]*Institute of Technical Physics and Materials Science, Centre for Energy Research,*
Hungarian Academy of Sciences, P.O. Box 49, H-1525 Budapest, Hungary
[2]*Faculty of Natural Sciences and Mathematics, University of Maribor,*
Koroška cesta 160, SI-2000 Maribor, Slovenia;
CAMTP-Center for Applied Mathematics and Theoretical Physics, University of Maribor,
Mladinska 3, SI-2000 Maribor, Slovenia;
and Complexity Science Hub, Josefstädterstraße 39, A-1080 Vienna, Austria

Economic experiments have shown that punishment can increase public goods game contributions over time. However, the effectiveness of punishment is challenged by second-order free-riding and antisocial punishment. The latter implies that noncooperators punish cooperators, while the former implies unwillingness to shoulder the cost of punishment. Here, we extend the theory of cooperation in the spatial public goods game by considering four competing strategies, which are traditional cooperators and defectors, as well as cooperators who punish defectors and defectors who punish cooperators. We show that if the synergistic effects are high enough to sustain cooperation based on network reciprocity alone, antisocial punishment does not deter public cooperation. Conversely, if synergistic effects are low and punishment is actively needed to sustain cooperation, antisocial punishment does is viable, but only if the cost-to-fine ratio is low. If the costs are relatively high, cooperation again dominates as a result of spatial pattern formation. Counterintuitively, defectors who do not punish cooperators, and are thus effectively second-order free-riding on antisocial punishment, form an active layer around punishing cooperators, which protects them against defectors that punish cooperators. A stable three-strategy phase that is sustained by the spontaneous emergence of cyclic dominance is also possible via the same route. The microscopic mechanism behind the reported evolutionary outcomes can be explained by the comparison of invasion rates that determine the stability of subsystem solutions. Our results reveal an unlikely evolutionary escape from adverse effects of antisocial punishment, and they provide a rationale for why second-order free-riding is not always an impediment to the evolutionary stability of punishment.

Subject Areas: Complex Systems,
Interdisciplinary Physics

I. INTRODUCTION

Cooperation is widespread in human societies [1–7]. Like no other species, we champion personal sacrifice for the common good [8,9]. Not only are people willing to incur costs to help unrelated others, economic experiments have shown that many are also willing to incur costs to punish those that do not cooperate [10–16]. Unfortunately, cooperation is jeopardized by selfish incentives to free-ride on the selfless contributions of others. What is more, individuals that abstain from punishing such free-riders are often called second-order free-riders for their failure to

bear the additional costs of punishment [17,18]. Several evolutionary models have been developed to study the effects of punishment on cooperation [19–27], and it has been pointed out that second-order free-riding is among the biggest impediments to the evolutionary stability of punishment [28–31].

In addition to second-order free-riding, the effectiveness of punishment is challenged by antisocial punishment. The fact that noncooperators sometimes punish cooperators has been observed experimentally in different human societies [32–37], and it has been shown theoretically that this antisocial punishment can prevent the successful coevolution of punishment and cooperation [38,39]. In fact, if antisocial punishment is an option, prosocial punishment may no longer increase cooperation, deteriorating instead to a self-interested tool for protecting oneself against potential competitors [40]. While the punishment of free-riders is aimed at increasing cooperation, antisocial punishment can be a form of retaliation for punishment received in

*szolnoki.attila@energia.mta.hu
†matjaz.perc@uni-mb.si

repeated games [32,41], or is simply aimed at cooperators without a retaliatory motive [36,37].

Given the potential drawbacks associated with punishment related to second-order free-riding and antisocial punishment, it has been rightfully pointed out that the maintenance of cooperation may be better achievable through less destructive means. In particular, rewards may be as effective as punishment and lead to higher total earnings without potential damage to reputation or fear from retaliation [42,43]. Although many evolutionary models confirm the effectiveness of positive incentives for the promotion of cooperation [44–52], in this case too the challenges associated with second-order free-riding and antisocial rewarding persist [53,54].

Here, we use methods of statistical physics to show how the two long-standing problems—namely, second-order free-riding and antisocial punishment—cancel each other out in an unlikely and counterintuitive evolutionary outcome, and in doing so restore the effectiveness of prosocial punishment to promote cooperation. We extend the theory of cooperation by considering the spatial public goods game with nonpunishing cooperators and defectors, as well as with cooperators who punish defectors and defectors who punish cooperators. As we show in detail, spatial pattern formation leads to unconditional defectors forming an active layer around punishing cooperators, which protects them against defectors that punish cooperators. This is a new evolutionary escape from adverse effects of antisocial punishment, which in turn also reveals unexpected benefits stemming from second-order free-riding.

In what follows, we first present the spatial public goods game with prosocial and antisocial punishment, and then proceed with the results and a discussions of their implications for the successful coevolution of cooperation and punishment.

II. PUBLIC GOODS GAME WITH PROSOCIAL AND ANTISOCIAL PUNISHMENT

The traditional version of the public goods game is simple and intuitive, and it captures the essence of the puzzle that is human cooperation [55,56]. In a group of players, each one can decide whether to cooperate (C) or defect (D). Cooperators contribute a fixed amount (equal to 1 without loss of generality) to the common pool, while defectors contribute nothing. The sum of all contributions is multiplied by a multiplication factor $r > 1$, which takes into account synergistic effects of cooperation, and the resulting amount of public goods is divided equally among all group members irrespective of their strategies. Defection thus yields the highest short-term individual payoffs, while cooperation is best for the group as a whole.

Here, we extend this game by introducing two additional strategies, namely, cooperators that punish defectors (P_C) and defectors that punish cooperators (P_D). The former represent prosocial punishment, while the latter represent antisocial punishment. Technically, P_C players punish D and P_D players, while P_D players punish C and P_C players. In a g group of size G the resulting payoffs are

$$\Pi_D^g = \Pi_{PGG} - \beta \frac{N_{P_C}}{G-1}, \tag{1}$$

$$\Pi_C^g = \Pi_{PGG} - \beta \frac{N_{P_D}}{G-1} - 1, \tag{2}$$

$$\Pi_{P_D}^g = \Pi_{PGG} - \beta \frac{N_{P_C}}{G-1} - \gamma \frac{N_C + N_{P_C}}{G-1}, \tag{3}$$

$$\Pi_{P_C}^g = \Pi_{PGG} - \beta \frac{N_{P_D}}{G-1} - \gamma \frac{N_D + N_{P_D}}{G-1} - 1, \tag{4}$$

where

$$\Pi_{PGG}^g = \frac{r(N_C + N_{P_C})}{G}, \tag{5}$$

and N_C, N_D, N_{P_C}, and N_{P_D} are, respectively, the number of nonpunishing cooperators, nonpunishing defectors, punishing cooperators, and punishing defectors in the g group. In addition to the multiplication factor r, we have two additional parameters, which are β, as the maximal fine imposed on a player if all other players within the group punish her, and γ, as the maximal cost of punishment that can apply. Importantly, the values of both β and γ are kept the same for prosocial and antisocial punishment so as to not give either a default evolutionary advantage or disadvantage.

This public goods game is staged on a square lattice with periodic boundary conditions where L^2 players are arranged into overlapping groups of size $G = 5$ such that everyone is connected to its $G - 1$ nearest neighbors. Accordingly, each player belongs to $g = 1, \ldots, G$ different groups, each of size G. Notably, the square lattice is the simplest of networks that takes into account the fact that the interactions among us are inherently structured rather than random. By using the square lattice, we continue a longstanding tradition that began with the work of Nowak and May [57], and which has since emerged as a default setup to reveal all evolutionary outcomes that are feasible within a particular version of the public goods game [56]. We should note, however, that our observations are robust and are not restricted to this interaction topology. The only crucial criteria are that players should have limited and stable connections with others, which allows network reciprocity to work.

Monte Carlo simulations are carried out as follows. Initially, each player on site x is designated as either a nonpunishing cooperator, a nonpunishing defector, a punishing cooperator, or a punishing defector with equal probability. The following elementary steps are then iterated repeatedly until a stationary solution is obtained,

i.e., until the average fractions of strategies on the square lattice become time independent. During an elementary step a randomly selected player x plays the public goods game in all the G groups where she is a member, whereby her overall payoff Π_{s_x} is thus the sum of all the payoffs $\Pi_{s_x}^g$ acquired in each individual group, as described in Eqs. (1)–(5). Next, a randomly selected neighbor of player x acquires her payoff Π_{s_y} in the same way. Lastly, player y imitates the strategy of player x with a probability given by the Fermi function,

$$\Gamma(s_x \rightarrow s_y) = 1/\{1 + \exp[(\Pi_{s_y} - \Pi_{s_x})/K], \qquad (6)$$

where K quantifies the uncertainty by strategy adoptions [58], implying that better performing players are readily adopted, although it is not impossible to adopt the strategy of a player performing worse. In the $K \rightarrow 0$ limit, player y imitates the strategy of player x if and only if $\Pi_{s_x} > \Pi_{s_y}$. Conversely, in the $K \rightarrow \infty$ limit, payoffs cease to matter and strategies change as per a flip of a coin. Between these two extremes players with a higher payoff will be readily imitated, although the strategy of underperforming players may also be occasionally adopted, for example, due to errors in decision making, imperfect information, and external influences that may adversely affect the evaluation of an opponent. Without loss of generality, we use $K = 0.5$, in agreement with previous research that showed this to be a fully representative value [58–60]. Repeating all described elementary steps L^2 times constitutes one full Monte Carlo step (MCS), thus giving a chance to every player to change their strategy once on average.

We note that imitation is a fundamental process by means of which humans change their strategies [61–64]. The application of imitation-based strategy updating based on the Fermi function is thus appropriate and justified, although, as we show, our results are robust to changes in the details that determine the microscopic dynamics of the studied public goods game. In terms of the application of the square lattice, we emphasize that, despite its simplicity, it fully captures the most relevant aspect of human interactions—namely, the fact that nobody interacts randomly with everybody else, not even in small groups, and that our interaction range is thus inherently limited. Applications of more complex interaction topologies are of course possible, but this does not affect our results. This is because our key argument is based on the limited number of interactions a players has, but it does not in any way rely on the specific properties of the square lattice topology.

The average fractions of all four strategies on the square lattice are determined in the stationary state after a sufficiently long relaxation time. Depending on the proximity to phase transition points and the typical size of emerging spatial patterns, the linear system size is varied from $L = 400$ to 6000, and the relaxation time is varied from 10^4 to 10^6 MCS to ensure that the statistical error is

comparable with the size of the symbols in the figures. We emphasize that the usage of a sufficiently large system size is a decisive factor that allows us to identify the correct evolutionary stable solutions. Using a too small system size may easily prevent this, for example, if the linear size of the lattice is comparable to or smaller than the typical size of the emerging spatial patterns.

III. RESULTS

Before presenting the main results, we briefly summarize the evolutionary outcomes in a well-mixed population. In the absence of a limited interaction range, the behavior is largely trivial and resembles that reported before for the traditional two-strategy public goods game [29,55]. In particular, if r exceeds the group size G, then both cooperative strategies dominate while all defectors die out. Conversely, below this threshold both defector strategies dominate while all cooperators die out. This behavior is also in agreement with the well-mixed results published in Ref. [39]. In short, all the nontrivial evolutionary solutions we report here in the continuation are due to the consideration of a structured population and remain completely hidden if a well-mixed population is assumed.

In what follows, we focus on two representative values of the multiplication factor that cover two relevantly different public goods game scenarios. First, we use $r = 3.8$, where the spatial selection allows cooperators to survive even in the absence of punishment—this is the well-known manifestation of network reciprocity, where the limited interactions among players allow cooperators to organize themselves into compact clusters, which confers them competitive payoffs in comparison to defectors [65]. Subsequently, we also use a sufficiently small, $r = 3.0$, value, where cooperators can no longer survive solely due to network reciprocity and thus require additional support [58]. For both values of r we determine the stationary fractions of strategies dependent on the punishment fine β and the punishment cost γ, and we pinpoint the location and type of phase transitions from the Monte Carlo simulation data.

In Fig. 1, we show the full β-γ phase diagram, as obtained for $r = 3.8$. As discussed above, we refer to this as the strong network reciprocity region. Presented results reveal that antisocial punishment is hardly viable, with P_D players surviving only in a tiny region of the β-γ parameter space. Conversely, as the fine β increases, punishing cooperators subvert nonpunishing cooperators, first via a discontinuous phase transition from the two-strategy $D + C$ phase to the two-strategy $D + P_C$ phase, and subsequently via a continuous phase transition to the absorbing P_C phase. The discontinuous phase transition is due to indirect territorial competition, which emerges between C and P_C players competing against defectors [29], while the continuous phase transition is due to an
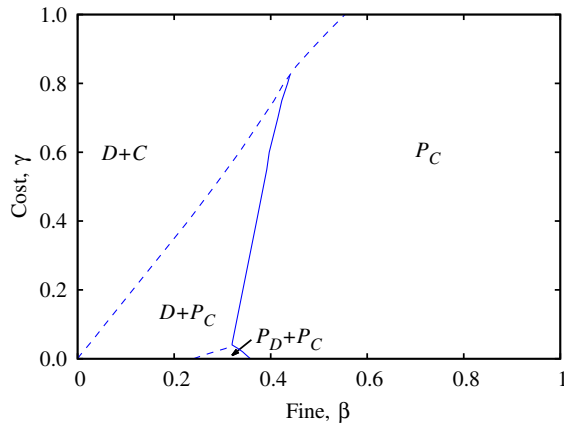
FIG. 1.   Full β-γ phase diagram of the spatial public goods game with prosocial and antisocial punishment, as obtained for $r = 3.8$. Solid lines denote continuous phase transitions, while dashed lines denote discontinuous phase transitions. Two representative cross sections of this phase diagram are presented in Fig. 2.

increasing effectiveness of punishment that stems from the larger fines.

The two representative cross sections of the phase diagram in Fig. 2 provide a more quantitative insight into the nature of these phase transitions. In both cases the application of small punishment fines yields a punishment-free state, where traditional cooperators and defectors coexist due to network reciprocity. If the cost of punishment is considerable, as in Fig. 2(a), the $D + C$ phase suddenly gives way to the $D + P_C$ phase at a critical value of the punishment fine $\beta_c = 0.229$, and by increasing β further, a defector-free state is reached at $\beta_c = 0.361$. This succession of phase transitions remains the same if the cost of punishment is tiny, shown in Fig. 2(b), apart from a narrow intermediate region of β, where antisocial punishment replaces nonpunishing defectors via a discontinuous $D + P_C \rightarrow P_D + P_C$ phase transition at $\beta_c = 0.284$. Interestingly, this phase transition is qualitatively identical to the preceding $D + C \rightarrow D + P_C$ phase transition—in both cases nonpunishing strategies are subverted by their punishing counterparts on the grounds of increasing punishment fines. It can also be observed that the emergence of a stable $P_D + P_C$ phase involves a slight decay of the fraction of P_C players, although they quickly recover to full dominance as the punishment fine is increased further.

To sum up, in the strong network reciprocity region, antisocial punishment has a negligible impact on the evolution of cooperation. In a small region of the β-γ parameter space, antisocial punishers can outperform nonpunishing defectors to form a stable coexistence with prosocial punishers. But apart from this, and despite the fact that both forms of punishment are implemented equally effectively (the values of both β and γ are kept the same for prosocial and antisocial punishment), antisocial punishment fails and is evolutionary unsuccessful.
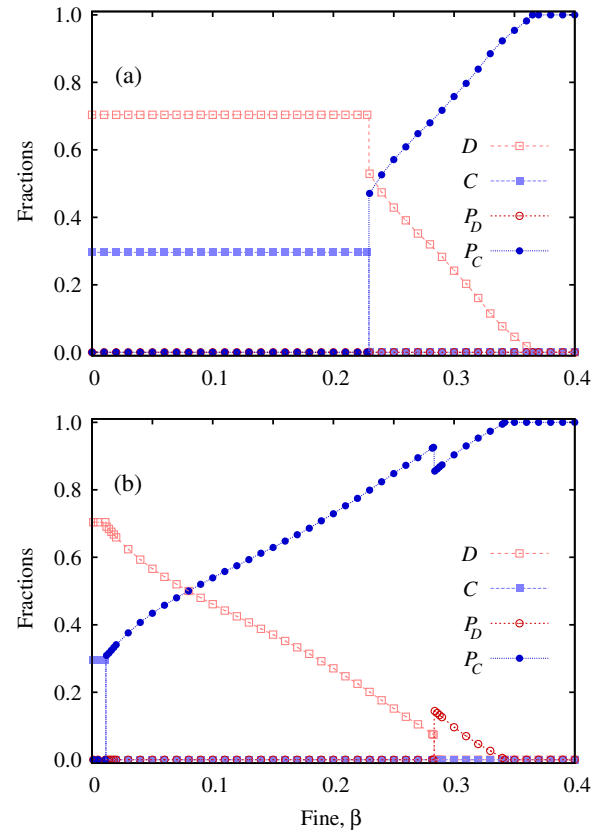

FIG. 2.   Two representative cross sections of the phase diagram depicted in Fig. 1, as obtained for the punishment cost $\gamma = 0.4$ (a) and $\gamma = 0.02$ (b). Depicted are stationary fractions of the four competing strategies dependent on the punishment fine β.

An exciting question now is what if the network reciprocity alone is not strong enough to support the coexistence of cooperators and defectors? Although previous research has shown that prosocial punishment can be effective if the imposed fines are sufficiently high [66,67], this result was obtained in the absence of antisocial punishment. However, if cooperators can also be punished, the situation changes significantly. A subsystem analysis of the public goods game entailing only punishing coopera-tors and punishing defectors actually reveals that at $r = 3.0$ cooperation is unable to survive regardless of the values of β and γ, and regardless of the fact that punishing cooper-ators also benefit from network reciprocity.

Quite remarkably, the evolutionary outcome of the full four-strategy public goods game can be very different. The full β-γ phase diagram presented in Fig. 3 reveals that punishing cooperators can actually dominate completely in a sizable region of the parameter plane. Of course, if the cost of punishment is too high in relation to the imposed fines, defectors dominate. More precisely, C and P_C players die out fast, with only D and P_D players remaining. In the absence of the two cooperative strategies, the relation between D and P_D players is neutral, since the latter do not need to bear the punishment cost. This yields a
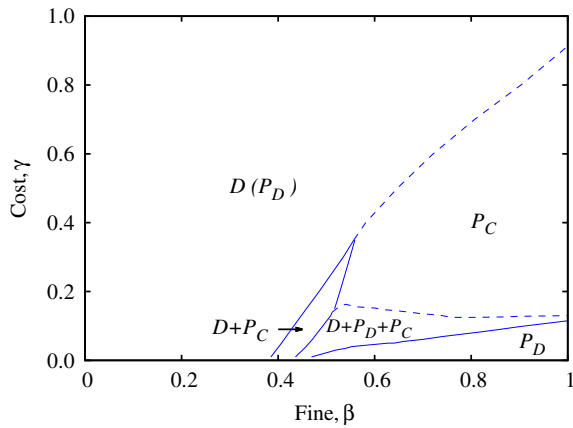
FIG. 3. Full β-γ phase diagram of the spatial public goods game with prosocial and antisocial punishment, as obtained for $r = 3.0$. Solid lines denote continuous phase transitions, while dashed lines denote discontinuous phase transitions. Representative spatial evolutions of the four competing strategies are presented in Figs. 4–6, while two representative cross sections of this phase diagram are presented in Fig. 7.
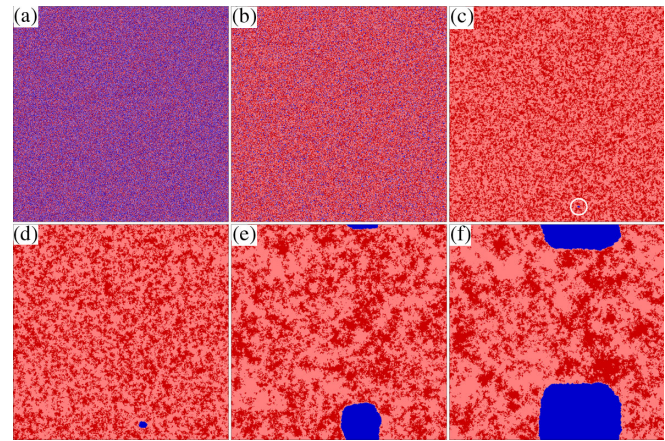

FIG. 4. Representative spatial evolution of the four competing strategies. Depicted are snapshots of the square lattice, as obtained for $\beta = 0.8$, $\gamma = 0.36$, and $r = 3$. Nonpunishing (punishing) cooperators are depicted in light blue (dark blue), while nonpunishing (punishing) defectors are depicted in light red (dark red). From a random initial state (a) both cooperative strategies start vanishing quickly (b). The only chance for cooperation to survive is if a lucky P_C seed starts growing in the sea of D players, as indicated in (c) by a white circle. It turns out, however, that P_C players, surrounded by a thin active layer of D players, can rise to complete dominance over time, as shown in (d)–(f) (the final state, where only P_C remain after 6000 MCS, is not shown). The linear size of the lattice is $L = 800$.

logarithmically slow coarsening without surface tension, as in the voter model [68,69]. In this case, the probability to reach either the absorbing D or the absorbing P_D phase depends on the fraction of these two strategies [70] when all cooperators die out, which is typically higher for D, and hence the $D(P_D)$ notation in Fig. 3.

Returning to the absorbing P_C phase, since network reciprocity at $r = 3.0$ is weak, an additional mechanism must be at work that allows the dominance of cooperation despite the low multiplication factor and despite antisocial punishment. This mechanism is illustrated in Fig. 4, where we show a representative spatial evolution of the four competing strategies from a random initial state for parameter values that yield the absorbing P_C phase. It is important to emphasize that a sufficiently large square lattice must be used, since otherwise the evolutionary process can quickly lead to a misleading outcome, i.e., to a solution that is not stable in the large population size limit, or to a solution that is highly sensitive on the initial fraction of strategies, as reported in Ref. [39]. This is, however, just a finite-size effect because the evolutionary stable solution can spread in the whole population if it has a chance to emerge somewhere locally. Because of the random initial state the number of nonpunishing and punishing cooperators starts dropping fast because support from network reciprocity is lacking, both because the value of r is low and even more so because compact clusters are not yet formed so early in the process. During this stage, if the population would be small, an accidental extinction of P_C players would be very likely. Indeed, even with $L = 800$ they manage to just barely survive, as indicated in Fig. 4(c) by a white circle. At this point the temporary winners appear to be D and P_D players, which in the

absence of cooperators are neutral, and hence perform a logarithmically slow coarsening [69].

However, the unlikely evolutionary twist is yet to come and reveals itself in Figs. 4(d)–4(f). Since P_C players are weaker than P_D players, the only chance for the former to survive is if they form a compact cluster inside a D domain (C cannot survive either way because $r = 3.0$ is too small). Although one might suspect that this "hanging by a thread" type of survival of P_C players is merely temporary because the superior P_D players will eventually invade their cluster, this does in fact never happen. On the contrary, punishing cooperators eventually rise to complete dominance (the final state is not shown in Fig. 4).

Crucial for the understanding of this counterintuitive evolutionary outcome is the realization that punishing defectors suffer from second-order free-riding of nonpunishing defectors as soon as they both meet in the vicinity of punishing cooperators. More precisely, when D players meet with P_D players in the vicinity of P_C players, P_D players have to bear the additional cost of punishment while D players are, of course, free from this burden. The same argument is traditionally put forward when it is time to explain why punishing cooperators are uncompetitive next to nonpunishing cooperators near defectors, and why, in fact, punishment is evolutionary unstable. When antisocial punishment is present, however, this very same reasoning helps punishing cooperators to beat defectors that punish them. As a result, D players start invading P_D
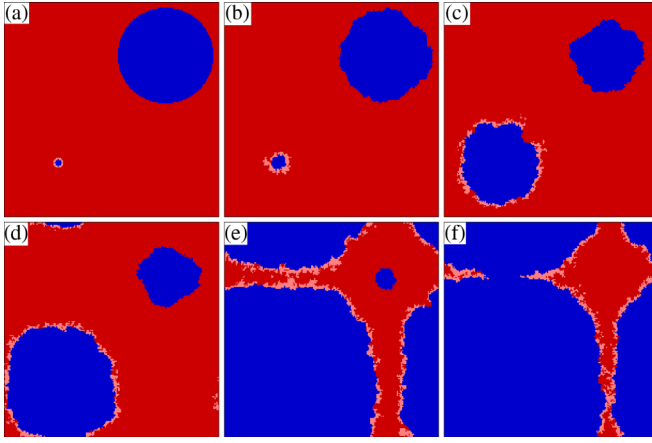
FIG. 5. An illustration of how second-order free-riding on antisocial punishment restores the effectiveness of prosocial punishment. A small domain of P_C players (dark blue), surrounded by a thin layer of D players (light red), is inserted into the sea of P_D players (dark red) in the bottom left corner of the lattice (a). Similarly, a sizable domain of P_C players, but without the protective D layer, is inserted into the sea of P_D players in the upper right-hand corner of the lattice (a). While the large P_C domain without the protective layer shrinks over time, the small P_C domain with the protective D layer grows (b)–(f). The absorbing P_C phase is reached after 2000 MCS (not shown). The linear size of the square lattice in this case is $L = 200$. Other parameters are $\beta = 0.8$, $\gamma = 0.3$, and $r = 3$.

domains, but in parallel, P_C players also invade D players from the other side of the interface. The thin active layer of D players thus acts as a protection, shielding P_C players from a direct invasion of P_D. As can be seen in Figs. 5(d)–5(f), the shield is not passive, but expands permanently because D players become successful when meeting P_D players close to cooperators. (This process is quantified via an effective invasion rate in the following section.) At the end, when P_D players die out, the D shield falls victim to the invasion of P_C players, which thus rise to complete dominance. A direct illustration of this mechanism is shown in Fig. 5, where a prepared initial state is used for clarity. The comparison of the evolution of a large but lonely P_C domain and a tiny but D-protected seed of

the same strategy illustrates nicely that the previously described "activated layer" mechanism is effective to overcome the danger of the simultaneous presence of antisocial punishment.

We show the above-described pattern formation in the animation provided in Ref. [71]. At this point, we also emphasize that the key mechanism that is responsible for the recovery of prosocial punishment is not restricted to the application of imitation-based strategy updating and is in fact robust to changes in the microscopic dynamics. For example, if we apply the so-called "score-dependent viability" strategy updating [39,72], the trajectory of evolution remains the same [73]. The only visible difference is that, in the latter case, the interfaces that separate the competing domains are rugged and strongly fluctuating, which in turn decelerates the evolutionary dynamics and prolongs the time needed to arrive at the same final outcome.

As we point out, snapshots in Fig. 4 illustrate clearly that the size of the lattice plays a decisive role in reaching the correct evolutionary outcome from a random initial state. For some parameter values that bring the population closer to a phase transition point or because of the large fluctuations of strategy abundance during the pattern formation, even $L = 6000$ (linear system size) can turn out to be too small. In such cases a prepared initial state, as depicted in Fig. 6(a), consisting of sizable patches of the four competing strategies, can help to determine the correct composition of the stationary solution. We use this approach to determine the stability of the three-strategy $D + P_D + P_C$ phase, which according to the phase diagram in Fig. 3 also forms an important part of the solution. Figure 6 illustrates that such a solution can be observed even if using a very small lattice size, if only suitable initial conditions are used. The alternating oscillations of red and blue indicate that this three-strategy phase is sustained by cyclic dominance. Indeed, due to using a different set of β and γ values from those used in Fig. 4, here P_D players beat P_C players because of the low value of r, P_C players beat D players because of prosocial punishment, and D players beat P_D players near cooperators because of second-order free-riding. However, the balance of these invasions is such that
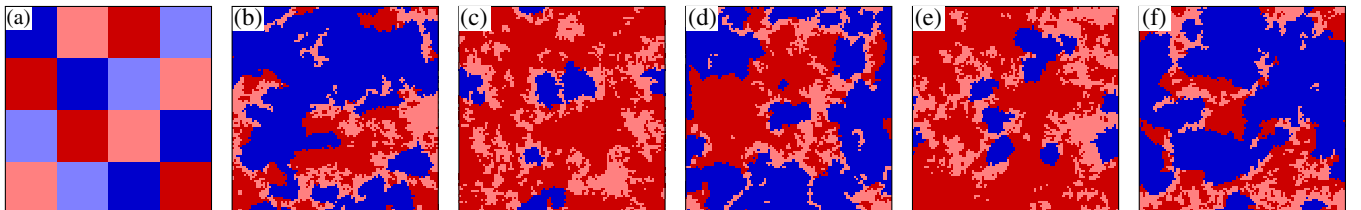

FIG. 6. Representative spatial evolution of the four competing strategies from a prepared initial state towards the three-strategy $D + P_D + P_C$ phase that is sustained by cyclic dominance. Note that blue and red colors dominate cyclically over the course of 16 000 MCS from left to right. The colors used are the same as in Figs. 4 and 5. Depicted are snapshots of the square lattice, as obtained for $\beta = 0.52$, $\gamma = 0.065$, and $r = 3$. Since a prepared initial state is used, a small square lattice with linear size $L = 100$ can be used for demonstration.

neither strategy dies out, and hence the three-strategy $D + P_D + P_C$ phase is stable.

The two representative cross sections of the phase diagram in Fig. 7 show that the average fraction of competing strategies changes similarly as in the canonical rock-paper-scissors model [74–77]. The stability of the three-strategy $D + P_D + P_C$ phase hinges strongly on the continuous, albeit oscillating and sometimes nearly vanishing, presence of all three strategies. As the inset in the bottom panel of Fig. 7 shows, the average fraction of D players can be extremely low, and therefore this three-strategy phase can be stable only if the size of the lattice is large enough. We use $L = 5400$ to produce the results presented in Fig. 7. If the size of the lattice would be smaller, a strategy could easily die out due to random fluctuations, in which case the evolution would terminate into a single-strategy phase.

The invasion rates within the three-strategy $D + P_D + P_C$ phase, which exists under conditions that are more favorable for the survival of the P_D strategy—specifically, if the

cost of punishment γ is lower—hence stabilizing the three-strategy phase, can be measured directly by monitoring the fractions of strategies when the evolution is initialized from straight domain interfaces [78,79]. While the meaning of w_1 and w_2 (see the inset in Fig. 8) is clear from the payoff differences, the determination of w_3 requires further clarification. Namely, if only P_D and D players would be present along the interface, then we would of course measure a net zero invasion rate because the two strategies are neutral in the absence of cooperators. However, since we are interested in their relation when P_C players are present too, we use parallel interfaces of P_C and P_D players, separated by thin (width of 5 lattice sites) layers of D players. In this way, although P_C and P_D players do not interact directly, the setup properly describes the movement of a D layer that is followed by punishing cooperators. This "effective" invasion, which emerges only in the presence of the third party, is highlighted by a dashed arrow in the legend of Fig. 8.

The decay of w_2 in the main panel of Fig. 8 highlights that the $P_C \to D$ invasions are relevant and in fact occur rather frequently, but also that their intensity deteriorates as the cost of punishment increases. Similarly, the $P_D \to P_C$ invasions are also a recurring phenomenon based on the positive value of the corresponding invasion rate w_1, which indicates that P_D players would dominate P_C players during a direct competition as a consequence of the small value of r. Nevertheless, this invasion rate also decays slightly as γ increases, because the costs associated with the main public goods game become less decisive comparing to the costs of punishment that both these strategies should
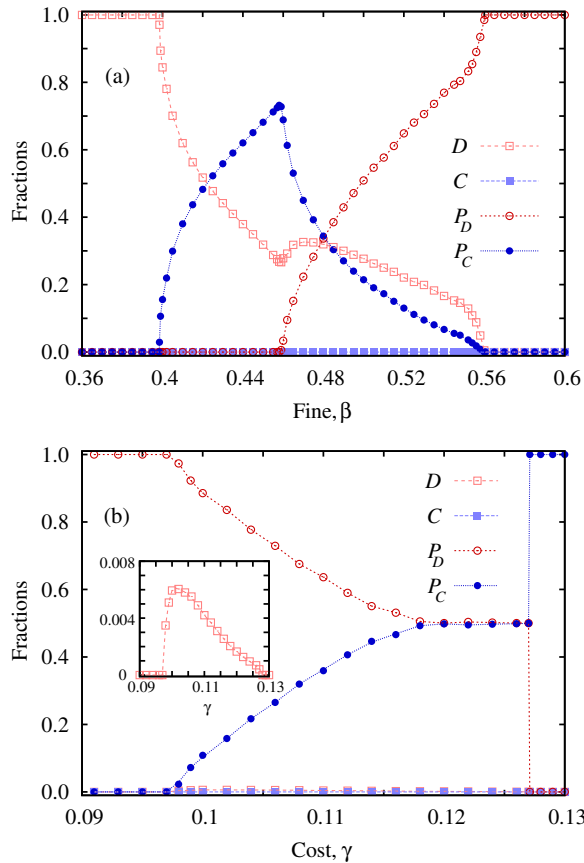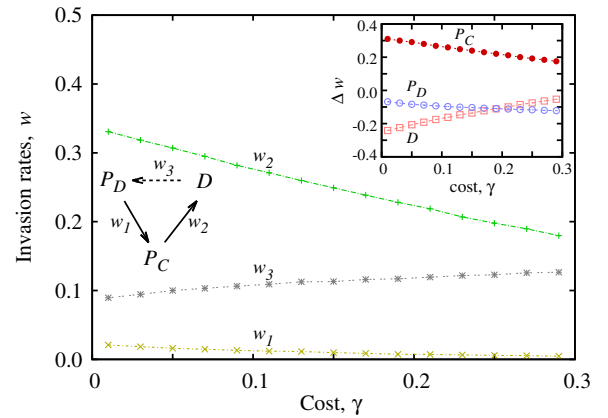

FIG. 7. Two representative cross sections of the phase diagram depicted in Fig. 7, as obtained for the punishment cost $\gamma = 0.04$ (a) and the punishment fine $\beta = 0.9$ (b). Depicted are stationary fractions of the four competing strategies dependent on the other punishment parameter from the one used for the cross section. The inset in (b) shows just how tiny the fraction of nonpunishing defectors in the stationary state can be.

FIG. 8. Invasion rates w_i between three competing strategies (see inset) dependent on the punishment cost γ, as obtained for $\beta = 0.8$ and $r = 3.0$. We note that D invades P_D only in the vicinity of P_C players, since otherwise the two defecting strategies are neutral. This fact is indicated by a dashed w_3 arrow. For details on how this specific invasion rate is measured, see the main text. The inset shows the difference Δw between different pairs of invasion rates, which explains how and why the fraction of strategies changes as a result of the increasing punishment cost (see main text for details).

bear. Lastly, the $w_3(\gamma)$ function is also always positive, because even a small cost evokes the second-order free-riding effect (in this case, of course, associated with avoiding the costs of antisocial punishment), such that in the presence of P_C players D players can invade P_D players. Accordingly, as the value of γ increases, so does the w_3 invasion rate, as shown in Fig. 8, which illustrates directly that second-order free-riding on antisocial punishment is responsible for the evolutionary success of prosocial punishment, specifically for the survival or even for the complete dominance of the P_C strategy.

The differences between these three invasion rates, depicted in the inset of Fig. 8, allow us to understand how the relative abundance of strategies changes as a result. For example, $w_3 - w_2$ quantifies how the fraction of nonpunishing defectors changes when we vary the cost of punishment. An increase in the value of γ will support strategy D, but the actual beneficiary will be her predator, which is strategy P_C. This seemingly paradoxical response of the population to the increase of the punishment cost is a well-known consequence of cyclic dominance, i.e., when directly supporting a particular species will actually support her predator [80]. This in turn explains why P_C players rise to full dominance when we increase the value of γ, as well as why P_D players dominate when we decrease it. Namely, decreasing the punishment cost supports the P_C strategy, which is the prey of punishing defectors.

IV. DISCUSSION

We show how second-order free-riding on antisocial punishment restores the effectiveness of prosocial punishment, thus providing an unlikely and counterintuitive evolutionary escape from adverse effects of antisocial punishment. When the synergistic effects of cooperation are low to the point of network reciprocity failing to sustain it, cooperators that punish defectors can still rise to dominance because nonpunishing defectors enable their evolutionary success by capitalizing on second-order free-riding and eliminating antisocial punishers as a result. If conditions for punishment are somewhat more lenient, we show that a three-strategy phase consisting of nonpunishing defectors, punishing cooperators, and punishing defectors becomes stable. The relations within this phase, and its termination to an absorbing punishing cooperator or an absorbing punishing defector phase, can be fully understood in terms of invasion rates along straight interfaces that separate different strategy domains. We demonstrate that these results are robust to changes in the microscopic dynamics, and we emphasize that the only important property of the interaction structure is the limited interaction range rather than its topological details. Indeed, the mechanism relies solely on spatial pattern formation, and is the first stand-alone remedy against adverse effects of antisocial punishment, not relying on any additional strategic complexity or other assumptions limiting its general

validity. Paradoxically, it turns out that antisocial punishment is vulnerable to the same second-order free-riding that is traditionally held responsible for preventing evolutionary stability of prosocial punishment.

We emphasize that these phenomena cannot be observed in well-mixed populations. Furthermore, a reliable study of competing subsystem solutions requires a careful finite-size analysis of the spatial system. Additionally, the usage of random initial conditions may be misleading, especially if using a small system size, because it does not necessarily allow for all possible subsystem solutions to emerge (before they could compete with one another). These difficulties can be overcome by using suitable prepared initial states, which allow the evolutionary stable subsystem solution to form before competition between them unfolds. Since pattern formation and invasions of propagating fronts are general features of multistrategy complex systems, such an analysis is a must when determining the consequences of spatiality.

Our research also reveals that under conditions that favor cooperation, for example, when the multiplication factor of the public goods game is sufficiently high for the spatial selection alone to sustain cooperation, antisocial punishment is overall uncompetitive. In fact, even though we use a fully symmetrical implementation of prosocial and antisocial punishment throughout our paper, antisocial punishers could survive only in very small regions of the parameter space. We may thus conclude that under such conditions cooperators that punish defectors should not be afraid of retaliatory antisocial punishment by defectors.

In comparison to previous findings concerning the symmetrical implementation of prosocial and antisocial rewarding [54], we find that with punishment there is no lower bound on the multiplication factor that would be impossible to compensate with a sufficiently effective punishment system. But there is a lower bound in case of rewarding: i.e., below a critical value of the multiplication factor full defection is unavoidable, and this regardless of just how efficient the rewarding system might be. In the case of punishment, ever lower values of the multiplication factor simply require ever higher fines at a given cost for cooperation to be sustained. Quite remarkably, the very same process puts a noose around antisocial punishers, which are defeated by second-order free-riding in their own ranks.

Although social preference models of economic decision making predict that antisocial punishment should not occur [81,82], and despite the fact that antisocial punishment is also inconsistent with rational self-interest and the hypothesis that punishment facilitates cooperation, it is nevertheless remarkably common across human societies [32–37]. In the light of this fact, it is important to extend the theory of cooperation in the spatial public goods game with the option that noncooperators can punish cooperators. Rather unexpectedly, the detrimental effects of such

antisocial punishment on the coevolution of punishment and cooperation turn out to be minor simply by taking into account the fact that the interactions among humans are inherently structured, entailing a limited number of frequently used links, rather than being random or well mixed.

ACKNOWLEDGMENTS

[1] R. Axelrod, *The Evolution of Cooperation* (Basic Books, New York, 1984).

[2] J. Henrich, Richard Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath, *In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies*, Am. Econ. Rev. **91**, 73 (2001).

[3] E. Fehr and U. Fischbacher, *Social Norms and Human Cooperation*, Trends Cognit. Sci. **8**, 185 (2004).

[4] S. Gächter, B. Herrmann, and C. Thöni, *Culture and Cooperation*, Phil. Trans. R. Soc. B **365**, 2651 (2010).

[5] D. G. Rand and M. A. Nowak, *Human Cooperation*, Trends Cognit. Sci. **17**, 413 (2013).

[6] V. Capraro, C. Smyth, K. Mylona, and G. A. Niblo, *Benevolent Characteristics Promote Cooperative Behaviour among Humans*, PLoS One **9**, e102881 (2014).

[7] O. P. Hauser, D. G. Rand, A. Peysakhovich, and M. A. Nowak, *Cooperating with the Future*, Nature (London) **511**, 220 (2014).

[8] S. B. Hrdy, *Mothers and Others: The Evolutionary Origins of Mutual Understanding* (Harvard University Press, Cambridge, MA, 2011).

[9] M. A. Nowak and R. Highfield, *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed* (Free Press, New York, 2011).

[10] T. Yamagishi, *The Provision of a Sanctioning System as a Public Good*, J. Pers. Soc. Psychol. **51**, 110 (1986).

[11] E. Fehr and S. Gächter, *Cooperation and Punishment in Public Goods Experiments*, Am. Econ. Rev. **90**, 980 (2000).

[12] O. Gurerk, B. Irlenbusch, and B. Rockenbach, *The Competitive Advantage of Sanctioning Institutions*, Science **312**, 108 (2006).

[13] B. Rockenbach and M. Milinski, *The Efficient Interaction of Indirect Reciprocity and Costly Punishment*, Nature (London) **444**, 718 (2006).

[14] S. Gächter, E. Renner, and M. Sefton, *The Long-Run Benefits of Punishment*, Science **322**, 1510 (2008).

[15] A. Ule, A. Schram, A. Riedl, and T. N. Cason, *Indirect Punishment and Generosity toward Strangers*, Science **326**, 1701 (2009).

[16] J. J. Jordan, M. Hoffman, P. Bloom, and D. G. Rand, *Third-Party Punishment as a Costly Signal of Trustworthiness*, Nature (London) **530**, 473 (2016).

[17] E. Fehr, *Don't Lose Your Reputation*, Nature (London) **432**, 449 (2004).

[18] M. Milinski and B. Rockenbach, *Punisher Pays*, Nature (London) **452**, 297 (2008).

[19] R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson, *The Evolution of Altruistic Punishment*, Proc. Natl. Acad. Sci. U.S.A. **100**, 3531 (2003).

[20] J. H. Fowler, *Altruistic Punishment and the Origin of Cooperation*, Proc. Natl. Acad. Sci. U.S.A. **102**, 7047 (2005).

[21] M. Nakamaru and Y. Iwasa, *The Coevolution of Altruism and Punishment: Role of the Selfish Punisher*, J. Theor. Biol. **240**, 475 (2006).

[22] C. Hauert, A. Traulsen, H. Brandt, M. A. Nowak, and K. Sigmund, *Via Freedom to Coercion: The Emergence of Costly Punishment*, Science **316**, 1905 (2007).

[23] K. Sigmund, *Punish or Perish? Retaliation and Collaboration among Humans*, Trends Ecol. Evol. **22**, 593 (2007).

[24] R. Boyd, H. Gintis, and S. Bowles, *Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare*, Science **328**, 617 (2010).

[25] T. Sasaki, A. Brännström, U. Dieckmann, and K. Sigmund, *The Take-It-or-Leave-It Option Allows Small Penalties to Overcome Social Dilemmas*, Proc. Natl. Acad. Sci. U.S.A. **109**, 1165 (2012).

[26] M. Perc and A. Szolnoki, *Self-Organization of Punishment in Structured Populations*, New J. Phys. **14**, 043013 (2012).

[27] J. J. Jordan and D. G. Rand, *Third-Party Punishment as a Costly Signal of High Continuation Probabilities in Repeated Games*, J. Theor. Biol. **421**, 189 (2017).

[28] K. Panchanathan and R. Boyd, *Indirect Reciprocity Can Stabilize Cooperation without the Second-Order Free Rider Problem*, Nature (London) **432**, 499 (2004).

[29] D. Helbing, A. Szolnoki, M. Perc, and G. Szabó, *Evolutionary Establishment of Moral and Double Moral Standards through Spatial Interactions*, PLoS Comput. Biol. **6**, e1000758 (2010).

[30] C. Hilbe and A. Traulsen, *Emergence of Responsible Sanctions without Second Order Free Riders, Antisocial Punishment or Spite*, Sci. Rep. **2**, 458 (2012).

[31] X. Chen, A. Szolnoki, and M. Perc, *Probabilistic Sharing Solves the Problem of Costly Punishment*, New J. Phys. **16**, 083016 (2014).

[32] L. Denant-Boemont, D. Masclet, and C. N. Noussair, *Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment*, Economic theory **33**, 145 (2007).

[33] M. A. Janssen, R. Holahan, A. Lee, and E. Ostrom, *Lab Experiments for the Study of Social-Ecological Systems*, Science **328**, 613 (2010).

[34] L. Balafoutas, L. Grechenig, and N. Nikiforakis, *Third-Party Punishment and Counter-Punishment in One-Shot Interactions*, Economics letters **122**, 308 (2014).

[35] J.-J. Wu, B.-J. Zhang, Z.-X. Zhou, Q.-Q. He, X.-D. Zheng, R. Cressman, and Y. Tao, *Costly Punishment Does Not Always Increase Cooperation*, Proc. Natl. Acad. Sci. U.S.A. **106**, 17448 (2009).

[36] S. Gächter and B. Herrmann, *The Limits of Self-Governance When Cooperators Get Punished: Experimental Evidence from Urban and Rural Russia*, European economic review **55**, 193 (2011).

[37] B. Herrmann, C. Thoni, and S. Gächter, *Antisocial Punishment across Societies*, Science **319**, 1362 (2008).

[38] D. G. Rand, H. Ohtsuki, and M. A. Nowak, *Direct Reciprocity with Costly Punishment: Generous Tit-for-Tat Prevails*, J. Theor. Biol. **256**, 45 (2009).

[39] D. G. Rand, J. J. Armao, M. Nakamaru, and H. Ohtsuki, *Anti-Social Punishment Can Prevent the Co-Evolution of Punishment and Cooperation*, J. Theor. Biol. **265**, 624 (2010).

[40] D. G. Rand and M. A. Nowak, *The Evolution of Antisocial Punishment in Optional Public Goods Games*, Nat. Commun. **2**, 434 (2011).

[41] M. Cinyabuguma, T. Page, and L. Putterman, *Can Second-Order Punishment Deter Perverse Punishment?*, Experimental Economics **9**, 265 (2006).

[42] A. Dreber, D. G. Rand, D. Fudenberg, and M. A. Nowak, *Winners Don't Punish*, Nature (London) **452**, 348 (2008).

[43] D. G. Rand, A. Dreber, T. Ellingsen, D. Fudenberg, and M. A. Nowak, *Positive Interactions Promote Public Cooperation*, Science **325**, 1272 (2009).

[44] C. Hilbe and K. Sigmund, *Incentives and Opportunism: From the Carrot to the Stick*, Proc. R. Soc. B **277**, 2427 (2010).

[45] C. Hauert, *Replicator Dynamics of Reward & Reputation in Public Goods Games*, J. Theor. Biol. **267**, 22 (2010).

[46] A. Szolnoki and M. Perc, *Reward and Cooperation in the Spatial Public Goods Game*, Europhys. Lett. **92**, 38003 (2010).

[47] T. Sasaki and T. Unemi, *Replicator Dynamics in Public Goods Games with Reward Funds*, J. Theor. Biol. **287**, 109 (2011).

[48] A. Szolnoki and M. Perc, *Evolutionary Advantages of Adaptive Rewarding*, New J. Phys. **14**, 093016 (2012).

[49] X. Chen, T. Sasaki, Å. Brännström, and U. Dieckmann, *First Carrot, Then Stick: How the Adaptive Hybridization of Incentives Promotes Cooperation*, J. R. Soc. Interface **12**, 20140935 (2015).

[50] T. Sasaki and S. Uchida, *Rewards and the Evolution of Cooperation in Public Good Games*, Biol. Lett **10**, 20130903 (2014).

[51] I. Okada, H. Yamamoto, F. Toriumi, and T. Sasaki, *The Effect of Incentives and Meta-Incentives on the Evolution of Cooperation*, PLoS Comput. Biol. **11**, e1004232 (2015).

[52] T. Sasaki, S. Uchida, and X. Chen, *Voluntary Rewards Mediate the Evolution of Pool Punishment for Maintaining Public Goods in Large Populations*, Sci. Rep. **5**, 8917 (2015).

[53] M. dos Santos, *The Evolution of Anti-Social Rewarding and Its Countermeasures in Public Goods Games*, Proc. R. Soc. B **282**, 20141994 (2015).

[54] A. Szolnoki and M. Perc, *Antisocial Pool Rewarding Does Not Deter Public Cooperation*, Proc. R. Soc. B **282**, 20151975 (2015).

[55] K. Sigmund, *The Calculus of Selfishness* (Princeton University Press, Princeton, NJ, 2010).

[56] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, *Statistical Physics of Human Cooperation*, Phys. Rep. **687**, 1 (2017).

[57] M. A. Nowak and R. M. May, *Evolutionary Games and Spatial Chaos*, Nature (London) **359**, 826 (1992).

[58] A. Szolnoki, M. Perc, and G. Szabó, *Topology-Independent Impact of Noise on Cooperation in Spatial Public Goods Games*, Phys. Rev. E **80**, 056109 (2009).

[59] A. Szolnoki, G. Szabó, and M. Perc, *Phase Diagrams for the Spatial Public Goods Game with Pool Punishment*, Phys. Rev. E **83**, 036101 (2011).

[60] A. Szolnoki and M. Perc, *Correlation of Positive and Negative Reciprocity Fails to Confer an Evolutionary Advantage: Phase Transitions to Elementary Strategies*, Phys. Rev. X **3**, 041021 (2013).

[61] A. M. Colman, *Cooperation, Psychological Game Theory and Limitations of Rationality in Social Interactions*, Behav. Brain Sci. **26**, 139 (2003).

[62] R. Cook, G. Bird, G. Lunser, S. Huck, and C. Heyes, *Automatic Imitation in a Strategic Context: Players of Rock-Paper-Scissors Imitate Opponents' Gesture*, Proc. R. Soc. B **279**, 780 (2012).

[63] M. Naber, M. V. Pashkam, and K. Nakayama, *Unintended Imitation Affects Success in a Competitive Game*, Proc. Natl. Acad. Sci. U.S.A. **110**, 20046 (2013).

[64] M. Buckert, J. Oechssler, and C. Schwieren, *Imitation under Stress*, J. Econ. Behav. Organ. **139**, 252 (2017).

[65] M. A. Nowak, *Evolutionary Dynamics* (Harvard University Press, Cambridge, MA, 2006).

[66] D. Helbing, A. Szolnoki, M. Perc, and G. Szabó, *Punish, but Not Too Hard: How Costly Punishment Spreads in the Spatial Public Goods Game*, New J. Phys. **12**, 083005 (2010).

[67] H. Brandt, C. Hauert, and K. Sigmund, *Punishment and Reputation in Spatial Public Goods Games*, Proc. R. Soc. B **270**, 1099 (2003).

[68] T. M. Liggett, *Interacting Particle Systems* (Springer, New York, 1985).

[69] I. Dornic, H. Chaté, J. Chave, and H. Hinrichsen, *Critical Coarsening without Surface Tension: The Universality Class of the Voter Model*, Phys. Rev. Lett. **87**, 045701 (2001).

[70] J. T. Cox and D. Griffeath, *Diffusive Clustering in the Two Dimensional Voter Model*, Ann. Probab. **14**, 347 (1986).

[71] See http://dx.doi.org/10.6084/m9.figshare.5403133.

[72] M. Nakamaru, H. Matsuda, and Y. Iwasa, *The Evolution of Cooperation in a Lattice-Structured Population*, J. Theor. Biol. **184**, 65 (1997).

[73] See http://dx.doi.org/10.6084/m9.figshare.5402647.

[74] A. Szolnoki, M. Mobilia, L.-L. Jiang, B. Szczesny, A. M. Rucklidge, and M. Perc, *Cyclic Dominance in Evolutionary Games: A Review*, J. R. Soc. Interface **11**, 20140735 (2014).

[75] T. Reichenbach, M. Mobilia, and E. Frey, *Mobility Promotes and Jeopardizes Biodiversity in Rock-Paper-Scissors Games*, Nature (London) **448**, 1046 (2007).

[76] J. Juul, K. Sneppen, and J. Mathiesen, *Clonal Selection Prevents Tragedy of the Commons When Neighbors Compete in a Rock-Paper-Scissors Game*, Phys. Rev. E **85**, 061924 (2012).

[77] B. Intoy and M. Pleimling, *Synchronization and Extinction in Cyclic Games with Mixed Strategies*, Phys. Rev. E **91**, 052135 (2015).

[78] A. Szolnoki, J. Vukov, and M. Perc, *From Pairwise to Group Interactions in Games of Cyclic Dominance*, Phys. Rev. E **89**, 062125 (2014).

[79] J. Vukov, A. Szolnoki, and G. Szabó, *Diverging Fluctuations in a Spatial Five-Species Cyclic Dominance Game*, Phys. Rev. E **88**, 022123 (2013).

[80] M. Frean and E. D. Abraham, *Rock-Scissors-Paper and the Survival of the Weakest*, Proc. R. Soc. B **268**, 1323 (2001).

[81] E. Fehr and K. M. Schmidt, *A Theory of Fairness, Competition, and Cooperation*, Q. J. Econ. **114**, 817 (1999).

[82] M. Rabin, *Incorporating Fairness into Game Theory and Economics*, Am. Econ. Rev. **83**, 1281 (1993).