Suspiciousness and the option not to interact facilitates trustful cooperation in prisoners dilemma

Tadeas Priklopil*[a,b], Krishnendu Chatterjee[b] and Martin Nowak[c]

[a]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[b]Institute of Science and Technology IST Austria (IST Austria), Am Campus 1, A-3400 Klosterneuburg

[c]Program for Evolutionary Dynamics, Department of Mathematics, and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

Keywords: evolutionary game theory, optional interactions, evolution of cooperation, non-social behaviour, partial information

*Corresponding author. E-mail: tadeas.priklopil@unil.ch

Abstract

In evolutionary game theory interactions between individuals are often assumed obligatory. However, in many real-life situations, individuals can decide to opt out of an interaction depending on the information they have about the opponent. We consider a simple evolutionary game theoretic model to study such a scenario, where at each encounter between two individuals the type of the opponent (cooperator/defector) is known with some probability, and where each individual either accepts or opts out of the interaction. If the type of the opponent is unknown, a trustful individual accepts the interaction, whereas a suspicious individual opts out of the interaction. If either of the two individuals opt out both individuals remain without an interaction. We show that in the prisoners dilemma optional interactions along with suspicious behaviour facilitates the emergence of trustful cooperation.

INTRODUCTION

Evolutionary games provide a general framework to study frequency dependent selection, where the fitness (payoff) of each individual is determined by playing a game with other individuals in the population. In the standard formulation, games between individuals are considered compulsory in the sense that individuals have no choice of whom they encounter, and are then forced to execute their strategy with the encountered individual (e.g. Weibull 1995). In nature, however, this is usually not the case. Various models have accounted for this by allowing individuals to be selective about their opponents either in terms of partner choice (”pre-interaction decisions”, e.g. Hruschka and Henrich 2006, Fu et al. 2008) and/or partner switching (”post-interaction decisions”, e.g. Hruschka and Henrich 2006, McNamara et al. 2008, Fu et al. 2008, Fujiwara-Greve and Okuno-Fujiwara 2009, Izquierdo et al. 2010, Wubs et al. 2016, Zheng et al. 2017), or allowing individuals to opt out of interactions altogether (”optional interactions”, e.g. Miller 1967, Vanberg and Congleton 1992, Orbell and Dawes 1993, Stanley et al. 1995, Batali and Kitcher 1995, Sherratt and Roberts 1998, Hauert et al. 2002a, Mathew and Boyd 2009, Ghang and Nowak 2015). Some models make both assumptions, individuals have the ability to influence the choice of their opponents as well as have the option to opt out, or to be forced to opt out, of interactions (Noë and Hammerstein 1994, Batali and Kitcher 1995, Hruschka and Henrich 2006). Here, we focus on optional interactions whilst assuming that opponents are chosen at random.

An extremely simple form of optional interactions is to accept no interactions, which is the so-called loners strategy (Hauert et al. 2002a,b, 2007, Fowler 2005, Brandt et al. 2006, Mathew and Boyd 2009, Cardinot et al. 2016). Individuals who adopt a loners strategy opt out of all interactions and receive a fixed ”loners payoff”. In evolutionary games with loners along with cooperators and defectors, where cooperators and defectors are assumed to accept every interaction, the evolutionary trajectories approach a cycle between the three strategies (Hauert et al. 2002a,b). This is an interesting result, particularly because in such models individuals have no information about their opponents.

While no information and no interaction represents an extreme scenario, in many situations individuals can base their decision to interact on partial information about their opponent. (deleted: In other words, individuals can predict with some accuracy the upcoming action of their opponent.) Classical examples where individuals in the population have at least some information about each other are as follows: (a) models of direct reciprocity: individuals have encountered their opponent in the past (Trivers 1971, Batali and Kitcher 1995, Sherratt and Roberts 1998, Castro and Toro 2008, Spichtig et al. 2013, Kurokawa 2017); (b) models of indirect reciprocity: the opponent has build a reputation of its past actions with other individuals (Nowak and Sigmund 1998a,b, Panchanathan and Boyd 2003, Nowak and Sigmund 2005, Fu et al. 2008, Ghang and Nowak 2015); or (c) the opponent appears or behaves a certain way before an interaction takes place that indicates its intended actions (Frank et al. 1993, Yamagishi et al. 1999, Reed et al. 2012, DeSteno et al. 2012). For example, the ability of correctly evaluating mate selection-related strategies of other individuals is common (Zahavi 1975, Iwasa et al. 1991, Jennions and Petrie 1997, Andersson and Simmons 2006). In such situations, and in contrast to loners strategy of always opting out, the decision of opting out or accepting the interaction ought to depend on the available partial information.

In this work we introduce a simple evolutionary game-theoretical model where the individuals encounter each other at random (no choice of opponents), but at each encounter they are given the option to either accept or opt out of the interaction based on partial information about their opponent. If either of the two individuals opt out, both individuals remain without an interaction. In our model the type of the opponent (cooperator or defector) is known with some fixed probability. If the type of the opponent is known, then individuals take a decision (accept or opt out) that yields a greater payoff. If the type of the opponent is unknown, then individuals can be either trustful or suspicious (Panchanathan and Boyd 2003, Sigmund 2010). A trustful individual accepts an interaction with the trust that the opponent will provide a greater payoff than opting out, and a suspicious individual opts out of an interaction suspecting that the opponent will provide a lesser payoff than what opting out yields. The strategy of an individual is thus a combination of its type (cooperator/defector) and a decision rule that dictates whether to accept or opt out of an interaction (trustful/suspicious).

We formally introduce our modeling framework in the following section, and then as an example, study the evolution of cooperation by working out the game of prisoners dilemma in detail. We succinctly summarize our key findings below.

- First, if the probability of knowing the type of the opponent is above a certain threshold, a threshold that is given in terms of payoffs, then trustful cooperation is an ESS. A similar condition was derived in (Nowak and Sigmund 1998a,b, Suzuki and Toquenaga 2005, Ghang and Nowak 2015). Interestingly, and in contrast to the previous findings, if opting out yields an equal or greater payoff than mutual defection, then trustful cooperation is a globally convergent ESS, i.e., trustful cooperation is reached from any initial state of the population. In particular, even an (almost) entirely defective population will be eventually replaced by trustful cooperators.

- Second, we consider that the probability of knowing the type of the opponent is below the required threshold. If opting out is at least as beneficial as mutual defection, then the evolutionary dynamics approaches a rock-paper-scissors cycle of trustful cooperation, trustful defection and suspicious cooperation. However, if opting out is strictly better than mutual defection, then for a low probability of knowing the type of the opponent, trustful cooperation, trustful defection and suspicious cooperation coexist at a globally stable equilibrium. We note that suspicious defection is always (eventually) selected against and thus eradicated form the population.

To summarize, we introduce a simple mathematically tractable model that enables us to study the interplay between social and non-social behavior. We apply our model to the game of prisoners dilemma where we show that the option of non-social behaviour of opting out of interactions, a ”natural precondition” of partner formation, allows for the emergence of (social and) cooperative behaviour. Moreover, we find that non-social behaviour together with the ability to recognise the behaviour of each other leads not only to stable cooperative populations but also to trustful behaviour that accepts interactions with potentially defective players.

MODEL DESCRIPTION

Consider a large and well-mixed population with two types of players, cooperators and defectors. Players are assumed to encounter each other at random, such that at each encounter they can either accept or reject each other for an interaction. If both players accept, a game is played and a payoff is received: if both players are cooperators both receive R, if both players are defectors both receive P, and if one is a defector and the other is a cooperator then the defector receives T and the cooperator S, such that $S < P < R < T$. A game is not played if at least one of the two players rejects the interaction (opt out), in which case both players receive a payoff L, where L can be any value relative to the payoffs S, P, R, T. Without loss of generality we set $L = 0$ and scale the other payoffs accordingly (SI). The payoffs S, P, R, T thus need to be reinterpreted as the difference between the particular social interaction and non-social behaviour. We note that each player knows its own type as well as the ordering of payoffs.

The decision to accept or opt out of an interaction is made based on the type of the opponent, which is known to the player with some fixed probability q. If the type of the opponent is known the decision to interact is obvious – a game that yields a greater payoff than opting out will be accepted and with a smaller payoff rejected. This is illustrated with the left branch in Figure 1 where a player of type A has identified the type of the encountered opponent B. The question is what to do when the opponent is unknown (the right branch in Figure 1). Since players have no information about the composition of the population (frequency distribution of cooperators and defectors) they have only two options, either *trust* that by accepting the interaction the unknown player will yield them a greater payoff than if they chose to opt out, or be *suspicious* that the interaction will be advantageous and reject the unknown opponent (Panchanathan and Boyd 2003, Sigmund 2010). All in all we obtain four strategies, trustful cooperation, suspicious cooperation, trustful defection and suspicious defection, keeping in mind that for some payoff configurations not all strategies are rational and hence will not be considered. For example, if mutual defection yields greater payoff than non-social behaviour $0 < P$, then defectors will always receive a greater payoff by accepting an interaction, known and unknown, and thus the strategy of suspicious defection will be disregarded.

We immediately observe that the cases $0 \leq S$ and $R \leq 0$ lead to trivial evolutionary dynamics (Batali and Kitcher 1995). If $0 \leq S$, then any interaction is at least as good as no interaction and thus all games should be accepted, and if $R \leq 0$, then cooperators receive always the maximum payoff by not interacting and so all games end up being rejected. In the first case we recover the dynamics of the prisoners dilemma with obligatory interactions where defective strategy is the evolutionary outcome. In the latter case players of both types opt out of all interactions. Thus, the task is to work out the evolutionary dynamics for the two remaining cases, $S < 0 \leq P < R < T$ and $S < P < 0 < R < T$. We remark that the non-generic case $P = 0$ is of special interest and will be considered separately, not only due to its simple evolutionary dynamics but also because a donation game, the central model in the literature of evolution of cooperation (Sigmund 2010), falls into this category of models when the benefit of defection $T - R$ and the cost of

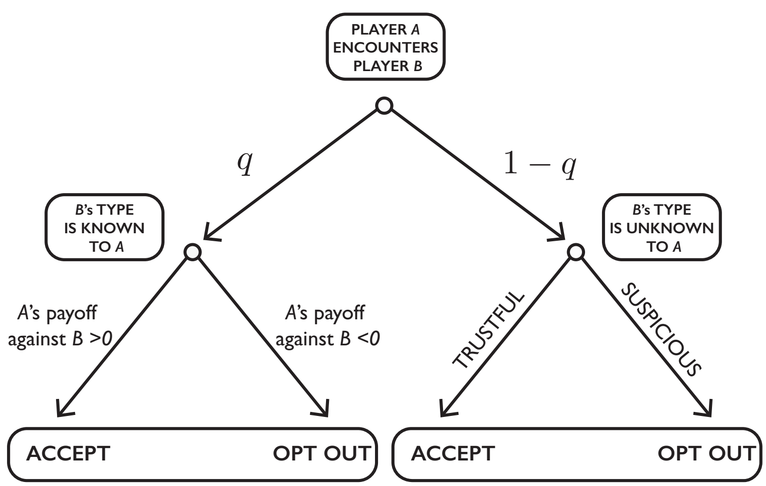

Figure 1: Decision tree for a player of type A. At the top node (yellow dot) nature decides whether the player A identifies the type of the encountered opponent B or not, which happens with probabilities q and $1 - q$, respectively. If player A identifies the type of the encountered opponent (left branch, blue node), the player chooses the action that maximizes its payoff. Thus player A will accept the interaction if the payoff of A against B is greater than 0, otherwise player A will opt out of the interaction. If player A doesn't identify the type of its encountered opponent (right branch, red node), player A can either be trustful or suspicious and will either accept or opt out of the interaction, respectfully.

cooperation $S - P$ are equal.

We may interpret mutual defection as some social interaction that provides basic income P, where the potentially harmful effect of the interaction is already factored in the payoff. Depending on the level of harm defection causes to the co-player, the payoff for mutual defection may be greater or smaller than the payoff for non-social (solitary) behaviour. Equivalently, and this is the terminology we use throughout the paper, we say that opting out is costly when $P > 0$ and beneficial when $P < 0$.

RESULTS

We will first work out a model for the two limiting cases where players have either zero information $q = 0$ or perfect information $q = 1$ about their opponents. In the following sections we will consider games with partial information $0 < q < 1$ and first deal with the special case $P = 0$ where opting out and mutual defection results in equal payoff. Lastly we solve the two remaining cases, $S < 0 < P$ where opting out is costly and $P < 0 < R$ where opting out is beneficial. For each model we analyse the evolutionary dynamics represented with a continuous-time replicator equation

$$\dot{x}_A = x_A \left(E_A - \bar{E} \right) \tag{1}$$

where the dot denotes a time derivative, x_A is the frequency and E_A is the expected payoff of strategy A, and $\bar{E} = \sum_B x_B E_B$ is the average payoff in the population.

Games with zero and perfect information

Let us first consider the case where players have zero information about the type of the opponent $q = 0$ and so all interactions are between unknown players. In both non-trivial cases $S < 0 \leq P < R < T$ and $S < P < 0 < R < T$ we have $S < 0 < R$, and so the decision for a cooperator to accept or opt out of an interaction with an (always) unknown opponent depends whether the opponent is likely to be a cooperator or a defector. If the unknown opponent is likely to be a defector it pays off to opt out, but if the opponent is likely to be a cooperator it pays off to accept the interaction. We thus need to consider both suspicious and trustful cooperators, where suspicious cooperators opt out of all interactions, while trustful cooperators accept every interaction. Similarly, if $P < 0 < R$ defectors may either be suspicious and opt out of all interactions or be trustful and always defect. However, for $S < 0 \leq P$ all defectors ought to be trustful and accept every interaction. In this case suspicious defectors will not be considered. We thus need to consider only three simple strategies, suspicious strategies (i.e. suspicious cooperators and for $P < 0 < R$ also suspicious defectors) who opt out of every interaction, trustful cooperators and trustful defectors who accept every interaction. The expected payoff for suspicious strategies is always 0 while for trustful strategies the payoffs are

$$\begin{aligned} f_1 &= x_1 R + y_1 S \\ g_1 &= x_1 T + y_1 P, \end{aligned} \tag{2}$$

where f_1, g_1 are the expected payoffs and x_1, y_1 are the frequencies of trustful cooperators and trustful defectors, respectfully. We will use subscript 1 to denote trustful players, and we reserve subscript 0 to denote suspicious players. The subscripts can be thought of representing the probability of accepting unknown opponents.

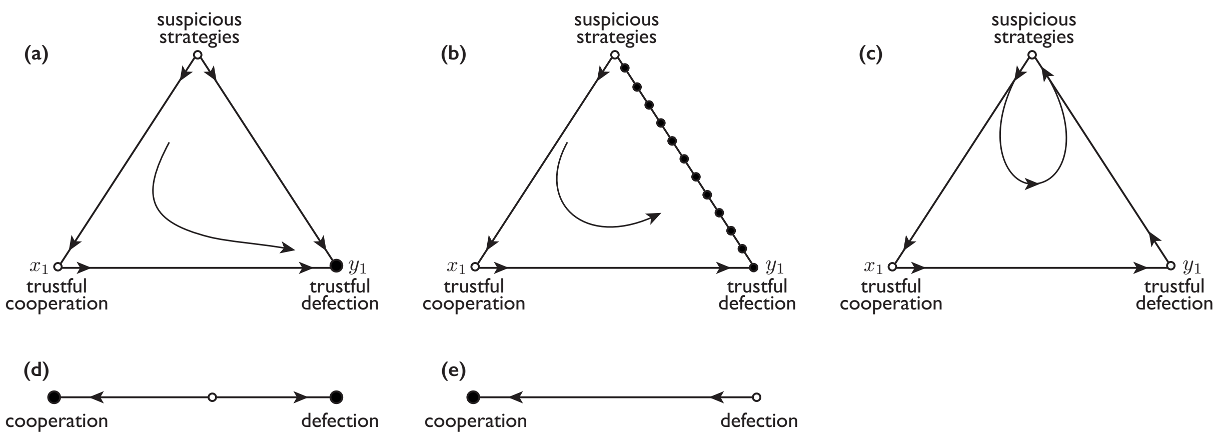

Figure 2: Top row: Evolutionary dynamics (1) for a model (2) with zero information $q = 0$. In (a) $0 < P$ all trajectories approach trustful defection (b) $P = 0$ all trajectories approach the line of equilibria spanned by suspicious strategies and trustful defection (c) $P < 0$ all trajectories approach suspicious strategies. Note that the boundary is a heteroclinic cycle. Bottom row: Evolutionary dynamics for a model with perfect information $q = 1$. In (d) $0 < P$ cooperation and defection are locally attracting, separated by an unstable equilibrium. In (e) $P \leq 0$ all trajectories approach cooperation.

The evolutionary dynamics of this model can be solved fully analytically and the results are depicted in Figure 2. In Figure 2(a) where $0 < P$, all trajectories approach trustful defection. In Figure 2 (b) where $P = 0$, all trajectories approach the line of equilibria spanned by suspicious strategies and trustful defection, and in Figure 2(c) where $P < 0$, all trajectories approach suspicious strategies. Note that in the last case the boundary is a heteroclinic cycle. This model was analysed in the context of public goods game by Hauert et al. (2002a,b).

If players have perfect information about the type of the opponent $q = 1$, it is nonsensical to distinguish between suspicious and trustful strategies as all opponents are known. In both non-trivial cases $S < 0 \leq P < R < T$ and $S < P < 0 < R < T$ cooperators will only accept interactions with other cooperators, while defectors will accept defectors only if $0 < P$. For all payoffs no games between defectors and cooperators are played. The analysis of the evolutionary dynamics is straightforward. If $0 < P$ cooperation and defection are both locally attracting states separated by an unstable equilibrium (Figure 2(d)), and if $P \leq 0$ cooperation is globally attracting (Figure 2(e)).

Games with partial information

In this section we consider models with partial information $0 < q < 1$. The first model we analyze is where opting out of interactions yields no benefits nor costs to the player and so $P = 0$. We analyze this case first because of its simple evolutionary dynamics and because it contains the donation game, a version of prisoners dilemma that has a central role in the literature of the evolution of cooperation (Sigmund 2010).

Opting out yields no benefits nor costs

In this section we assume that opting out yields players the same payoff as mutual defection, i.e. $P = 0$. In such a case, defectors should always accept unknown players since accepting a game guarantees them a payoff that is at least $0\ (\leq P, T)$. Suspicious defection is therefore not a rational strategy and will not be considered. Cooperators, however, may want to accept or opt out of an interaction with an unknown player: if the opponent is likely to be a cooperator, accepting is more beneficial than opting out $0 < R$, but if the opponent is likely to be a defector it is better to opt out $S < 0$. We thus consider three strategies, trustful cooperators who accept a known cooperator and an unknown opponent but reject a known defector, suspicious cooperators who accept a known cooperator but reject everyone else, and trustful defectors who accept all opponents.

To investigate the evolutionary dynamics (1) we calculate the expected payoffs for each strategy

$$\begin{aligned} f_0 &= \left(x_0 q^2 + x_1 q\right) R \\ f_1 &= \left(x_0 q + x_1\right) R + y_1 (1-q) S \\ g_1 &= x_1 (1-q) T, \end{aligned} \tag{3}$$

where similarly to previous section f_0, f_1, g_1 are the expected payoffs and x_0, x_1, y_1 are the frequencies of suspicious cooperators, trustful cooperators and trustful defectors, respectfully.

The evolutionary dynamics (1) with the expected payoffs given in (3) can be analysed fully analytically (see SI) and the results are depicted in Figure 3. In Figure 3(a), where $0 < q < \frac{T-R}{T}$, all trivial equilibria are saddles and because the interior trimorphic equilibrium (x_0, x_1, y_1) is an unstable spiral all trajectories approach the heteroclinic cycle of trustful cooperation, trustful defection and suspicious cooperation (see SI for the exact expression of the interior trimorphic equilibrium and the stability analysis). In Figures 3(b) where $\frac{T-R}{T} < q < \frac{T}{(R(1+\frac{R}{4T})+T}$, trustful cooperation turns into a stable equilibrium, and so all trajectories approach the equilibrium of trustful cooperation. In Figure 3(c) where $\frac{T}{R(1+\frac{R}{4T})+T} < q < \frac{T+R}{T}$, the interior trimorphic equilibrium (x_0, x_1, y_1) changes from an unstable spiral to an unstable node, and in Figure 3 (d) where $\frac{T+R}{T} < q < 1$, the trimorphic equilibrium (x_0, x_1, y_1) exits the interior. In both cases all trajectories approach the equilibrium of trustful cooperation. We remark that in the limiting cases where q approaches 0 or 1 we recover the model with zero $q = 0$ and perfect information $q = 1$, respectfully: as q approaches 0 the trimorphic equilibrium (x_0, x_1, y_1) approaches the equilibrium of suspicious cooperation x_0 and the line spanned by suspicious cooperators x_0 and trustful defectors y_1 turns into a line of equilibria (Figure 2(b)), and as q approaches 1 the unstable dimorphic equilibrium (x_1, y_1) approaches the equilibrium of trustful defection y_1 and so all trajectories approach the equilibrium of trustful cooperation.

The model (with partial information) contains two qualitatively different evolutionary outcomes. First, when

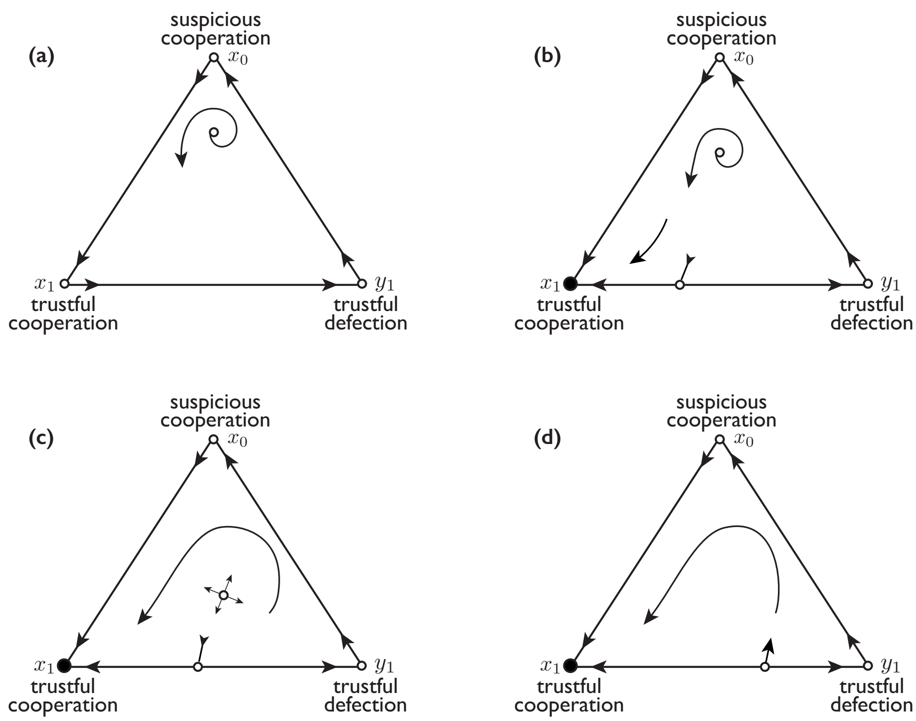

Figure 3: Evolutionary dynamics (1) for a model (3) where opting out is not costly nor beneficial $P = 0$. The parameter values are **(a)** $0 < q < \frac{T-R}{T}$ **(b)** $\frac{T-R}{T} < q < \frac{T}{(R(1+\frac{R}{4T})+T}$ **(c)** $\frac{T}{R(1+\frac{R}{4T})+T} < q < \frac{T+R}{T}$ **(d)** $\frac{T+R}{T} < q < 1$. In each panel in the top node all the players are suspicious cooperators ($x_0 = 1$), in the bottom left node all the players are trustful cooperators ($x_1 = 1$) and in the bottom right node all the players are trustful defectors ($y_1 = 1$). The analytical expressions for the dimorphic and trimorphic equilibria, and their stability conditions, are given in the SI. There are two qualitatively different evolutionary trajectories: In panel (a) $0 < q < \frac{T-R}{T}$ every trajectory approaches the rock-paper-scissors cycle of trustful cooperation, trustful defection and suspicious cooperation, and in panels (b)-(d) $\frac{T-R}{T} < q < 1$ all trajectories converge to a fully trustful cooperation.

$0 < q < \frac{T-R}{T}$, the evolutionary dynamics approaches a heteroclinic rock-paper-scissors cycle of trustful cooperation, trustful defection and suspicious cooperation (Figure 3(a)). This is because for lower values of q most encounters are between unknown players. Therefore (i) almost all games between trustful defectors and trustful cooperators are accepted, and the situation is (almost) identical to the donation game with obligatory interactions where trustful defection beats trustful cooperation (ii) when trustful cooperators are absent both suspicious cooperators and trustful defectors play only amongst themselves, and because cooperative interaction yields higher payoff than defective interactions suspicious cooperators beat trustful defectors (iii) if most players are cooperators, trustful cooperators beat suspicious cooperators because trustful cooperators play more cooperative games by accepting unknown, and therefore cooperative, opponents.

Second, when $\frac{T-R}{T} < q < 1$, the evolutionary outcome is a population of trustful cooperation, independently of the initial (strictly positive) frequency distribution of strategies (Figures 3(b)-(d)). Trustful cooperation is an ESS because for higher values of q a population of trustful cooperators efficiently refuse defective games. This implies that trajectories nearby converge to a fully trustful cooperation. The global convergence is due to the existence of suspicious cooperators as they can invade a population of defectors, and then be eventually replaced by trustful cooperators.

We remark that a similar ESS condition has been derived in Nowak and Sigmund (1998a), Nowak and Sigmund (1998b), Suzuki and Toquenaga (2005) and Ghang and Nowak (2015). There are however two notable differences. Firstly, the condition given in the previous work was derived for a donation game stating that cooperation is an ESS if the probability of knowing the type of the opponent q is greater than the cost to benefit ratio of cooperation. However, our model is derived for the general prisoners dilemma allowing us to make a distinction between the cost of cooperation $P - S$ and the benefit of defection $T - R$ (in the donation game they are equal). The interpretation of the ESS condition then becomes a ratio between the benefit of defection $T - R$, rather than cost of cooperation, and a payoff value which is the difference between unknown and known defectors encountering a trustful cooperator, i.e. T (recall the reinterpretation of the payoff values). Secondly, but more importantly, our condition implies global convergence to trustful cooperation. This is a consequence of allowing decision rules that are optimal when trustful behaviour is not, and therefore, when population consist mainly of defectors, suspicious behaviour becomes the outcompeting social norm which eventually enables the dominance of trustful cooperation.

Opting out is costly

Lets now suppose that players who opt out are strictly worse off than players who mutually defect $S < 0 < P$. Because defectors should accept every interaction whenever $0 \leq P$, the strategies under consideration are identical to the previous model ($P = 0$). The expected payoffs are

$$
\begin{aligned}
f_0 &= \left(x_0 q^2 + x_1 q\right) R \\
f_1 &= \left(x_0 q + x_1\right) R + y_1 (1-q) S \\
g_1 &= x_1 (1-q) T + y_1 P.
\end{aligned}
\tag{4}
$$

The evolutionary dynamics (1) with the expected payoffs given in (4) can be analysed fully analytically (see SI for detailed analysis) and we summarise the results in Figure 4.

In contrast with the previous model with $P = 0$, a trimorphic equilibrium (x_0, x_1, y_1) enters the interior of the state space whenever (deleted: the condition) $\frac{P}{-S} < 1$ holds. There are thus three cases to consider that depend on whether the trimorphic equilibrium enters the interior of the state space, and if it does, whether at the time of entry the

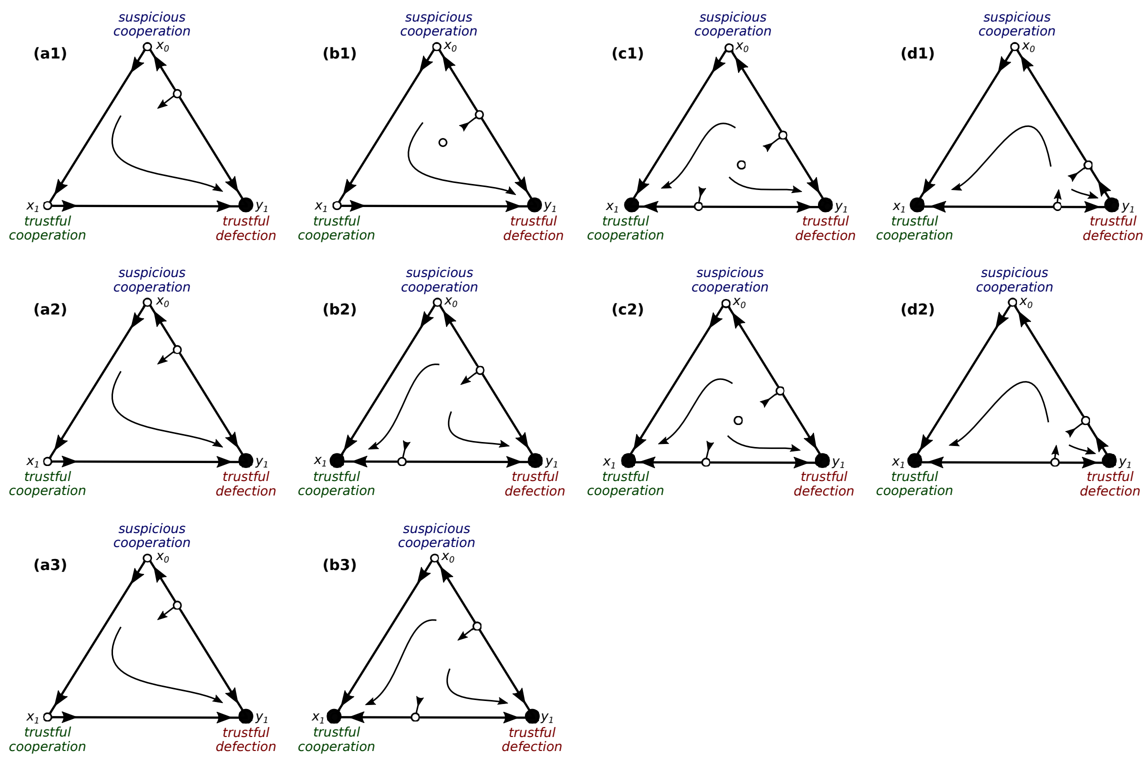

Figure 4: Evolutionary dynamics (1) for a model (4) where rejected interactions are costly $S < 0 < P$. We distinguish three cases (a1)-(d1), (a2)-(d2) and (a3)-(b3), depending on the relationship between the trimorphic equilibrium (x_0, x_1, y_1) and the equilibrium of trustful cooperation (see the main text). The parameter values are **(a1)** $0 < q < \frac{P}{-S}$ **(b1)** $\frac{P}{-S} < q < \frac{T-R}{T}$ **(c1)** $\frac{T-R}{T} < q < \frac{PR-ST}{-S(R+T)}$ **(d1)** $\frac{PR-ST}{-S(R+T)} < q < 1$, **(a2)** $0 < q < \frac{T-R}{T}$ **(b2)** $\frac{T-R}{T} < q < \frac{P}{-S}$ **(c2)** $\frac{P}{-S} < q < \frac{PR-ST}{-S(R+T)}$ **(d2)** $\frac{PR-ST}{-S(R+T)} < q < 1$, **(a3)** $0 < q < \frac{T-R}{T}$ **(b3)** $\frac{T-R}{T} < q < 1$. Notation is identical to Figure 3. There are two qualitatively different evolutionary outcomes: in panels where $0 < q < \frac{T-R}{T}$ all trajectories approach trustful defection, and in panels where $\frac{T-R}{T} < q < 1$ all trajectories approach either trustful defection or trustful cooperation depending on the initial frequency distribution. See SI for a detailed analysis.

equilibrium of trustful cooperation is stable or not. In the first case the trimorphic equilibrium (x_0, x_1, y_1) enters the interior while trustful cooperation is an unstable equilibrium $0 < \frac{P}{-S} < \frac{T-R}{T}$ (Figure 4, top row (a1)-(d1)). In (a1) $0 < q < \frac{P}{-S}$ the trimorphic equilibrium (x_0, x_1, y_1) is in the exterior of the state space and the only stable equilibrium is the equilibrium of trustful defection y_1. In (b1) $\frac{P}{-S} < q < \frac{T-R}{T}$ the unstable trimorphic equilibrium (x_0, x_1, y_1) enters the interior, and in (c1) $\frac{T-R}{T} < q < \frac{PR-ST}{-S(R+T)}$ the equilibrium of trustful cooperation x_1 becomes stable. In (d1) $\frac{PR-ST}{-S(R+T)} < q < 1$ the trimorphic equilibrium (x_0, x_1, y_1) leaves the interior. We have that in (a1)-(b1) all evolutionary trajectories approach the equilibrium of trustful defection y_1 (globally convergent ESS), and in (c1)-(d1) it depends on the initial frequency distribution of strategies whether evolutionary trajectories approach the equilibrium of trustful cooperation x_1 or trustful defection y_1 (both locally convergent ESS).

In the second case trustful cooperation is stable as the trimorphic equilibrium (x_0, x_1, y_1) enters the interior $\frac{T-R}{T} < \frac{P}{-S} < 1$ (Figure 4, middle row (a2)-(d2)). In (a2) $0 < q < \frac{T-R}{T}$ the trimorphic equilibrium (x_0, x_1, y_1) is in the exterior of the state space and the only stable equilibrium is the equilibrium of trustful defection y_1. In (b2) $\frac{T-R}{T} < q < \frac{P}{-S}$ the equilibrium of trustful cooperation becomes stable, in (c2) $\frac{P}{-S} < q < \frac{PR-ST}{-S(R+T)}$ the trimorphic equilibrium (x_0, x_1, y_1) enters the interior and in (d2) $\frac{PR-ST}{-S(R+T)} < q < 1$ the trimorphic equilibrium (x_0, x_1, y_1) leaves the interior. We have that in (a2) all trajectories approach the equilibrium of trustful defection y_1 and in (b2)-(d2) it depends on the initial frequency distribution of strategies whether trajectories approach the equilibrium of trustful cooperation x_1 or trustful defection y_1. In the third case the trimorphic equilibrium never enters the interior $1 < \frac{P}{-S}$ (Figure 4, bottom row (a3)-(b3)). In (a3) $0 < q < \frac{T-R}{T}$ the only stable equilibrium is the equilibrium of trustful defection y_1 and so all trajectories approach trustful defection and in (b3) $\frac{T-R}{T} < q < 1$ the equilibrium of trustful cooperation becomes stable and so depending on the initial frequency distribution of strategies all trajectories approach the equilibrium of trustful cooperation x_1 or trustful defection y_1. We remark that as q approaches 0 or 1 this model simplifies to the model with zero $q = 0$ (Figure 2a) and perfect information $q = 1$, respectfully.

We observe that in this model trustful defection is an ESS for all values of q. This is because opting out is costly $0 < P$ and so both trustful and suspicious cooperators are at a disadvantage for sufficiently high frequency of defectors. This means that all trajectories converge to a fully defective population whenever $0 < q < \frac{T-R}{T}$. When $\frac{T-R}{T} < q < 1$ trustful cooperation is also an ESS, but contrary to the previous model ($P = 0$) it is not a globally convergent ESS. However, the basin of attraction increases with q and for large q only trajectories close to full defection are unable to reach the ESS of trustful cooperation.

Opting out is beneficial

In this section we suppose that opting out yields a strictly greater payoff than mutual defection $P < 0 < R$. In contrast to the previous two cases, defectors ought to avoid each other and so in addition to trustful cooperators, trustful defectors and suspicious cooperators we must also consider suspicious defectors, having in total four strategies. Note that since in this model mutual defection is worse than opting out, defective strategies will reject known defectors. The expected payoffs are

$$\begin{aligned} f_0 &= \left(x_0 q^2 + x_1 q\right) R \\ f_1 &= \left(x_0 q + x_1\right) R + (y_0 q + y_1)(1-q)S \\ g_0 &= x_1(1-q)T \\ g_1 &= x_1(1-q)T + y_1(1-q)^2 P, \end{aligned} \tag{5}$$

where y_0 is the frequency and g_0 the expected payoff of suspicious defectors. The evolutionary dynamics (1) with the expected payoffs given in (5) can be analysed analytically, except for intermediate values of q where we couldn't

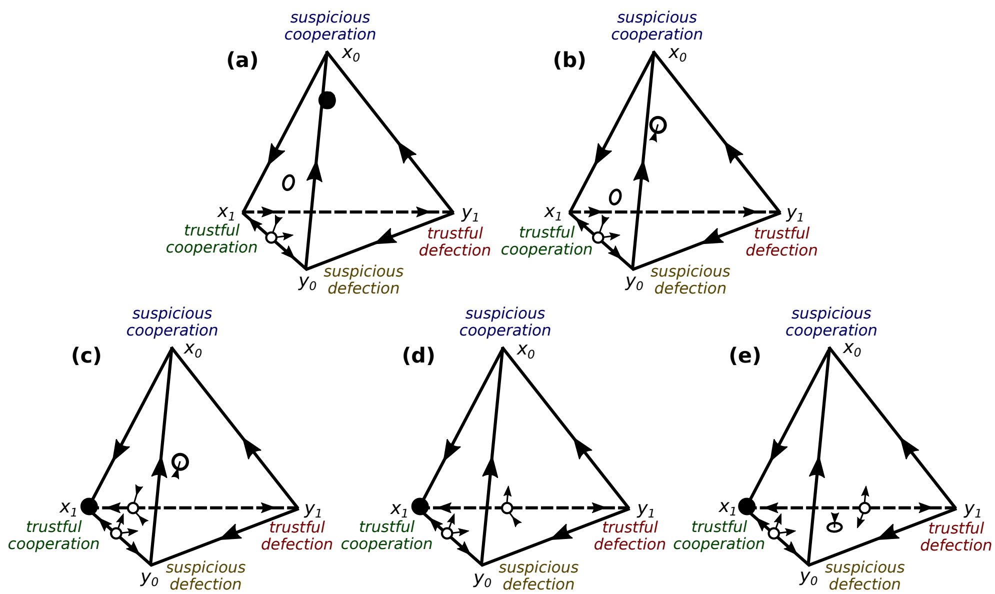

Figure 5: Evolutionary dynamics (1) for a model (5) where opting out is beneficial $P < 0 < R$, and when $T < 4R$: **(a)** $0 < q < q^{\text{stab.}}_{x_0 x_1 y_1}$ **(b)** $q^{\text{stab.}}_{x_0 x_1 y_1} < q < q_0$ **(c)** $q_0 < q < q^{\text{exit}}_{x_0 x_1 y_1}$ **(d)** $q^{\text{exit}}_{x_0 x_1 y_1} < q < q^{\text{entry}}_{x_1 y_0 y_1}$ **(e)** $q^{\text{entry}}_{x_1 y_0 y_1} < q < 1$. The filled circles are stable equilibria, i.e. all the eigenvalues are negative (see SI for details). For simplicity no arrows are drawn for the trimorphic equilibria, unless the equilibrium is an unstable equilibrium but also has negative eigenvalues in which case the stable direction(s) is drawn. There are three different evolutionary outcomes. **1**. All trajectories approach the equilibrium of suspicious cooperation, trustful cooperation and trustful defection (x_0, x_1, y_0) (panel (a)). **2**. All trajectories approach one of the two heteroclinic cycles, either $x_0 \to x_1 \to y_1$ or $x_0 \to x_1 \to y_1 \to y_0$. Numerical investigation shows it is the first one (panel (b)) . **3**. All trajectories approach the equilibrium of trustful cooperation x_1 (panels (c)-(d)).

determine which of the two, when $T < 4R$, or three, when $4R \leq T$, possible heteroclinic cycles evolutionary trajectories approach to (see below for the precise condition; a more detailed analysis is in SI). Figure 5 summarizes the results for the case $T < 4R$ and Figure 6 summarizes the case $4R \leq T$.

The threshold values at which we transition between panels in Figures 5 and 6 are

$$q^{\text{stab.}}_{x_0x_1y_1} = \frac{1}{-2P(T-R)}\left[-2P(T-R) - SR - \sqrt{R^2S^2 + 4SPRT - 4SPR^2}\right] \tag{6}$$

$$q_0 = \frac{T-R}{T} = q^{\text{enter}}_{x_1y_1} = q^{\text{exit}}_{x_0x_1y_0} \tag{7}$$

$$q^{\text{exit}}_{x_0x_1y_1} = \frac{-1}{2PR}[S(R+T) - 2PR + \sqrt{S^2(R+T)^2 - 4PSR^2}] \tag{8}$$

$$q^{\text{entry}}_{x_1y_0y_1} = \frac{1}{-2PT}[T(S-P) + \sqrt{-4P^2RT + T^2(P+S)^2}] \tag{9}$$

$$\tag{10}$$

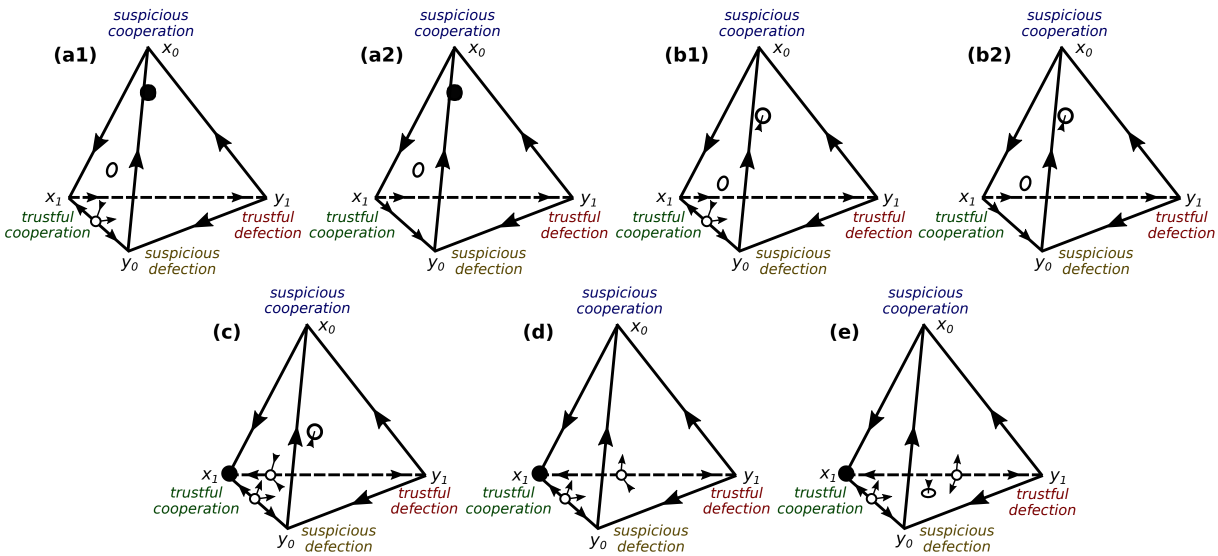

Figure 6: Evolutionary dynamics (1) for a model (5) where opting out is beneficial $P < 0 < R$, and when $4R \leq T$: In contrast with the case in Figure 5, the unstable equilibrium (x_1, y_0) exits the interior for $q^{\text{exit}}_{x_1y_0} < q < q^{\text{entry}}_{x_1y_0}$. We distinguish three cases based on the order in which we transition between the panels when q increases. For the case **(i)** $q^{\text{stab.}}_{x_0x_1y_1} < q^{\text{exit}}_{x_1y_0}$, we transition between (a1), (b1), (b2), (b1), after which continue to (c), (d) and (e) **(ii)** $q^{\text{exit}}_{x_1y_0} < q^{\text{stab.}}_{x_0x_1y_1} < q^{\text{entry}}_{x_1y_0}$ we transition between (a1), (a2), (b2), (b1), after which continue to (c), (d) and (e), and **(iii)** $q^{\text{entry}}_{x_1y_0} < q^{\text{stab.}}_{x_0x_1y_1}$ we transition between (a1), (a2), (a1), (b1), after which continue to (c), (d) and (e). Similarly to Figure 5 we have **(a1a2)** $0 < q < q^{\text{stab.}}_{x_0x_1y_1}$ **(b1b2)** $q^{\text{stab.}}_{x_0x_1y_1} < q < q_0$ **(c)** $q_0 < q < q^{\text{exit}}_{x_0x_1y_1}$ **(d)** $q^{\text{exit}}_{x_0x_1y_1} < q < q^{\text{entry}}_{x_1y_0y_1}$ **(e)** $q^{\text{entry}}_{x_1y_0y_1} < q < 1$. Notation is identical to Figure 5. There are three different evolutionary outcomes: **1**. All trajectories approach the equilibrium of suspicious cooperation, trustful cooperation and trustful defection (x_0, x_1, y_0) (panels (a1, a2)) . **2**. All trajectories approach one of the three heteroclinic cycles, either $x_0 \to x_1 \to y_1$ or $x_0 \to x_1 \to y_1 \to y_0$ (panels b1,b2), or an additional cycle $x_0 \to x_1 \to y_0$ which is possible only in panel (b2). Numerical investigation shows it is the first one. **3**. All trajectories approach the equilibrium of trustful cooperation x_1 (panels (c)-(d)) .

where $q^{\text{stab.}}_{x_0x_1y_1} < q_0 < q^{\text{exit}}_{x_0x_1y_1} < q^{\text{entry}}_{x_1y_0y_1}$ for all payoff values S, P, R, T. In Figure 6 we need additional thresholds

$$q^{\text{exit}}_{x_1y_0} = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\frac{R}{T}} \tag{11}$$

$$q^{\text{entry}}_{x_1y_0} = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\frac{R}{T}}, \tag{12}$$

where $q^{\text{exit}}_{x_1y_0} < q^{\text{entry}}_{x_1y_0} < q_0$ for all R, T. However, the relative order between the thresholds $q^{\text{exit}}_{x_1y_0}, q^{\text{entry}}_{x_1y_0}$ and $q^{\text{stab.}}_{x_0x_1y_1}$ depends on S, P, R, T.

Let us first analyze the case $T < 4R$ (Figure 5). In panel (a) $0 < q < q^{\text{stab.}}_{x_0x_1y_1}$ there exists a stable trimorphic equilibrium (x_0, x_1, y_1). In panel (b) $q^{\text{stab.}}_{x_0x_1y_1} < q < q_0$ the trimorphic equilibrium (x_0, x_1, y_1) becomes unstable and there are no stable equilibria in the system. In panel (c) $q_0 < q < q^{\text{exit}}_{x_0x_1y_1}$ the dimorphic equilibrium (x_1, y_1) enters the interior and trustful cooperation x_1 becomes stable. In panel (d) $q^{\text{exit}}_{x_0x_1y_1} < q < q^{\text{entry}}_{x_1y_0y_1}$ the trimorphic equilibrium (x_0, x_1, y_1) exits the interior by passing through the dimorphic equilibrium (x_1, y_1), and in panel (e) $q^{\text{entry}}_{x_1y_0y_1} < q < 1$ an unstable trimorphic equilibrium (x_1, y_0, y_1) enters the interior by passing through the dimorphic equilibrium (x_1, y_0). Because there are no interior $4-$morphic equilibria (see SI) all evolutionary trajectories approach the boundary of the state space. As a consequence we get that in panel (a) all evolutionary trajectories approach the stable coexistence of suspicious cooperation, trustful cooperation and trustful defection at the equilibrium (x_0, x_1, y_1). In panel (b) all evolutionary trajectories approach one of the two heteroclinic cycles, either the cycle between suspicious cooperation, trustful cooperation and trustful defection ($x_0 \to x_1 \to y_1$) or the cycle between suspicious cooperation, trustful cooperation, trustful defection and suspicious defection ($x_0 \to x_1 \to y_1 \to y_0$). Our numerical investigation indicates it is the cycle $x_0 \to x_1 \to y_1$. Finally, in panels (c)-(e) all evolutionary trajectories approach trustful cooperation x_1.

In Figure 6, where $4R \leq T$, the phase planes are similar to the previous case except that the dimorphic unstable equilibrium (x_1, y_0) exits the interior for $q^{\text{exit}}_{x_1y_0} < q < q^{\text{entry}}_{x_1y_0}$. We need to distinguish three cases based on the order in which we transition between the panels when q increases. In the first case (i) $q^{\text{stab.}}_{x_0x_1y_1} < q^{\text{exit}}_{x_1y_0}$, we transition between (a1), (b1), (b2), (b1), after which we continue to (c), (d) and (e). In the second case (ii) $q^{\text{exit}}_{x_1y_0} < q^{\text{stab.}}_{x_0x_1y_1} < q^{\text{entry}}_{x_1y_0}$ we transition between (a1), (a2), (b2), (b1), after which we continue to (c), (d) and (e), and (iii) $q^{\text{entry}}_{x_1y_0} < q^{\text{stab.}}_{x_0x_1y_1}$ we transition between (a1), (a2), (a1), (b1), after which we continue to (c), (d) and (e). Otherwise the threshold values for which we transition between panels are similar to Figure 5. An important consequence of the dimorphic equilibrium exiting the interior is that in panel (b2) evolutionary trajectories may approach an additional heteroclinic cycle of suspicious cooperation, trustful cooperation and suspicious defection ($x_0 \to x_1 \to y_0$). However, our numerical investigation indicates all trajectories approach the cycle $x_0 \to x_1 \to y_1$. We remark that as q approaches 0 or 1 this model simplifies to the model with no $q = 0$ (Figure 2(c)) and perfect information $q = 1$, respectfully. As q approaches 0 then the globally stable trimorphic equilibrium (x_0, x_1, y_1) approaches the equilibrium of suspicious cooperation x_0 and when q approaches 1 then the unstable dimorphic equilibrium (x_1, y_1) approaches y_1 and so all trajectories approach trustful cooperation.

DISCUSSION

In this paper we introduced an evolutionary game theoretic model where individuals encounter each other at random, but have the option to opt out of interactions based on partial information about their encountered opponents. With a fixed probability, individuals are assumed to know whether the opponent is a cooperator or defector. This simple formulation allowed us to solve the model of prisoners dilemma with optional interactions fully analytically, with the exception of a specific parameter region where we were not able to determine which of the three or four heteroclinic cycles evolutionary trajectories approach to (see below).

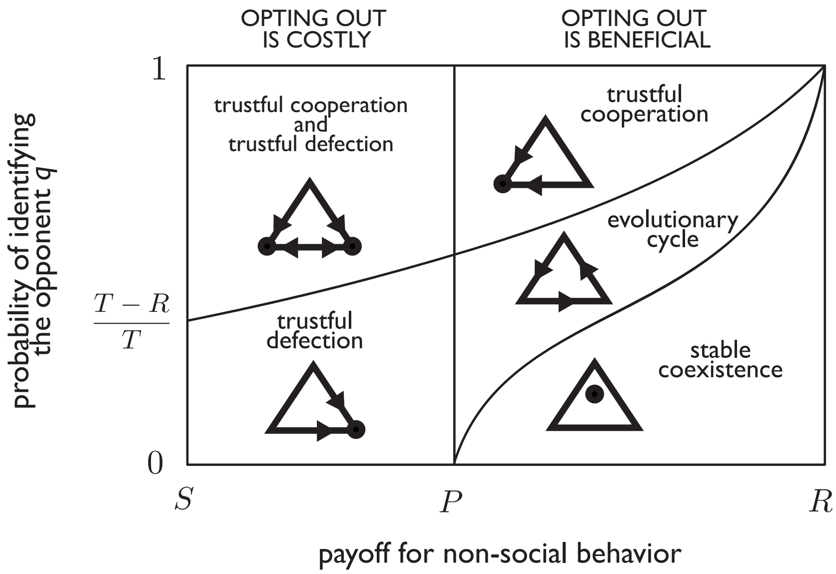

Figure 7: Summary of the results. On the vertical axis is the probability of knowing the type of the opponent q, and on the horizontal axis is the payoff for non-social behaviour 0 (opting out). The vertical black line in the middle represents the non-generic case where $P = 0$, while on the left of the vertical line opting out is costly $S \le 0 < P$ and on the right opting out is beneficial $P < 0 < R$. In each area with a different colour theme we draw a triangle that represents the phase plane for the parameter values in the area, such that in each triangle in the bottom left corner all the players are trustful cooperators x_1, in the bottom right corner all the players are trustful defectors y_1 and the upper corner all the players are suspicious cooperators x_1. Trustful cooperation is an ESS above the curve $q = \frac{T-R}{T}$ (the upper curve) and trustful defection is an ESS whenever $S \le 0 < P$. Thus for $S \le 0 < P$ and $0 \le q < \frac{T-R}{T}$ all trajectories approach trustful defection (red area), for $P \le 0S < 0$ and $\frac{T-R}{T} < q \le 1$ all trajectories approach trustful cooperation (green area) and for $S < 0 < P$ and $\frac{T-R}{T} < q \le 1$ all trajectories approach either trustful defection or trustful cooperation depending on the initial frequency distribution (purple area). For $P \le 0 \le R$ and $q^{\text{stab.}}_{x_0 x_1 y_1} < q < \frac{T-R}{T}$, where $q = q^{\text{stab.}}_{x_0 x_1 y_1}$ is the bottom curve (see the exact expression in (6)), all trajectories approach the rock-paper-scissors cycle of suspicious cooperation, trustful cooperation and trustful defection (numerical result; yellow area). For $P \le 0 \le R$ below the curve $q = q^{\text{stab.}}_{x_0 x_1 y_1}$ (blue area) all trajectories approach the stable coexistence of suspicious cooperation, trustful cooperation and trustful defection.

The results of our paper are summarised in Figure 7. First, we find that if the probability of identifying the type of the opponent is sufficiently high, $\frac{T-R}{T} < q \le 1$, then trustful cooperation is an ESS (similar condition was derived in Nowak and Sigmund 1998a,b, Suzuki and Toquenaga 2005, Ghang and Nowak 2015). Interestingly, and in contrast with previous findings, if opting out is at least as beneficial as mutual defection ($P \leq 0$), then trustful cooperation is a globally convergent ESS, i.e. trustful cooperation is reached from any initial frequency distribution of strategies (green area in Figure 7). In particular, even an (almost) entirely defective population will be replaced by trustful cooperators.

Secondly, we find that if the probability of knowing the type of the opponent is $0 \leq q \leq \frac{T-R}{T}$, and opting out is at least as beneficial as mutual defection ($P \leq 0$), then all evolutionary trajectories approach one of the three heteroclinic cycles given in model (5) (yellow area denoted ”evolutionary cycle” in Figure 7). Numerical investigation indicates that all trajectories approach the cycle of suspicious cooperation, trustful cooperation and trustful defection. Thirdly, if opting out is strictly worse than mutual defection ($S < 0 < P$) then trustful defection is always an ESS, either a locally convergent $\frac{T-R}{T} < q \leq 1$ (purple area denoted ”trustful defection and trustful cooperation” in Figure 7) or globally convergent ESS $0 \leq q \leq \frac{T-R}{T}$ (red area denoted ”trustful defection” in Figure 7). Lastly, if opting out is strictly beneficial ($P < 0 \leq R$), then for $0 \leq q < q^{\text{stab.}}_{x_0 x_1 y_1}$, trustful cooperators, trustful defectors and suspicious cooperators coexist at a globally stable equilibrium (see model (5) for the exact condition; blue area denoted ”stable coexistence” in Figure 7). Note that suspicious defectors are always (eventually) selected against and thus eradicated from the population. We remark that the models with zero $q = 0$ and perfect information $q = 1$ are (deleted: also) aligned with the Figure 7.

Our model can be extended in a straightforward manner to several intriguing directions. One possibility is to consider multiplayer games where each player has partial information about other players in the group. Here, a group of players may find themselves in a situation where only a fraction of players want to opt out while others would wish to continue the game, which may or may not be allowed depending on the biological motivation of the model. Ultimately, such situations would have to be accounted for by the model which consequently leads to more complex decision-rules as the group size increases. Another possibility is to allow errors in perception or execution of strategies (Molander 1985, Sigmund 2010). This scenario would also require to update our current strategies as even trustful individuals should either doubt the truthfulness of the observed type (errors in perception) or should be suspicious of the future action of the opponent (errors in execution). Yet another possibility is to consider a game where players don’t have the option of opting out if the opponent wants to interact. This case may apply for example in mating systems with forced copulations (Verrell 1998). However, the assumption of forced interactions may be better suited for games other than prisoners dilemma where we suspect its effect on the dynamics becomes trivial. This is because in prisoners dilemma the preference for opponents is unidirectional, and so the preferred cooperative players would be forced into harmful partnerships, consequently lowering the level of cooperation. Finally, instead of pure-decision rules a mixed decision could be used where accepting an unknown opponent happens with some probability. This set-up could be used, for example, to investigate the gradual evolution of trust in fully suspicious populations.

To conclude, our simple mathematically tractable evolutionary model with optional interactions, a model that can be readily extended to games other than prisoners dilemma, shows that the option of non-social behaviour facilitates the emergence of cooperative behaviour. Interestingly, the option of non-sociality facilitates not only stable cooperative populations but also trustful behaviour that accepts interactions with potentially harmful players.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007–2013) under REAGrant Agreement No. 291734, the Austrian Science Fund (FWF) S11407-N23 (RiSE/SHiNE) and the ERCStart Grant (279307: Graph Games).

References

Malte Andersson and Leigh W Simmons. Sexual selection and mate choice. *Trends in Ecology & Evolution*, 21(6): 296–302, 2006.

John Batali and Philip Kitcher. Evolution of altriusm in optional and compulsory games. *Journal of theoretical biology*, 175(2):161–171, 1995.

Hannelore Brandt, Christoph Hauert, and Karl Sigmund. Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences*, 103(2):495–497, 2006.

Marcos Cardinot, Maud Gibbons, Colm O'Riordan, and Josephine Griffith. Simulation of an optional strategy in the prisoner's dilemma in spatial and non-spatial environments. In *International Conference on Simulation of Adaptive Behavior*, pages 145–156. Springer, 2016.

Laureano Castro and Miguel A Toro. Iterated prisoners dilemma in an asocial world dominated by loners, not by defectors. *Theoretical population biology*, 74(1):1–5, 2008.

David DeSteno, Cynthia Breazeal, Robert H Frank, David Pizarro, Jolie Baumann, Leah Dickens, and Jin Joo Lee. Detecting the trustworthiness of novel partners in economic exchange. *Psychological science*, page 0956797612448793, 2012.

James H Fowler. Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(19):7047–7049, 2005.

Robert H Frank, Thomas Gilovich, and Dennis T Regan. The evolution of one-shot cooperation: An experiment. *Ethology and sociobiology*, 14(4):247–256, 1993.

Feng Fu, Christoph Hauert, Martin A Nowak, and Long Wang. Reputation-based partner choice promotes cooperation in social networks. *Physical Review E*, 78(2):026117, 2008.

Takako Fujiwara-Greve and Masahiro Okuno-Fujiwara. Voluntarily separable repeated prisoner's dilemma. *The Review of Economic Studies*, 76(3):993–1021, 2009. doi: 10.1111/j.1467-937X.2009.00539.x. URL + `http://dx.doi.org/10.1111/j.1467-937X.2009.00539.x`.

Whan Ghang and Martin A Nowak. Indirect reciprocity with optional interactions. *Journal of theoretical biology*, 365:1–11, 2015.

Christoph Hauert, Silvia De Monte, Josef Hofbauer, and Karl Sigmund. Replicator dynamics for optional public good games. *Journal of Theoretical Biology*, 218(2):187–194, 2002a.

Christoph Hauert, Silvia De Monte, Josef Hofbauer, and Karl Sigmund. Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570):1129–1132, 2002b.

Christoph Hauert, Arne Traulsen, Hannelore Brandt, Martin A Nowak, and Karl Sigmund. Via freedom to coercion: the emergence of costly punishment. *science*, 316(5833):1905–1907, 2007.

Daniel J Hruschka and Joseph Henrich. Friendship, cliquishness, and the emergence of cooperation. *Journal of Theoretical Biology*, 239(1):1–15, 2006.

Yoh Iwasa, Andrew Pomiankowski, and Sean Nee. The evolution of costly mate preferences ii. the'handicap'principle. *Evolution*, pages 1431–1442, 1991.

Segismundo S Izquierdo, Luis R Izquierdo, and Fernando Vega-Redondo. The option to leave: Conditional dissociation in the evolution of cooperation. *Journal of Theoretical Biology*, 267(1):76–84, 2010.

Michael D Jennions and Marion Petrie. Variation in mate choice and mating preferences: a review of causes and consequences. *Biological Reviews*, 72(2):283–327, 1997.

S. Kurokawa. The extended reciprocity: Strong belief outperforms persistence. *Journal of theoretical biology*, 421: 16–27, 2017.

Sarah Mathew and Robert Boyd. When does optional participation allow the evolution of cooperation? *Proceedings of the Royal Society of London B: Biological Sciences*, 276(1659):1167–1174, 2009.

John M McNamara, Zoltan Barta, Lutz Fromhage, and Alasdair I Houston. The coevolution of choosiness and cooperation. *Nature*, 451(7175):189–192, 2008.

Ralph R Miller. No play: a means of conflict resolution. *Journal of personality and social psychology*, 6(2):150, 1967.

Per Molander. The optimal level of generosity in a selfish, uncertain environment. *The Journal of Conflict Resolution*, 29(4):611–618, 1985. ISSN 00220027, 15528766. URL `http://www.jstor.org/stable/174244`.

Ronald Noë and Peter Hammerstein. Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral ecology and sociobiology*, 35(1):1–11, 1994.

Martin Nowak and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, pages 1291–1298, 2005.

Martin A Nowak and Karl Sigmund. The dynamics of indirect reciprocity. *Journal of theoretical Biology*, 194(4): 561–574, 1998a.

Martin A Nowak and Karl Sigmund. Evolution of indirect reciprocity by image scoring. *Nature*, 393(6685):573–577, 1998b.

John M Orbell and Robyn M Dawes. Social welfare, cooperators' advantage, and the option of not playing the game. *American sociological review*, pages 787–800, 1993.

Karthik Panchanathan and Robert Boyd. A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *Journal of theoretical biology*, 224(1):115–126, 2003.

Lawrence Ian Reed, Katharine N Zeglen, and Karen L Schmidt. Facial expressions as honest signals of cooperative intent in a one-shot anonymous prisoner's dilemma game. *Evolution and Human Behavior*, 33(3):200–209, 2012.

Thomas N Sherratt and Gilbert Roberts. The evolution of generosity and choosiness in cooperative exchanges. *Journal of Theoretical Biology*, 193(1):167–177, 1998.

Karl Sigmund. *The calculus of selfishness*. Princeton University Press, 2010.

Mathias Spichtig, Maurice W Sabelis, and Martijn Egas. Why conditional cooperators should play prisoner?s dilemma games instead of opting out. *Evolution of Altruism: Exploring Adaptive Landscapes*, page 37, 2013.

E Ann Stanley, Dan Ashlock, and Mark D Smucker. Iterated prisoner's dilemma with choice and refusal of partners: Evolutionary results. In *European Conference on Artificial Life*, pages 490–502. Springer, 1995.

Yukari Suzuki and Yukihiko Toquenaga. Effects of information and group structure on evolution of altruism: analysis of two-score model by covariance and contextual analyses. *Journal of theoretical biology*, 232(2):191–201, 2005.

Robert L Trivers. The evolution of reciprocal altruism. *The Quarterly review of biology*, 46(1):35–57, 1971.

Viktor J Vanberg and Roger D Congleton. Rationality, morality, and exit. *American Political Science Review*, 86(02): 418–431, 1992.

Paul A Verrell. The evolution of mating systems in insects and arachnids. *American Zoologist*, 38(3):585–587, 1998.

Jörgen W Weibull. Evolutionary game theory. 1995. *Massachusetts Institute of Technology*, 1995.

Matthias Wubs, Redouan Bshary, and Laurent Lehmann. Coevolution between positive reciprocity, punishment, and partner switching in repeated interactions. 283(1832):20160488, 2016.

Toshio Yamagishi, Masako Kikuchi, and Motoko Kosugi. Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, 2(1):145–161, 1999.

Amotz Zahavi. Mate selection?a selection for a handicap. *Journal of theoretical Biology*, 53(1):205–214, 1975.

Xiu-Deng Zheng, Cong Li, Jie-Ru Yu, Shi-Chang Wang, Song-Jia Fan, Bo-Yu Zhang, and Yi Tao. A simple rule of direct reciprocity leads to the stable coexistence of cooperation and defection in the prisoner's dilemma game. *Journal of Theoretical Biology*, 420:12 – 17, 2017. ISSN 0022-5193. doi: http://dx.doi.org/10.1016/j.jtbi.2017.02.036. URL `http://www.sciencedirect.com/science/article/pii/S0022519317300991`.

SUPPLEMENTARY INFORMATION

Rescaling the payoffs S, P, R, T

First we will show that only the difference in payoffs between social interactions and non-social behaviour matters. The expected payoff for each strategy A, upon encountering a random player, is

$$\begin{aligned} E_A =& \sum_B x_B \pi_{AB} u_{AB} + \sum_B x_B (1 - \pi_{AB}) L \\ =& \sum_B x_B \pi_{AB} (u_{AB} - L) + L, \end{aligned} \tag{13}$$

where x_B is the frequency of a strategy B, $\pi_{AB} \in [0, 1]$ is the probability that players A and B will play a game (function of q) and $u_{AB} \in \{S, P, R, T\}$ is the payoff to a player A when the interaction is accepted with a player B. The evolutionary dynamics is represented with the continuous-time replicator dynamics

$$\begin{aligned} \dot{x}_A =& x_A \left[E_A - \bar{E} \right] \\ =& x_A \left[\sum_B x_B \pi_{AB} (u_{AB} - L) + L - \sum_C x_C \left(\sum_B x_B \pi_{CB} (u_{CB} - L) + L \right) \right] \\ =& x_A \left[\sum_B x_B \pi_{AB} (u_{AB} - L) - \sum_{B,C} x_C x_B \pi_{CB} (u_{CB} - L) \right]. \end{aligned} \tag{14}$$

which shows that we only need to consider the difference in payoffs between accepted and rejected interactions $u_{AB} - L$. We thus scale the payoffs, and redefine the notation so that with T we denote $T - L$, etc.

Zero information

Case $0 < P$:

<u>1-morphic equilibria</u>

- $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$. The eigenvalues are
  - $\lambda_{x_0, x_1} = 0$.
  - $\lambda_{x_0, y_1} = 0$.
- $\hat{z}_{x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1, 0)$. The eigenvalues are
  - $\lambda_{x_1, x_0} = -R < 0$, and so $\hat{z}_{x_1}$ is always stable in the direction of $x_1 = 1$.

- $\lambda_{x_1,y_1} = T - R > 0$, and so $\hat{z}_{x_1}$ is always unstable in the direction of $y_1 = 1$.

- $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 0, 1)$. The eigenvalues are
  - $\lambda_{y_1,x_0} = -P < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $y_1 = 1$.
  - $\lambda_{y_1,x_1} = S - P < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $y_1 = 1$.

Since there are no (interior) 2-morphic nor (any) 3-morphic equilibria all trajectories approach the equilibrium of trustful defection $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 0, 1)$.

Case $0 = P$:

1-morphic equilibria

- $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$. For stability see below.
- $\hat{z}_{x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1, 0)$. The eigenvalues are
  - $\lambda_{x_1,x_0} = -R < 0$, and so $\hat{z}_{x_1}$ is always stable in the direction $x_1 = 1$.
  - $\lambda_{x_1,y_1} = T - R > 0$, and so $\hat{z}_{x_1}$ is always unstable in the direction $y_1 = 1$.
- $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 0, 1)$. For stability see below.

2-morphic equilibria

- $\hat{z}_{x_0 y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1 - y_1, y_1)$ gives a line of equilibria. The eigenvalues are
  - $\lambda_{x_0 y_1, x_0 y_1} = 0$, as this is a line of equilibria there is no (directional) dynamics along this line.
  - $\lambda_{x_0 y_1, y_1} = y_1 S \leq 0$, and so the line of equilibria $\hat{z}_{x_0 y_1}$ is stable w.r.t. to the interior of the phase-plane whenever $y_1 > 0$. For $y_1 = 0$ the equilibrium point $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$ is unstable in the direction of $x_1 = 1$.

Since there are no 3-morphic equilibria all trajectories approach the line of equilibria $\hat{z}_{x_0 y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1 - y_1, y_1)$, where $y_1 > 0$.

Case $P < 0$:

1-morphic equilibria

- $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$. The eigenvalues are

- $\lambda_{x_0,x_1} = 0$.
- $\lambda_{x_0,y_1} = 0$.

- $\hat{z}_{x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1, 0)$. The eigenvalues are
  - $\lambda_{x_1,x_0} = -R < 0$, and so $\hat{z}_{x_1}$ is always stable in the direction of $x_1 = 1$.
  - $\lambda_{x_1,y_1} = T - R > 0$, and so $\hat{z}_{x_1}$ is always unstable in the direction of $y_1 = 1$.
- $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 0, 1)$. The eigenvalues are
  - $\lambda_{y_1,x_0} = -P > 0$, and so $\hat{z}_{y_1}$ is always unstable in the direction of $x_0 = 1$.
  - $\lambda_{y_1,x_1} = S - P < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $y_1 = 1$.

Since there are no 3-morphic equilibria all trajectories approach the equilibrium of suspicious strategies $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$.

Perfect information

In the model with perfect information cooperators interact only amongst themselves and reject every interaction with a defector, whereas defectors interact amongst themselves if $P > 0$ and interact with no-one if $P \leq 0$.

Case $0 < P$:

The dynamics is captured by $\dot{x} = xR, \dot{y} = yP$ where x, y are cooperators and defectors, respectfully.

<u>1-morphic equilibria</u>

- $\hat{z}_x = (\hat{x}, \hat{y}) = (1, 0)$. The eigenvalue is $\lambda_{x,y} = -R$ and so this equilibrium is stable.
- $\hat{z}_y = (\hat{x}, \hat{y}) = (0, 1)$. The eigenvalue is $\lambda_{y,x} = -P$ and so this equilibrium is stable.

<u>2-morphic equilibrium</u>

- $\hat{z}_{x,y} = (\hat{x}, \hat{y}) = (\frac{P}{R+P}, \frac{R}{P+R})$. The eigenvalue is $\lambda_{xy} = \frac{RP}{R+P} > 0$ and so this equilibrium is always unstable whenever it is in the interior.

Both cooperation and defection are locally attracting strategies.

Case $P \geq 0$:

The dynamics is captured by $\dot{x} = xR, \dot{y} = 0$ where x, y are cooperators and defectors, respectfully. Since x increases for any $x > 0$ the dynamics approaches $x = 1$ for any initial condition $x > 0$.

Partial information: opting out is costly $S < 0 < P < R < T$

1-morphic equilibria

- $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$. The eigenvalues are
  - $\lambda_{x_0,x_1} = qR(1-q) > 0$, and so $\hat{z}_{x_0}$ is always unstable in the direction of $x_1 = 1$.
  - $\lambda_{x_0,y_1} = -q^2R < 0$, and so $\hat{z}_{x_0}$ is always stable in the direction of $y_1 = 1$.
- $\hat{z}_{x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1, 0)$. The eigenvalues are
  - $\lambda_{x_1,x_0} = -R(1-q) < 0$, and so $\hat{z}_{x_1}$ is always stable in the direction of $x_1 = 1$.
  - $\lambda_{x_1,y_1} = (1-q)T - R < 0$, and so $\hat{z}_{x_1}$ is stable in the direction of $y_1 = 1 \iff \frac{T-R}{T} < q < 1$. We denote $q_0 = \frac{T-R}{T}$.
- $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 0, 1)$. The eigenvalues are
  - $\lambda_{y_1,x_0} = -P < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $x_1 = 1$.
  - $\lambda_{y_1,x_1} = (1-q)S - P < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $y_1 = 1$.

2-morphic equilibria

- $(\hat{x}_0, \hat{x}_1, \hat{y}_1) = (\frac{1}{1-q}, \frac{-q}{1-q}, 0)$ which is never in the interior of the state space.
- $\hat{z}_{x_0y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (\frac{P}{P+Rq^2}, 0, \frac{Rq^2}{P+Rq^2})$, which is always in the interior of the state space. The eigenvalues are
  - $\lambda_{x_0y_1,x_0y_1} = \frac{RPq^2}{P+RPq^2} > 0$, and so $\hat{z}_{x_0y_1}$ is always unstable in the direction of $x_0 = 1$ and $y_1 = 1$.
  - $\lambda_{x_0y_1,x_0x_1y_1} = \frac{Rq}{P+Rq^2}\left(P - qP + qS - q^2S\right)$, which is always positive if $\frac{P}{-S} > 1$, and if $\frac{P}{-S} < 1$ it is positive iff $\frac{P}{-S} < q < 1$. Thus $\hat{z}_{x_0y_1}$ is unstable in the direction of the state space spanned by strategies (x_0, x_1, y_1) for all $0 < q < 1$ if $\frac{P}{-S} > 1$, and for $\frac{P}{-S} < q < 1$ if $\frac{P}{-S} < 1$. We denote $q^{\text{entry}}_{x_0x_1y_1} = \frac{P}{-S}$.
- $\hat{z}_{x_1y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, \frac{1}{A}(P - S(1-q)), \frac{1}{A}(R - T(1-q)))$, where $A = P + R - (T+S)(1-q)$. We get that the equilibrium is in the interior of the state space $\iff \frac{T-R}{T} < q < 1$. We denote $q^{\text{entry}}_{x_1y_1} = q_0$. The eigenvalues are

- $\lambda_{x_1y_1,x_1y_1} = \frac{B_1}{A}$, where $B_1 = q^2ST + PqT + qRS - 2qST + PR - PT - RS + ST$ and A is as above. Therefore, whenever the equilibrium is in the interior we must have $A > 0$, and so $\lambda_{x_1y_1,x_1y_1} > 0 \iff B_1 > 0$. We get $B_1 > 0 \iff \frac{T-R}{T} < q < 1$. That is, whenever $\hat{z}_{x_1y_1}$ is in the interior it is always unstable in the direction of $x_1 = 1$ and $y_1 = 1$.
- $\lambda_{x_1y_1,x_0x_1y_1} = \frac{B_2}{A}$, where $B_2 = q^2RS + q^2ST + PqR - qRS - 2qST - PR + ST$ and A is as above. Therefore, whenever the equilibrium is in the interior we must have $A > 0$, and so $\lambda_{x_1y_1,x_0x_1y_1} < 0 \iff B_2 < 0$. We get that $B_2 < 0$ is true for all $0 < q < 1$ if $\frac{P}{-S} > 1$, and is true for $0 < q < \frac{PR-ST}{-S(T+R)}$, if $\frac{P}{-S} < 1$. We denote $\frac{PR-ST}{-S(R+T)} = q^{\text{exit}}_{x_0x_1y_1}$.

3-morphic equilibrium

- $\hat{z}_{x_0x_1y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (\frac{1}{A}(PR - ST + qS(T + R)), \frac{1}{A}(-qR(qS + P)), \frac{1}{A}(qRT(1 - q)))$, where $A = (PR - ST + qR(S + T))(1 - q)$. We notice that $\hat{y}_1 > 0 \iff A > 0$ which is true for all $0 < q < 1$ if $\frac{T}{-S} > 1$ and is true for $0 < q < \frac{PR-ST}{-R(S+T)}$ if $\frac{T}{-S} < 1$. Also, we notice that $\hat{x}_1 > 0$, given $A > 0$, iff $\frac{P}{-S} < q$, and that $\hat{x}_0 > 0$, given $A > 0$, iff $0 < q < \frac{PR-ST}{-S(R+T)}$. Therefore $\hat{z}_{x_0x_1y_1}$ is in the interior iff $A > 0$ and $q^{\text{entry}}_{x_0x_1y_1} < q < q^{\text{exit}}_{x_0x_1y_1}$ where the latter condition is true whenever $\frac{P}{-S} < 1$. Because the former condition is true for all $0 < q < 1$ if $\frac{T}{-S} > 1$ and is true for $0 < q < \frac{PR-ST}{-R(S+T)}$ if $\frac{T}{-S} < 1$, we still need to confirm that $\frac{PR-ST}{-S(R+T)} < \frac{PR-ST}{-R(S+T)}$. Since this inequality is always true, we have that $\hat{z}_{x_0x_1y_1}$ is in the interior iff $\frac{P}{-S} < 1$ and $q^{\text{entry}}_{x_0x_1y_1} = \frac{P}{-S} < q < \frac{PR-ST}{-S(R+T)} = q^{\text{exit}}_{x_0x_1y_1}$.

  The eigenvalues are

  - $\lambda^{1,2}_{x_1y_1,x_0x_1y_1} = \frac{(1-q)qR}{2A}(B \pm \sqrt{\Delta})$, where A is as above, $B = P(T - R) - RSq > 0$ and $\Delta = q^2R^2S^2 + 4q^2RS^2T + 4q^2S^2T^2 + 2PqR^2S + 6PqRST + 4PqST^2 - 4qS^2T^2 + P^2R^2 + 2P^2RT + P^2T^2 - 4PST^2$. If the eigenvalues are complex, i.e. $\Delta < 0$, then the real part of $\lambda^{1,2}_{x_1y_1,x_0x_1y_1}$ is always positive because $B > 0$. If the eigenvalues are real, i.e. Δ is non-negative, then $\lambda^{1,2}_{x_1y_1,x_0x_1y_1}$ are positive when $q^{\text{entry}}_{x_0x_1y_1} = \frac{P}{-S} < q < \frac{PR-ST}{-S(R+T)} = q^{\text{exit}}_{x_0x_1y_1}$, i.e. whenever the equilibrium is in the interior. Therefore, the equilibrium $\hat{z}_{x_0x_1y_1}$ is unstable whenever it is in the interior of the state space.

    To see whether the equilibrium is an unstable node or a spiral we check for which values the eigenvalues are complex, i.e. $\Delta < 0$. The roots of $\Delta = 0$ are $q^{-,+}_{x_0x_1y_1,\text{complex}} = \frac{\alpha \pm 2\sqrt{\beta}}{\gamma}$, where $\alpha = PR^2 + 3RPT + 2PT^2 - 2ST^2$, $\beta = PRST^3 + 2PST^4 + S^2T^4$ and $\gamma = -S(R^2 + 4RT + 4T^2) > 0$. If $\beta < 0$, then $\Delta > 0$, i.e. the eigenvalues are always real, which is true when $\frac{T}{2T+R} < \frac{P}{-S}$. If $\beta \geq 0$, then

$\Delta < 0 \iff q^-_{x_0x_1y_1,\text{complex}} < q < q^+_{x_0x_1y_1,\text{complex}}$. Because we know that the eigenvalues must be real when the equilibrium enters and when it exits the interior of the state space, we obtain the following result: The equilibrium $\hat{z}_{x_0x_1y_1}$, is always unstable when it is in the interior of the state space (necessarily $\frac{P}{-S} < 1$), more precisely it is an

- unstable spiral iff $\frac{T}{2T+R} > \frac{P}{-S}$ and $q^-_{x_0x_1y_1,\text{complex}} < q < q^+_{x_0x_1y_1,\text{complex}}$
- unstable node iff $\frac{T}{2T+R} < \frac{P}{-S}$, or $\frac{T}{2T+R} > \frac{P}{-S}$ and $q^{\text{entry}}_{x_0x_1y_1} = \frac{P}{-S} < q < q^-_{x_0x_1y_1,\text{complex}}$ and $q^+_{x_0x_1y_1,\text{complex}} < q < \frac{PR-ST}{-S(R+T)} = q^{\text{exit}}_{x_0x_1y_1}$

We remark that if $\frac{1+\sqrt{5}}{2}R < T$ then $\frac{T}{2T+R} < \frac{T-R}{T}$, and when $R < T < \frac{1+\sqrt{5}}{2}R$ then $\frac{T}{2T+R} > \frac{T-R}{T}$. These conditions tell us the relationship between $\hat{z}_{x_1y_1}$ entering the interior and whether the eigenvalues of $\hat{z}_{x_0x_1y_1}$ are always real or not.

In summary, there are two qualitatively different evolutionary trajectories: if $0 < q < \frac{T-R}{T}$, then all trajectories tend towards trustful defection, and if $\frac{T-R}{T} < q < 1$, then trajectories tend to either trustful defection or cooperation, depending on the exact initial conditions.

Partial information: opting out yields no benefits nor costs $S < 0 = P < R < T$

Because the strategies for this model ($0 = P$) and the model where opting out is costly ($0 < P$) are identical, we obtain the evolutionary dynamics by simple setting $P = 0$ in the previous model. For completeness we work this case out.

1-morphic equilibria

- $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (1, 0, 0)$. The eigenvalues are
  - $\lambda_{x_0,x_1} = qR(1-q) > 0$, and so $\hat{z}_{x_0}$ is always unstable in the direction of $x_1 = 1$.
  - $\lambda_{x_0,y_1} = -q^2R < 0$, and so $\hat{z}_{x_0}$ is always stable in the direction of $y_1 = 1$.
- $\hat{z}_{x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 1, 0)$. The eigenvalues are
  - $\lambda_{x_1,x_0} = -R(1-q) < 0$, and so $\hat{z}_{x_1}$ is always stable in the direction of $x_0 = 1$.
  - $\lambda_{x_1,y_1} = (1-q)T - R < 0$, and so $\hat{z}_{x_1}$ is stable in the direction of $y_1 = 1 \iff \frac{T-R}{T} < q < 1$. We denote $q_0 = \frac{T-R}{T}$.
- $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, 0, 1)$. The eigenvalues are

- $\lambda_{y_1,x_0} = 0$, but since $\lambda_{x_0,y_1} < 0$ and there are no 2-morphic equilibria when $x_1 = 0$ (see below), the equilibrium $\hat{z}_{y_1}$ is always stable in the direction of $x_0 = 1$.
- $\lambda_{y_1,x_1} = (1-q)S < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $x_1 = 1$.

<u>2-morphic equilibria</u>

- $(\hat{x}_0, \hat{x}_1, \hat{y}_1) = (\frac{1}{1-q}, \frac{-q}{1-q}, 0)$ which is never in the interior of the state space.
- $\hat{z}_{x_1y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (0, \frac{1}{A}(-S(1-q)), \frac{1}{A}(R - T(1-q)))$, where $A = R - (T+S)(1-q)$. We get that the equilibrium is in the interior of the state space $\iff \frac{T-R}{T} < q < 1$. We denote $q^{\text{entry}}_{x_1y_1} = q_0$. The eigenvalues are
    - $\lambda_{x_1y_1,x_1y_1} = \frac{B_1}{A}$, where $B_1 = q^2ST + qRS - 2qST - RS + ST$ and A is as above. Therefore, whenever the equilibrium is in the interior we must have $A > 0$, and so $\lambda_{x_1y_1,x_1y_1} > 0 \iff B_1 > 0$. We get $B_1 > 0 \iff \frac{T-R}{T} < q < 1$. That is, whenever $\hat{z}_{x_1y_1}$ is in the interior it is always unstable in the direction of $x_1 = 1$ and $y_1 = 1$.
    - $\lambda_{x_1y_1,x_0x_1y_1} = \frac{B_2}{A}$, where $B_2 = q^2RS + q^2ST - qRS - 2qST + ST$ and A is as above. Therefore, whenever the equilibrium is in the interior we must have $A > 0$, and so $\lambda_{x_1y_1,x_0x_1y_1} < 0 \iff B_2 < 0$ which is true for $0 < q < \frac{T}{(T+R)}$. We denote $\frac{T}{(R+T)} = q^{\text{exit}}_{x_0x_1y_1}$.

<u>3-morphic equilibrium</u>

- $\hat{z}_{x_0x_1y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_1) = (\frac{1}{A}(-ST + qS(T+R)), \frac{1}{A}(-q^2RS), \frac{1}{A}(qRT(1-q)))$, where $A = (-ST + qR(S+T))(1-q)$. We notice that $\hat{y}_1 > 0 \iff A > 0$ which is true for all $0 < q < 1$ if $\frac{T}{-S} > 1$ and is true for $0 < q < \frac{-ST}{-R(S+T)}$ if $\frac{T}{-S} < 1$. Also, we notice that $\hat{x}_1 > 0$ only if $A > 0$, and if $A > 0$ then $\hat{x}_0 > 0$ iff $0 < q < \frac{T}{(R+T)}$. Therefore $\hat{z}_{x_0x_1y_1}$ is in the interior iff $A > 0$ and $0 < q < \frac{T}{(R+T)}$. Because the former condition is true for all $0 < q < 1$ if $\frac{T}{-S} > 1$ and is true for $0 < q < \frac{-ST}{-R(S+T)}$ if $\frac{T}{-S} < 1$, we still need to confirm that $\frac{T}{(R+T)} < \frac{-ST}{-R(S+T)}$. Since this inequality is always true, we have that $\hat{z}_{x_0x_1y_1}$ is in the interior iff $0 < q < \frac{T}{(R+T)}$.

  The eigenvalues are
    - $\lambda^{1,2}_{x_1y_1,x_0x_1y_1} = \frac{-(1-q)qSR}{2A}(B \pm \sqrt{\Delta})$, where A is as above, $B = Rq > 0$ and $\Delta = q^2R^2 + 4q^2RT + 4q^2T^2 - 4qT^2$. If the eigenvalues are complex, i.e. $\Delta < 0$, then the real part of $\lambda^{1,2}_{x_1y_1,x_0x_1y_1}$ is always

positive because $B > 0$. If the eigenvalues are real, i.e. Δ is non-negative, then $\lambda^{1,2}_{x_1y_1,x_0x_1y_1}$ are positive when $q^{\text{entry}}_{x_0x_1y_1} = 0 < q < \frac{T}{(R+T)} = q^{\text{exit}}_{x_0x_1y_1}$, i.e. whenever the equilibrium is in the interior. Therefore, the equilibrium $\hat{z}_{x_0x_1y_1}$ is unstable whenever it is in the interior of the state space.

To see whether the equilibrium is an unstable node or a spiral we check for which values the eigenvalues are complex, i.e. $\Delta < 0$. The roots of $\Delta = 0$ are $q^{-}_{x_0x_1y_1,\text{complex}} = 0$ and $q^{+}_{x_0x_1y_1,\text{complex}} = \frac{T}{R(1+\frac{R}{4T})+T}$. Because $\frac{T}{R(1+\frac{R}{4T})+T} < \frac{T}{R+T)}$ and since eigenvalues must be real when the equilibrium exits the interior we have that the eigenvalues are complex (equilibrium is a spiral) when $0 < q < \frac{T}{R(1+\frac{R}{4T})+T}$ and real (equilibrium is an unstable node) when $\frac{T}{R(1+\frac{R}{4T})+T} < q < \frac{T}{(R+T)}$.

In summary, there are two qualitatively different evolutionary trajectories: if $0 < q < \frac{T-R}{T}$, then all trajectories tend towards the heteroclinic cycle of trustful cooperation, trustful defection and suspicious cooperation, and if $\frac{T-R}{T} < q < 1$, then all trajectories tend to trustful cooperation.

Partial information: opting out is beneficial $S < P < 0 < R < T$

1-morphic equilibria

- $\hat{z}_{x_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (1, 0, 0, 0)$. The eigenvalues are
  - $\lambda_{x_0,x_1} = qR(1-q) > 0$, and so $\hat{z}_{x_0}$ is always unstable in the direction of $x_1 = 1$.
  - $\lambda_{x_0,y_0} = -q^2R < 0$, and so $\hat{z}_{x_0}$ is always stable in the direction of $y_0 = 1$.
  - $\lambda_{x_0,y_1} = -q^2R < 0$, and so $\hat{z}_{x_0}$ is always stable in the direction of $y_1 = 1$.
- $\hat{z}_{x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, 1, 0, 0)$. The eigenvalues are
  - $\lambda_{x_1,x_0} = -R(1-q) < 0$, and so $\hat{z}_{x_1}$ is always stable in the direction of $x_0 = 1$.
  - $\lambda_{x_1,y_0} = (1-q)T - R < 0$, and so for $\frac{T}{R} < 4$, $\hat{z}_{x_1}$ is always unstable, and for $\frac{T}{R} > 4$, $\hat{z}_{x_1}$ is unstable in the direction of $y_0 = 1 \iff q_- < q < q_+$ where $q_{-,+} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 4\frac{R}{T}}$.
  - $\lambda_{x_1,y_1} = (1-q)T - R < 0$, and so $\hat{z}_{x_1}$ is stable in the direction of $y_1 = 1 \iff \frac{T-R}{T} < q < 1$.
- $\hat{z}_{y_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, 0, 1, 0)$. The eigenvalues are
  - $\lambda_{y_0,x_0} = 0$, and since there is no $\hat{z}_{x_0y_0}$ interior equilibrium the stability is determined by the eigenvalue λ_{x_0,y_0}; $\hat{z}_{y_0}$ is always unstable in the direction of $x_0 = 1$.
  - $\lambda_{y_0,x_1} = q(1-q)S < 0$, and so $\hat{z}_{y_0}$ is always stable in the direction of $x_1 = 1$.

- $\lambda_{y_0,y_1} = 0$, and since there is no $\hat{z}_{y_0y_1}$ interior equilibrium the stability is determined by the eigenvalue λ_{y_1,y_0}; $\hat{z}_{y_0}$ is always stable in the direction of $y_1 = 1$.

- $\hat{z}_{y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, 0, 0, 1)$. The eigenvalues are
  - $\lambda_{y_1,x_0} = -(1-q)^2 P > 0$, and so $\hat{z}_{y_1}$ is always unstable in the direction of $x_0 = 1$.
  - $\lambda_{y_1,x_1} = (1-q)S - (1-q)^2 P < 0$, and so $\hat{z}_{y_1}$ is always stable in the direction of $x_1 = 1$.
  - $\lambda_{y_1,y_0} = -(1-q)^2 P > 0$, and so $\hat{z}_{y_1}$ is always unstable in the direction of $y_0 = 1$.

2-morphic equilibria

- $\hat{z}_{x_0x_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (\frac{1}{1-q}, \frac{-q}{1-q}, 0, 0)$ which is never in the interior of the state space.

- $\hat{z}_{x_0y_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, 0, 1, 0)$ which coincides with the 1-morphic equilibrium and is never in the strict interior of the state space.

- $\hat{z}_{x_0y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (\frac{P(1-q)^2}{Pq^2+q^2R-2Pq+P}, 0, 0, \frac{q^2R}{Pq^2+q^2R-2Pq+P})$ which is never in the interior of the state space.

- $\hat{z}_{y_0y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, 0, 1, 0)$ which coincides with the 1-morphic equilibrium and is never in the strict interior of the state space.

- $\hat{z}_{x_1y_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, \frac{-(q-1)qS}{(q^2S+q^2T-qS-qT+R}, \frac{q^2T-qT+R}{q^2S+q^2T-qS-qT+R}, 0)$ which is always in the interior when $\frac{T}{R} < 4$, and when $\frac{T}{R} > 4$ it leaves the interior $\iff q_- < q < q_+$ where $q_{-,+} = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 4\frac{R}{T}}$. We denote $q^{\text{exit}}_{x_1y_0} = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\frac{R}{T}}$ and $q^{\text{entry}}_{x_1y_0} = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\frac{R}{T}}$.
  The eigenvalues $\lambda_{x_1y_0}$: all eigenvalues and the equilibrium have the same denominator, and so the numerator of the eigenvalues determines the stability. The numerators of the eigenvalues are
  - $\tilde{\lambda}_{x_1y_0,x_1y_0} = qS(q^3T - 2q^2T + qR + qT - R)$, where the term in the brackets has roots $1, \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 4\frac{R}{T}}$: the equilibrium, whenever in the interior, is always unstable in the direction of the state space spanned by strategies x_1, y_0.
  - $\tilde{\lambda}_{x_1y_0,x_0x_1y_0} = q^2S(q^2T + qR - 2qT - R + T)$, where the term in the brackets has roots $\frac{T-R}{T}, 1$: the equilibrium, whenever in the interior, is stable in the direction of the state space spanned by strategies x_0, x_1, y_0 iff $0 < q < \frac{T-R}{T}$.

- $\tilde{\lambda}_{x_1y_0,x_0y_0y_1} = qST(q^3 - 3q^2 + 3q - 1)$, where the term in the brackets has a triple root 1: the equilibrium, whenever in the interior, is always unstable in the direction of the state space spanned by strategies x_0, y_0, y_1.

- $\hat{z}_{x_1y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, \frac{Pq^2-2Pq+qS+P-S}{Pq^2-2Pq+qS+qT+P+R-S-T}, 0, \frac{qT+R-T}{Pq^2-2Pq+qS+qT+P+R-S-T})$ which is in the interior whenever $0 < q < \frac{T-R}{T}$.

  The eigenvalues $\lambda_{x_1y_1}$: all eigenvalues and the equilibrium have the same denominator, and so the numerator of the eigenvalues determines the stability. The numerators of the eigenvalues are

  - $\tilde{\lambda}_{x_1y_1,x_1y_1} = Pq^3T + Pq^2R - 3Pq^2T + q^2ST - 2PqR + 3PqT + qRS - 2qST + PR - PT - RS + ST$, which has the roots $1 - \frac{S}{P}, \frac{T-R}{T}, 1$: the equilibrium, whenever in the interior, is always unstable in the direction of the state space spanned by strategies x_1, y_1.
  - $\tilde{\lambda}_{x_1y_1,x_1y_0y_1} = -(Pq^4T - 3Pq^3T + q^3ST + Pq^2R + 3Pq^2T - 3q^2ST - 2PqR - PqT + 3qST + PR - ST)$, which has the roots $\frac{1}{-2PT}(-PT + ST \pm \sqrt{-4P^2RT + P^2T^2 + 2PST^2 + S^2T^2}), 1$: the equilibrium, whenever in the interior, is unstable in the direction of the state space spanned by strategies x_1, y_0, y_1 iff $\frac{1}{-2PT}(-PT + ST + \sqrt{-4P^2RT + P^2T^2 + 2PST^2 + S^2T^2}) < q < 1$.
  - $\tilde{\lambda}_{x_1y_1,x_0x_1y_1} = Pq^3R - 3Pq^2R + q^2RS + q^2ST + 3PqR - qRS - 2qST - PR + ST$, which has the roots $\frac{1}{-2PR}(-2PR + S(R+T) \pm \sqrt{-4PR^2S + S^2(R+T)^2}), 1$: the equilibrium, whenever in the interior, is unstable in the direction of the state space spanned by strategies x_0, x_1, y_1 iff $q^{\text{exit}}_{x_0x_1y_1} = \frac{1}{-2PR}(-2PR + S(R+T) + \sqrt{-4PR^2S + S^2(R+T)^2}) < q < 1$.

<u>3-morphic equilibria</u>

- $\hat{z}_{x_0y_0y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (0, 0, 1, 0)$ which coincides with the 1-morphic equilibrium and is never in the strict interior of the state space.

- $\hat{z}_{x_1y_0y_1} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = \frac{1}{A}(0, SPq(1-q), (Pq^2T - PqT + qST + PR - ST), -STq(1-q))$, where $A = q^2ST - Pq^2S - Pq^2T + PqS + PqT - 2qST - PR + ST$. The equilibrium is in the interior iff

$$q^{\text{entry}}_{x_1y_0y_1} = \frac{1}{-2PT}[ST - PT + \sqrt{-4P^2RT + T^2(P+S)^2}] < q < 1. \tag{15}$$

  - The eigenvalues $\lambda_{x_1y_0y_1,x_1y_0y_1}$ corresponding to the direction spanned by strategies x_1, y_0, y_1 are of the form $\frac{-S(1-q)q}{2A}[B \pm \sqrt{\Delta}]$, where A is as above and $B = PqT + PR - TP > 0 \iff \frac{T-R}{T} < q < 1$ and

$\Delta = 4P^2q^4T^2 - 12P^2q^3T^2 + 4Pq^3ST^2 + 4P^2q^2RT + 13P^2q^2T^2 - 12Pq^2ST^2 - 6P^2qRT - 6P^2qT^2 + 12PqST^2 + P^2R^2 + 2P^2RT + P^2T^2 - 4PST^2$. Because $\frac{T-R}{T} < q^{\text{entry}}_{x_1y_0y_1}$, then when the equilibrium is in the interior at least one of the eigenvalues is always positive and thus the equilibrium is always unstable.

  – The eigenvalue corresponding to the direction spanned by strategies $x_0x_1y_0y_1$ is $\lambda_{x_1y_0y_1,x_0x_1y_0y_1} = \frac{1}{A}[-SPq^2(T(1-q)^2 + qR)] < 0$.

- $\hat{z}_{x_0x_1y_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = \frac{1}{A}(-S(T-qT-R), -SRq, RT(1-q), 0)$, where $A = (-S(T-R)+RT)(1-q) > 0$. The equilibrium is in the interior of the state space iff $0 < q < \frac{T-R}{T}$.

  – The eigenvalues $\lambda_{x_0x_1y_0,x_0x_1y_0}$ corresponding to the direction spanned by strategies x_0, x_1, y_0 are of the form $\frac{-S(1-q)q}{2A}[R \pm \sqrt{\Delta}]$, where A is as above and $\Delta = R^2 - 4qT(T-qT-R)$. Because either $\sqrt{\Delta} < R$ or the eigenvalues are complex, then both eigenvalues (or their real part) are positive and the equilibrium is always unstable (either a node or a spiral).

  – The eigenvalue corresponding to the direction spanned by strategies x_0, x_1, y_0, y_1 is $\frac{-STR(1-q)q}{-S(T-R)+RT} > 0$.

- $\hat{z}_{x_0x_1y_0} = (\hat{x}_0, \hat{x}_1, \hat{y}_0, \hat{y}_1) = (\frac{-qR}{(1-q)A}(Pq^2 - 2Pq + qS + P), \frac{1}{(1-q)A}(Pq^2R - 2PqR + qRS + qST + PR - ST), \frac{1}{A}(qRT), 0)$ where $A = Pq^2R - 2PqR + qRS + qRT + PR - ST$. We find that the equilibrium is in the interior iff

$$0 < q < \frac{-1}{2PR}[S(R+T) - 2PR + \sqrt{S^2(R+T)^2 - 4PSR^2}] = q^{\text{exit}}_{x_0x_1y_1}. \tag{16}$$

  – Next, we show that in the state space spanned by strategies x_0, x_1, y_1 the equilibrium is (i) a spiral sink for q close to 0 (ii) it changes into a spiral source at $q = q^{\text{stab.}}_{x_0x_1y_1}$, where $0 < q^{\text{stab.}}_{x_0x_1y_1} < q_0$ (iii) the equilibrium $\hat{z}_{x_1y_1}$ enters the interior at $q = q_0$, where $q^{\text{stab.}}_{x_0x_1y_1} < q_0 < q^{\text{exit}}_{x_0x_1y_1}$ (iv) the equilibrium $\hat{z}_{x_0x_1y_0}$ leaves the interior by passing the equilibrium $\hat{z}_{x_1y_1}$ at $q^{\text{exit}}_{x_0x_1y_1}$. The equilibrium $\hat{z}_{x_0x_1y_0}$ becomes an unstable node before it exists the interior either in between (ii) and (iii) or (iii) and (iv), depending on the relationship between q_0 and $q^{\text{exit}}_{x_0x_1y_1}$.

  Proof. The eigenvalues $\lambda_{x_0x_1y_1,x_0x_1y_1}$ corresponding to the direction spanned by strategies $x_0x_1y_1$ are of the form

$$\lambda = \frac{qR}{2A}\left[B \pm \sqrt{\Delta}\right] \tag{17}$$

  where $A = Pq^2R - 2PqR + qRS + qRT + PR - ST$, $B = -Pq^2R + Pq^2T + 2PqR - 2PqT - qRS - PR + PT$ and Δ is some lengthy expression. Because $\frac{qR}{2A}$ is positive whenever the equilibrium is in the interior, the equilibrium changes stability while in the interior if either (a) $\Delta > 0$ and the sign of

$\left[B \pm \sqrt{\Delta}\right]$ changes between both being negative and one of them becoming positive, or (b) $\Delta < 0$ and B changes sign (i.e. the real part of the eigenvalue).

Lets first consider (a) and solve for which q the expression $\left[B \pm \sqrt{\Delta}\right]$ is zero. We get

$$q_{1,2} = \frac{1}{2PR}\left[2PR - S(R+T) \pm \sqrt{S^2(R+T^2) - 4PSR^2}\right]$$

and $q_{3,4} = \frac{1}{2P}\left[2P - S \pm \sqrt{S^2 - 4PS}\right]$. We notice that for $q_{3,4}$ to be real numbers we must have $S < 4P$, but then q_3 and q_4 are both negative. Also, we notice that $q_{1,2}$ with a plus sign is always negative and so the only candidate for which the stability may change (given $\Delta > 0$) is $q = \frac{-1}{2PR}\left[S(R+T) - 2PR + \sqrt{S^2(R+T)^2 - 4PSR\right.$ which is the value at which the equilibrium leaves the interior. Thus if the equilibrium changes stability while in the interior, it must happen when $\Delta < 0$ and when B changes sign (because purely real eigenvalues don't change sign for $0 < q < q^{\text{exit}}_{x_0x_1y_1}$)

Before we calculate the change of sign in B, we first find that when the equilibrium leaves the interior one of the purely real eigenvalues at $q^{\text{exit}}_{x_0x_1y_1}$ is zero, and the other one is positive, by evaluating B at $q^{\text{exit}}_{x_0x_1y_1}$:

$$\frac{ST}{2PR^2}\left[-2PR^2 + ST(R+T) + T\sqrt{-S(4PR^2 - S(R+T)^2)}\right] > 0 \iff \tag{18}$$

$$-ST^2(4PR^2 - S(R+T)^2) > 4P^2R^4 - 4PSTR^2(R+T) + S^2T^2(T+R)^2 \iff \tag{19}$$

$$PR - ST > 0 \tag{20}$$

which is always true. And so at the moment of leaving the interior the eigenvalues are $\lambda_1 = \frac{qR}{2A}\left[B - \sqrt{\Delta}\right] = 0$ and $\lambda_2 = \frac{qR}{2A}\left[B + \sqrt{\Delta}\right] > 0$. The equilibrium thus changes from being an unstable node to a (unstable) saddle. Note that we don't know whether the equilibrium is an unstable saddle ($\lambda_1 < 0$ and $\lambda_2 > 0$) or node ($\lambda_1, \lambda_1 > 0$) while the equilibrium is still in the interior. However, if we find that for some $0 \leq q < q^{\text{exit}}_{x_0x_1y_1}$ the eigenvalues are complex, then necessarily the equilibrium must be a node ($\lambda_1, \lambda_2 > 0$) while the equilibrium is still in the interior. This is because the eigenvalues are continuous and at the value q where they change from complex to real the real eigenvalues must be of the same sign. If in addition we find that for some $0 \leq q < q^{\text{exit}}_{x_0x_1y_1}$ the eigenvalues are complex and the real part is negative, the stability must change whenever $B = 0$ which implies $\Delta < 0$ (since we know, again, that purely real eigenvalues don't change sign for $0 \leq q < q^{\text{exit}}_{x_0x_1y_1}$).

Lets calculate the eigenvalues (17) for small q by taking the taylor expansion at $q = 0$ up to the second order, and we get

$$\frac{Rq}{2(PR-ST)}\left[P(T-R)\pm\sqrt{P^2(R+T)^2-4PST^2)}\right]. \tag{21}$$

The expression in front of brackets is positive and the expression in front of the square root is negative. Since the discriminant is negative, we get that for small q the eigenvalues are complex and the real part $(P(T-R))$ is negative. Furthermore, the solution of $B=0$ is

$$q_{1,2}=\frac{1}{2P(R-T)}\left[2PR-2PT-RS\pm\sqrt{R^2S^2+4PRST-4PSR^2}\right] \tag{22}$$

and we check that the value with a plus sign is always greater than 1 and so B changes sign only once and this happens at the value

$$q^{\text{stab.}}_{x_0x_1y_1}=\frac{1}{2P(R-T)}\left[2PR-2PT-RS-\sqrt{R^2S^2+4PRST-4PSR^2}\right]. \tag{23}$$

We also confirm that $q^{\text{stab.}}_{x_0x_1y_1}<\frac{T-R}{T}<q^{\text{exit}}_{x_0x_1y_1}$.

The claims (i)-(iv) are thus being shown correct: (i) $0<q<q^{\text{stab.}}_{x_0x_1y_1}$ the equilibrium $\hat{z}_{x_0x_1y_0}$ has complex eigenvalues with a negative real part and is thus a stable spiral (ii) $q^{\text{stab.}}_{x_0x_1y_1}<q<\frac{T-R}{T}$ the equilibrium $\hat{z}_{x_0x_1y_0}$ becomes unstable since the real part passes zero but $\hat{z}_{x_1y_1}$ has not yet entered the interior (iii) $\frac{T-R}{T}<q<q^{\text{exit}}_{x_0x_1y_1}$ the equilibrium $\hat{z}_{x_1y_1}$ has entered the interior (iv) $q^{\text{exit}}_{x_0x_1y_1}<q<1$ the equilibrium $\hat{z}_{x_0x_1y_0}$ has left the interior and changed from being an unstable node to a saddle. The equilibrium $\hat{z}_{x_0x_1y_0}$ turned from an unstable spiral to an unstable node either at (ii) or (iii), depending on the relationship between q_0 and $q^{\text{exit}}_{x_0x_1y_1}$.

– The eigenvalue corresponding to the state space spanned by all four strategies is $\lambda_{x_0x_1y_1,x_0x_1y_0y_1}=\frac{RTq^2}{A}[S(1-q)-P(1-q)^2]<0$.

We conclude that $\hat{z}_{x_0x_1y_0}$ is a stable equilibrium for $0<q<q^{\text{stab.}}_{x_0x_1y_1}$, where $q^{\text{stab.}}_{x_0x_1y_1}<\frac{T-R}{T}<q^{\text{exit}}_{x_0x_1y_1}$.

4-morphic equilibria Lets show that this model doesn't contain an (interior) 4-morphic equilibrium. The (only) $4-$morphic equilibrium is

$$\hat{z}_{x_0x_1y_0y_1}=(\hat{x}_0,\hat{x}_1,\hat{y}_0,\hat{y}_1)=(\frac{PRSq}{(1-q)A},\frac{PS(T(1-q)-R)}{(1-q)A}, \tag{24}$$

$$\frac{-RSTq}{(1-q)A}, \frac{-RT(P(1-q)-S)}{(1-q))A}) \tag{25}$$

where $A = (PST + RST - PRS - PRT)$. Clearly, for the equilibrium to be in the interior, all the numerators must be of the same sign. However, this is never true, because $S - P(1-q) < 0$ and so the numerator of $\hat{y}_0$ is always negative while for example the numerator of $\hat{x}_1$ is always positive $PSRq > 0$.

<u>evolutionary trajectories</u>

Since the system doesn't contain an interior equilibrium then every trajectory must converge to the boundary (by theorem $5.2.1$ in Hofbauer and Sigmund (1998), and by noting that every replicator equation is equivalent to some Lotka-Volterra equation, see theorem $7.5.1$.).

We have summarised all the possible evolutionary trajectories in Figures 5 and 6. Lets collect threshold values that are important for the phase plane analysis: If $T > 4R$, then

$$q^{\text{stab.}}_{x_0 x_1 y_1} = \frac{1}{-2P(T-R)}\left[-2P(T-R) - SR - \sqrt{R^2S^2 + 4SPRT - 4SPR^2}\right] \tag{26}$$

$$q_0 = \frac{T-R}{T} = q^{\text{enter}}_{x_1 y_1} = q^{\text{exit}}_{x_0 x_1 y_0} \tag{27}$$

$$q^{\text{exit}}_{x_0 x_1 y_1} = \frac{-1}{2PR}[S(R+T) - 2PR + \sqrt{S^2(R+T)^2 - 4PSR^2}] \tag{28}$$

$$q^{\text{entry}}_{x_1 y_0 y_1} = \frac{1}{-2PT}[T(S-P) + \sqrt{-4P^2RT + T^2(P+S)^2}] \tag{29}$$

$$\tag{30}$$

where always $q^{\text{stab.}}_{x_0 x_1 y_1} < q_0 < q^{\text{exit}}_{x_0 x_1 y_1} < q^{\text{entry}}_{x_1 y_0 y_1}$. If $T > 4R$ then also thresholds

$$q^{\text{exit}}_{x_1 y_0} = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\frac{R}{T}} \tag{31}$$

$$q^{\text{entry}}_{x_1 y_0} = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 4\frac{R}{T}} \tag{32}$$

are relevant, s.t. $q^{\text{exit}}_{x_1 y_0} < q^{\text{entry}}_{x_1 y_0} < q_0$. However, it depends on the payoffs what is the relative order between $q^{\text{exit}}_{x_1 y_0}, q^{\text{entry}}_{x_1 y_0}$ and $q^{\text{stab.}}_{x_0 x_1 y_1}$.