University of Leicester

Department of Economics

Assortativity evolving from social dilemmas

Heinrich H. Nax, ETH Zürich
Alexandros Rigos, University of Leicester

Working Paper No. 15/19

August 2015

Assortativity evolving from social dilemmas

Heinrich H. Nax* Alexandros Rigos†

December 22, 2015

Abstract

Assortative mechanisms can overcome tragedies of the commons that otherwise result in dilemma situations. Assortativity criteria include various forms of kin selection, greenbeard genes, and reciprocal behaviors, usually presuming an exogenously fixed matching mechanism. Here, we endogenize the matching process with the aim of investigating how assortativity itself, jointly with cooperation, is driven by evolution. Our main finding is that full-or-null assortativities turn out to be long-run stable in most cases, independent of the relative speeds of both processes. The exact incentive structure of the underlying social dilemma matters crucially. The resulting social loss is evaluated for general classes of dilemma games, thus quantifying to what extent the tragedy of the commons may be endogenously overcome.

Keywords: cooperation, (co-)evolution, assortativity, democratic consensus

1 Introduction

What happens when a population would collectively benefit from cooperative behavior by all its individuals, while each individual has a private incentive to defect? In some such 'social dilemma' situations, collective action (Olson, 1965) may fail and the tragedy of the commons (Hardin, 1968) may result. However, many mechanisms in nature exist through which cooperative behaviors evolve (see Sachs et al., 2004; West et al., 2007, 2011, for reviews).

*Department of Social Sciences, ETH Zürich, Clausiusstrasse 37-C3, 8092 Zurich, Switzerland, email: hnax@ethz.ch.

†Department of Economics, University of Leicester, Leicester, LE1 7RH, UK, email: ar374@le.ac.uk.

Hence, the 'puzzle of cooperation' (Darwin, 1871) is that nature, involving humans and animals alike, provides us with many examples of social dilemma situations that are successfully resolved by suitable mechanisms, but also with many other examples that result in the tragedy of the commons.

Perhaps the best methodology to study the evolution of cooperation is provided by game theory (von Neumann and Morgenstern, 1944; Nash, 1951). Without suitable mechanisms, game theory predicts non-cooperative behavior in social dilemmas. The game-theoretic literature has addressed this issue at length (beginning with Hamilton, 1963, 1964a,b; Axelrod, 1984). It was shown that cooperation is not favored if interactions in the population are well-mixed/random (Nash, 1950; Lehmann and Keller, 2006; Young, 2011).

The class of mechanisms that we study in this paper function by assorting cooperators. The first assortative mechanisms date back to Wright (1921, 1922, 1965). Indeed, such mechanisms can lead to cooperative behavior in social dilemma situations; well-known examples include kin selection (Hamilton, 1964a,b; Domingue et al., 2014) via limited dispersal/locality ('spatial interactions'; Nowak and May, 1992; Eshel et al., 1998; Skyrms, 2004; Hauert, 2006; Abdellaoui et al., 2014), greenbeard genes (Dawkins, 1976; Frank, 2010; Jansen and Baalen, 2006; Sinervo et al., 2006; Brown and Buckling, 2008; Fletcher and Doebeli, 2009, 2010; Gardner and West, 2010), preferences ('homophily'; Alger and Weibull, 2012, 2013; Xie et al., 2015), or are based on behavior ('reciprocal/meritocratic matching'; Clutton-Brock, 2010; Gunnthorsdottir et al., 2010; Rabanal and Rabanal, 2014; Nax et al., 2014, 2015). Importantly, assortment based on behavior is key for (but not restricted to) sustaining cooperation in humans as both theoretical models (Biernaskie et al., 2011) and experiments (Wang et al., 2012) show. In this study we focus on this class of behavior-assortative mechanisms.

Under sufficiently assortative mechanisms, high levels of cooperation are predicted (e.g. Hamilton and Taborsky, 2005a,b; Bergstrom, 2003; Jensen and Rigos, 2014; Nax et al., 2014). It is unlikely, however, that assortativity fell from the sky. More likely, it evolved driven by evolutionary dynamics within the population and across populations. In this paper, we contribute to the assortativity literature by providing a model to endogenize the evolution of assortativity, in particular of behavior.

In our model, assortativity evolves by 'democratic consensus', a standard mechanism to reach consensus in humans. Democratic consensus therefore is a natural candidate for studying the evolution of behavior assortativity, which is particularly relevant for human interactions (Biernaskie et al., 2011). Known as range, average, cardinal, utility or score voting in voting theory, such decision-making rules are used by numerous (proto-)democratic human collectives (Staveley, 1972). Voting by clapping/shouting, financial lobbying, and other

mechanisms resembling a tug-of-war-like competition in two opposite directions are examples. The basic feature of democratic consensus in our model is that the underlying mechanism of our interaction gets more or less assortative depending on which direction yields greater surplus. Democratic consensus is also similar to biological auctions, which are aggregation rules used by many animal species (Couzin et al., 2011; Chatterjee et al., 2012) such as bees selecting hive-locations (Seeley and Visscher, 2004) or ants choosing nest sites (Franks et al., 2002). To the best of our knowledge there exists no comparable prior study of evolving assortativity based on democratic consensus dynamics. In biology, other models have been proposed based on different factors such as invasion by mutants (Dieckmann and Doebeli, 1999; Jiang et al., 2013; Dyson-Hudson and Smith, 1978; Bearhop et al., 2005). Related is also Newton (2014) who studies evolving assortativity in the indirect evolutionary models by Alger and Weibull (2012, 2013, 2014, 2015). Other ways of endogenizing the matching rule such as dynamical networks may lead to different results, and these are avenues for further research we shall sketch in our concluding discussion.

In terms of underlying games, we focus on a class of symmetric two-player social dilemmas that nests the standard prisoners' dilemma (PD) (Rapoport and Chammah, 1965) but also includes other games. All agents are of the same kind, one whose strategy choices are driven by his own material self-interest alone. All social dilemmas we consider, not just the PD, are important situations that often occur in reality with potential detrimental consequences to cooperation.

The PD is the best-known example of social dilemmas, that is, of situations with the common characteristic that individuals have an incentive to defect when facing cooperators. The evolution of cooperation amongst humans and animals in social dilemma situations has received enormous attention, and the PD in particular has been studied widely in this context beginning with Trivers (1971); Maynard Smith and Price (1973); Maynard Smith (1987) (see also (Axelrod and Hamilton, 1981)). Beyond the PD, there are related, less well-known social dilemmas of comparable practical importance. All our social dilemmas share the public goods character, but games differ with respect to which outcomes (i) are Nash equilibria and (ii) maximize total payoffs.

Our social dilemma situations include the prisoners' dilemma, the snowdrift game (also known as the hawk-dove game, the game of chicken, or the volunteer's dilemma (Maynard Smith and Price, 1973; Doebeli and Hauert, 2005; Diekmann, 1985; Myatt and Wallace, 2008; Raihani and Bshary, 2011)), the missing hero dilemma (Schelling, 1971) and the underprovision dilemma. As a byproduct of our operationalization, we introduce the 'underprovision dilemma', a variant of the snowdrift game, which to the best of our knowledge

has not previously been considered but certainly also represents an important class of games deserving investigation.

Our dynamical analyses rely on standard evolutionary replicator equations (Taylor and Jonker, 1978; Taylor, 1979). In the standard mathematical formulation of such a dynamic (e.g. Eshel, 1983; Helbing, 1992; Weibull, 1995; Eshel et al., 1997), we would assume a well-mixed population, that is, pairs would be drawn uniformly at random from the population. Here, we shall focus on action-assortative matching instead, using recently introduced methods (Bergstrom, 2003; Jensen and Rigos, 2014). In our dilemma games, such a rule is 'meritocratic' as it 'rewards' ('punishes') cooperators (defectors) by matching them with other cooperators (defectors). Assortativity itself evolves by democratic consensus. In the PD game, for example, cooperators prefer more assortativity in order to be matched less often with defectors, while defectors prefer less assortativity for the opposite reason. In which direction this struggle evolves depends on how many people stand on either side, and by how much they benefit from either change.

Our analysis proceeds in three steps. First, we study the stability of equilibria given an exogenous level of assortativity. Second, we endogenize the evolution of assortativity and investigate the stability of regimes under our voting dynamic. Finally, we evaluate which outcome is more stable in the long run.

2 The model

2.1 Social dilemmas

We start by laying out the general setup. Here, we have an infinite population taken to be the closed interval $[0,1]$ that can follow one of two strategies, either 'cooperate' (C) or 'defect' (D). (Alternative labels could be 'contribute' and 'free-ride'.) Denote by x the proportion of individuals playing C. Individuals in the population follow one of the two strategies, get matched to one other individual in the population, and then carry out their strategy in their pair. The exact process by which they get selected in pairs will be discussed in the next section.

Social dilemma A social dilemma game in our setting is represented by a matrix of the form shown in Table 1.

Table 1: The payoff matrix of a social dilemma

	C	D
C	r, r	$a, 1$
D	$1, a$	0,0

Hence a social dilemma is defined by $G = (r, a)$. To ensure that C-C is not an equilibrium under random matching, we impose $0 < r < 1$ for all G, which defines the common 'public goods character': defection is always an individual best response against cooperation, but cooperation always increases the opponent's payoff. Moreover, we restrict $a \in (-1, r)$, so that C-D outcomes are associated with either higher or lower total payoffs than C-C, while D-D remains the outcome with lowest total payoffs in all cases. We therefore investigate the following four different types of (well-known) social dilemma games:

Prisoners' dilemma (Rapoport and Chammah, 1965) The PD game is obtained by setting $2r > 1+a$ and $a < 0$. Defection is a strictly dominant strategy, and total payoffs are highest in C-C. The unique Nash equilibrium is D-D.

Snowdrift game The SD game is obtained by setting $2r < 1 + a$ and $a > 0$. Cooperation is a best response against defection, and the outcome where exactly one player contributes maximizes total payoffs. The game is a symmetric anti-coordination game with two (efficient) C-D Nash equilibria in pure strategies and one in mixed strategies. The mixed-strategy equilibrium is the unique symmetric one.

Underprovision dilemma The UD game is obtained by setting $2r > 1 + a$ and $a > 0$. It is a natural variant of the SD, but, to the best of our knowledge, has not been treated formally previously. Like in the SD, cooperation is a best response against defection, but now the synergies to mutual cooperation (beyond cost sharing) are so high that C-C maximizes total payoffs. The game is a symmetric anti-coordination game with two (inefficient) C-D Nash equilibria in pure strategies and one in mixed strategies. As in SD, the mixed-strategy equilibrium is the unique symmetric one.

Missing hero dilemma The MHD game is obtained by setting $2r < 1 + a$ and $a < 0$ (see Schelling, 1971; Diekmann and Przepiorka, 2015). Defection is a strictly dominant strategy and the unique Nash equilibrium is D-D, but – other than in the standard PD – the C-D outcome maximizes total payoffs.

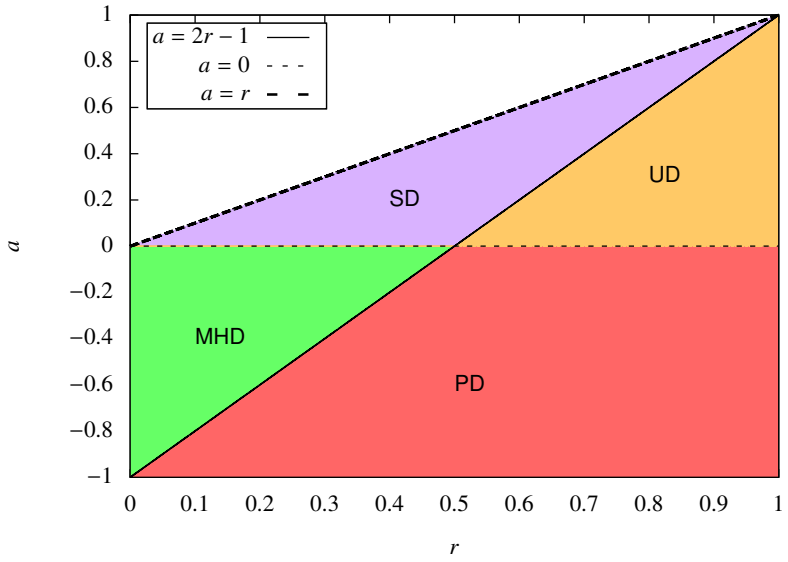

Figure 1: Types of social dilemmas.

Figure 1 illustrates how the different social dilemmas live in r–a space. The differences in the nature of the dilemmas are summarized in Table 2.

Table 2: Social dilemma classification

		Outcome with maximal total payoffs	
		C-D ($a > 2r-1$)	C-C ($a < 2r-1$)
Best Reply versus D	C ($a > 0$)	SD	UD
	D ($a < 0$)	MHD	PD

Note that only in the SD game it is the case that the pure-strategy Nash equilibria of the baseline normal-form game and the outcomes that maximize total payoffs coincide (the 'baseline loss' of the Nash equilibrium prediction is 0). In all other cases, there is a positive 'baseline loss': either C-C maximizes total payoffs but D-D (prisoners')/ C-D (underprovision) is equilibrium, or C-D maximizes total payoffs but D-D is the equilibrium (missing hero dilemma). Since we consider a one-population matching protocol, only the symmetric, mixed-strategy Nash equilibrium of the normal form game can be achieved as an equilibrium population strategy whereas the asymmetric pure-strategy equilibria (C-D) cannot.

2.2 Action assortativity

We follow Nax et al. (2014) in the definition of an action-assortative matching rule represented by a *constant index of assortativity* $\alpha \in [0,1]$ (Bergstrom, 2003). For $\alpha \in [0,1]$, α represents the difference between the probability that a cooperator has to meet another cooperator and the probability that a defector has to meet a cooperator. At one extreme ($\alpha = 1$) is full assortativity, where cooperators are matched with cooperators (and defectors with defectors) with probability one. The other extreme ($\alpha = 0$) is random matching – the standard assumption in the literature – where individuals are uniformly randomly matched with each other independently of their actions.

Environment The environment E is defined by a social dilemma, $G = (r, a)$, together with a given level of assortativity, α; $E = (G, \alpha)$.

We use the formalization introduced by Jensen and Rigos (2014) so that any matching rule can be described by a vector $\mathbf{f} = (f_1, f_2, f_3)$ where f_i represents the proportion of type-i pairs formed after the matching process. Pairs of type 1 contain two cooperators, pairs of type 2 contain one cooperator and one defector, and pairs of type 3 contain two defectors. For the matching rule to be consistent, the number of cooperators in the population must be equal to the number of cooperators that are found in the matched pairs. Therefore, the accounting identity $x = f_1 + f_2/2$ has to hold (similarly for defectors, $1-x = f_3 + f_2/2$).

Given the above, the probability that a cooperator meets another cooperator is simply $p_{CC} = f_1/x$ whereas the probability that a defector meets a cooperator is $p_{DC} = f_2/(1-x)$. Under a matching rule with a constant index of assortativity α, the difference $p_{CC} - p_{DC} = \alpha$ is constant for all values of x. Therefore, the components of $\mathbf{f}$ are given by

$$\begin{aligned} f_1(x) &= \alpha x + (1-\alpha)x^2 \\ f_2(x) &= 2(1-\alpha)x(1-x) \\ f_3(x) &= \alpha(1-x) + (1-\alpha)(1-x)^2. \end{aligned}$$

The average payoff of a cooperator is $\pi_C = (r f_1 + a f_2/2)/x$, and that of a defector is $\pi_D = (f_2/2)/(1-x)$. The average payoff in the population is therefore $x\pi_C + (1-x)\pi_D$. We refer to this average payoff as *efficiency* and discuss it in subsection 2.4.

The proportion of cooperators in the population x evolves according to the replicator dynamics where the average fitness of cooperators and defectors are π_C and π_D respectively. Thus, the dynamics of x are given by

$$\dot{x} = x(1-x)(\pi_C - \pi_D). \tag{1}$$

Definition 1 (Environment equlibrium) *Given environment $E = (G, \alpha)$, $x^* \in [0,1]$ is an environment equilibrium if it is asymptotically stable under the replicator dynamics* (1).

Lemma 1 *Almost all environments have an environment equilibrium.*

Proof. See A.1.

As shown in the proof of Lemma 1, there are some social dilemmas for which there are environments with multiple environment equilibria. These are the subset of Prisoners' Dilemmas with $a \leq r-1$. For these social dilemmas, when $\alpha \leq 1-r$ there is a unique equilibrium at $x=0$ and when $\alpha \geq \frac{a}{a-r}$, there is a unique equilibrium at $x=1$. However, for values of $\alpha \in (1-r, a/(a-r))$, both $x=0$ and $x=1$ are equlibria. In environments with $a-r+1=0$ and $\alpha = 1-r$, all $x \in [0,1]$ are neutrally stable and the replicator dynamics (1) have no asymptotically stable points. Therefore, such environments have no equilibrium. The set of these environments is "small" in the sense that it is of measure zero under any continuous probability measure over the space of environments.

2.3 Environment equilibrium robustness

We aim to evaluate how robust each of the environment equilibria is when more than one exist. In light of the previous paragraph, environments with $a < r-1$ and $\alpha \in (1-r, a/(a-r))$ have two equilibria (at $x=0$ and $x=1$). Our notion of robustness takes an 'invasion-barrier' approach (see for example Weibull, 1995, p. 42). Consider populations that evolve separately, in different demes, but under the same environment E. All such populations (unless they start from the interior rest point $x = \left(a-r+r(1-\alpha)^{-1}\right)/(a-r+1)$) will eventually be driven to either $x=0$ or $x=1$, depending on their respective initial conditions.

Now consider a population 1 which has reached the equilibrium at $x_1 = 1$ being invaded by members of another population 2 that has reached the equilibrium at $x_2 = 0$. Obviously, the resulting population will depend on the proportions with which populations 1 and 2 mix. Let λ be the proportion of individuals of population 2 in the resulting population. Then the proportion of individuals of population 1 in the resulting population will be $1-\lambda$. Obviously, the proportion of cooperators in the resulting population will be $x = \lambda \cdot x_2 + (1-\lambda) \cdot x_1 = 1-\lambda$. Our measure of robustness of the equilibrium at $x=1$ is the answer to the following question: "What is the biggest λ such that the dynamics will eventually bring the resulting population back to x_1?" So,

robustness of the equilibrium at $x=1$ is the largest shock that a population at $x=1$ can sustain. More formally, we provide Definition 2.

Definition 2 (Environment equilibrium robustness) *For any environment* $E=((r,a),\alpha)$ *with* $a<r-1$ *and* $\alpha\in\left(1-r,\frac{a}{a-r}\right)$*, the robustness* ϱ_E *of the equilibrium at* $x=1$ *is*

$$\varrho_E=\sup\left\{\lambda\in[0,1]:(1-\lambda)\in \textit{basin of attraction of } x=1\right\}.$$

We will say that the equilibrium at $x=1$ is more robust than the equilibrium at $x=0$ if $\varrho_E>1/2$. This happens for environments with $\alpha>(1-a-r)/(1-a+r)$. Notice that this invasion-barrier measure is conceptually related to but different from 'stochastic stability' (Foster and Young, 1990).

2.4 Efficiency

For a given environment $E=(G,\alpha)$ and a proportion of cooperators x, the average payoff in the population, *i.e. efficiency*, is given by

$$W(x,\alpha)=r f_1+\frac{1+a}{2}f_2+0\cdot f_3=r\,x+(1-\alpha)(1-x)x(1+a-r).$$

We now focus our attention to the highest level of efficiency that can be attained in an equilibrium of a given environment (G,α). Notice that for social dilemmas G for which there are values of α such that (G,α) has two equilibria (*i.e.* when $a<r-1$), the highest level of efficiency is reached when $x=1$. This is achieved in equilibrium for environments with $\alpha>1-r$. We refer to this as the *maximum environment equilibrium efficiency* of environment (G,α).

Lemma 2 *For all our social dilemmas* G*, maximum environment equilibrium efficiency is non-decreasing in the assortativity,* α*, of the environment.*

Proof. See A.2.

This result is important for our setting, because higher levels of assortativity will therefore typically mitigate or overcome the social dilemma that is associated with random interactions. The open question that remains is whether the endogenous dynamics that drive the evolution of assortativity will implement high levels of assortativity or not.

2.5 Full dynamics

The dynamics on assortativity α that we consider are motivated by utility voting. The tendency for α to increase/decrease is driven (i) by *the relative size* of the two populations that would benefit from an α-increase/decrease and (ii) by *the extent* of that benefit. In particular, α is governed by the following dynamics: each individual gets one vote to cast; either for higher or for lower α. In order to decide whether to vote for higher or lower assortativity, individuals use a logit choice rule (Blume, 1993) (also 'Fermi function') based on their most recently received payoff. The probability for an individual i, currently matched into a homogeneous pair (C-C or D-D), to vote for an increase of α is increasing in i's payoff. Similarly, the probability for an individual j who is currently matched into a heterogeneous pair (C-D) to vote for an increase of α is decreasing in j's payoff.

More specifically, let us denote by M and m respectively the highest and lowest payoff that can be attained by a player in a given social dilemma. If an individual gets M (m), then, with probability one, he votes for an increase (decrease) of α if in a homogeneous (heterogeneous) pair and for a decrease of α if in a heterogeneous (homogeneous) pair. If receiving a payoff of $u \in (m,M)$, an individual in a homogeneous pair votes for an increase of α with a probability given by

$$p^{+}_{\text{homog}}(u) = \frac{e^{g(u)}}{1+e^{g(u)}}, \tag{2}$$

where u is the payoff the individual received and $g(\cdot)$ is a normalizing function given by

$$g(u) = \frac{1}{M-u} - \frac{1}{u-m}. \tag{3}$$

Consequently this means that, for $u \in (m,M)$, the probability that the individual votes for a decrease of α is

$$p^{-}_{\text{homog}}(u) = \frac{1}{1+e^{g(u)}}. \tag{4}$$

Hence, the "excess" probability for an individual matched in a homogeneous pair to vote for an increase of α is

$$z_{\text{homog}}(u) = \frac{e^{g(u)}-1}{e^{g(u)}+1} \tag{5}$$

Similarly, for individuals in heterogeneous pairs, the excess probability for them to vote for an increase in α will be

$$z_{\text{heter}}(u) = \frac{1-e^{g(u)}}{e^{g(u)}+1}. \tag{6}$$

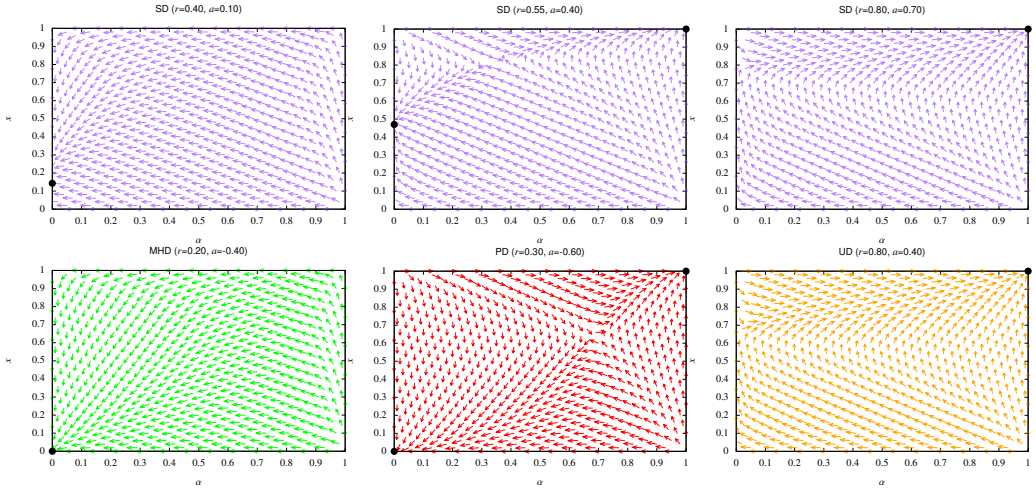

Figure 2: The principal patterns that result from the full dynamics for the different types of social dilemmas when the relative speed of the dynamic processes is $s = 1$. Black dots indicate the full equilibria. The figures are obtained by plotting the direction of the $(\dot{\alpha}, \dot{x})$ vector as given by equations (1) and (8).

Obviously, $z_{\text{heter}}(u) = -z_{\text{homog}}(u)$. In our calculations, we will use the function $z(\cdot)$ given by

$$z(u) = \begin{cases} -1 & \text{if } u = m \\ \frac{\exp(g(u))-1}{\exp(g(u))+1} & \text{if } u \in (m, M) \\ 1 & \text{if } u = M \end{cases} \tag{7}$$

Let v^+ and v^- denote the number of votes for an increase and for a decrease of the level of assortativity respectively. Aggregating the votes, these are $v^+ = f_1(x,\alpha)z(r) + f_3(x,\alpha)z(0)$ and $v^- = f_2(x,\alpha)(z(a) + z(1))/2$.

Now the exact form of the dynamics takes a replicator-style form:

$$\dot{\alpha} = s \cdot \alpha(1-\alpha)(v^+ - v^-) \tag{8}$$

where $s \in (0, \infty)$ denotes the relative speed of the α dynamics and the x dynamics. The higher the value of s, the faster α adjusts. At the limit of $s \to 0$, the whole system is governed by equation (1) whereas α remains constant and equal to its initial value.

The main patterns that arise under these dynamics for $s = 1$ are depicted in Figure 2.

2.6 Full equilibrium

We are interested in identifying states that are stable under the full dynamics. We define a *full equilibrium* as follows.

Definition 3 (Full equilibrium) *For any social dilemma $G=(r,a)$, a pair (x^*,α^*) will be called a full equilibrium if it is a stable node of the full dynamics (equations 1 and 8).*

Obviously, for (x^*,α^*) to be a full equilibrium, it is necessary for x^* to be an environment equilibrium of $E=(G,\alpha^*)$, and α^* to be an evolutionarily stable state of the voting dynamics given x^*.

Observation 1 *All full equilibria of any social dilemma G have either $\alpha^*=1$ or $\alpha^*=0$.*

Proof. For a proof and discussion see A.3.

So, for any social dilemma G there are two candidate full equilibria: one at $(x,\alpha)=(x_0^*,0)=\mathbf{x}_0$ (where $x_0^*=0$ when $a\leq 0$ and $x_0^*=a/(a-r+1)$ when $a>0$) and one at $(x,\alpha)=(1,1)=\mathbf{x}_1$.

In light of Observation 1, we want to assess and quantify how robust each of the two equilibria is. As done in section 2.3, we do so by following an invasion-barrier approach. Consider populations that evolve separately, in different demes, but under the same social dilemma $G=(r,a)$. All such populations (unless they start from the interior rest point where $\dot{x}=\dot{\alpha}=0$) will eventually be driven to either $\mathbf{x}_0$ or $\mathbf{x}_1$, depending on their respective initial conditions.

Now consider a population 1 which has reached the equilibrium at $\mathbf{x}_1$ being invaded by members of another population 2 that has reached the equilibrium at $\mathbf{x}_0$. Obviously, the resulting population will depend on the proportions with which populations 1 and 2 mix. Let λ be the proportion of individuals of population 2 in the resulting population. Then the proportion of individuals of population 1 in the resulting population will be $1-\lambda$. The proportion of cooperators in the resulting population will be $x=\lambda x_0^*+(1-\lambda)$. The level of assortativity in the resulting population (the tendency of its members to match assortatively) will be $\alpha=1-\lambda$. Our measure of robustness of the full equilibrium at $\mathbf{x}_1$ is the largest invasion that a population at $\mathbf{x}_1$ can sustain by a population at $\mathbf{x}_0$. Since full equilibria have either $\alpha=0$ or $\alpha=1$, we refer to the robustness of the full equilibrium at $\mathbf{x}_1$ as *full assortativity robustness*. More formally, we provide Definition 4.

Definition 4 (Full assortativity robustness) *For any social dilemma $G=(r,a)$ full assortativity robustness is*

$$\varrho_G=\sup\left\{\lambda\in[0,1]:\lambda\mathbf{x}_0+(1-\lambda)\mathbf{x}_1\in \textit{basin of attraction of}\,\mathbf{x}_1\right\}.$$

We will say that full assortativity is more robust than null assortativity if $\varrho_G > 1/2$. Whether full assortativity is 'more robust' than null assortativity depends on the type and exact parameter values of the underlying social dilemma G.[1] Note that assortativity robustness ϱ_G can be seen as a measure of the expected full equilibrium efficiency relative to full assortativity (the assortative optimum).[2] Figure 3 summarizes the robustness analysis in r–a space for various values of s.

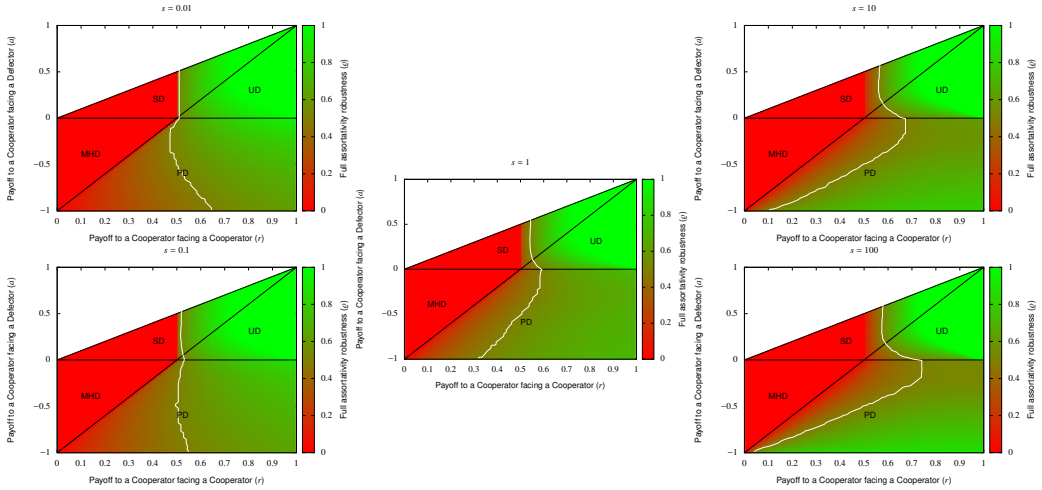

Figure 3: Robustness of full assortativity. The white line separates the cases where full/null assortativity is more robust ($\varrho = 0.5$). The figures are obtained by numerically integrating the system of differential equations (1) and (8) for each social dilemma (r,a) and finding the highest λ for which $\lambda \mathbf{x}_0 + (1-\lambda)\mathbf{x}_1$ is in the basin of attraction of $\mathbf{x}_1$ using a grid search algorithm.

We find that full assortativity is highly robust in most underprovision dilemmas and snowdrift games with $r > 0.5$. Cooperative outcomes close to socially optimal levels are therefore expected. By contrast, in missing-hero dilemmas and snowdrift games with $r < 0.5$, random matching is highly robust, and the efficient outcome is not reached. In prisoner's dilemmas, the robustness depends on the exact parameters of the game.

The intuition behind these results is the following. When $a > 0$ (SDs and UDs), individuals compare their payoff to the maximum and minimum payoffs in the game, 1 and 0 respectively. All defectors vote for a decrease in assor-

[1] Note that Bergstrom (2013) studies games where types with different assortativity levels (beyond full-or-null) compete.

[2] The expected full equilibrium efficiency is expressed by $\varrho_G r + (1-\varrho_G)(a/(1-r+a))$ when $a > 0$ and $\varrho_G r$ when $a \leq 0$ which is compared to r, the efficiency under full assortativity, yielding an expression that is linear in ϱ_G in both cases.

tativity for sure. If $r < 1/2$, then cooperators in C-C pairs are more likely to vote for a decrease in assortativity than for an increase as they do not receive a high enough payoff when matched to other cooperators. The only individuals who are more likely to vote for a higher assortativity are the cooperators who were matched to defectors. But if assortativity does increase, then the proportion of such individuals in the population decreases and thus assortativity is self-defeating. When $r > 1/2$ (UDs and some SDs), cooperators in C-C pairs are more likely to vote for more assortativity and thus if there are enough cooperators in the population, higher levels of assortativity can be reached which is now self-reinforcing. Notice that a higher value of r contribute to the increase of assortativity in two ways: it makes cooperators (i) grow faster and (ii) more likely to vote for higher assortativity. Similar arguments hold for MHDs and PDs.

Relative speeds Cooperation and assortativity co-evolve. The relative speeds of their two adjustment dynamics affect the robustness of full assortativity depending on the nature of the underlying social dilemma. In particular, faster (slower) adjustment of the assortativity level makes full assortativity more (less) robust for the class of social dilemmas for which there are environments with two equilibria (PDs with $a < r - 1$). In contrast, for social dilemmas with $a > r-1$, higher s leads to full assortativity being less robust. These results can be clearly seen in Figure 3. The intuition and analysis for this result is given in B.

3 Discussion

The 'puzzle of cooperation' in the sense of how and why cooperation amongst animals or humans emerges and survives in some social dilemmas but not in others, has kept scientists busy for many years. One strand of research in this area has been to understand the role of assortativity, through various mechanisms, in overcoming the inherent social dilemma. Indeed, some of the best-known mechanisms that lead to cooperative behavior such as kin selection and greenbeards belong to this family. In this paper, we focus on behavior assortativity and break with the assumption of a pre-existent, fixed level of assortativity. Instead, we propose a dynamic by which assortativity co-evolves with cooperation through democratic consensus. That way, we are able to study what assortativity-cooperation pairs are evolutionarily co-stable.

Our main findings summarize as follows. Only null-or-full assortativities are long-run stable, providing evolutionary support for models making either assumption depending on context as in Wright (1921, 1922, 1965). We

relate the resulting dynamics' long-run properties –in terms of assortativity and cooperativeness– to the exact incentive structure of the underlying social dilemma. Seemingly small differences between social dilemmas, even within the same class of strategic interactions, may matter crucially for convergence properties. Our analysis quantifies to what degree the tragedy of the commons is overcome through endogenous assortativity evolution. Depending on the nature of the game, higher levels of cooperation than what is achieved under random interactions may emerge. In underprovision dilemmas and snowdrift games with $r > 0.5$, cooperation reaches efficient levels. In prisoners' dilemmas, similarly high levels of assortativity and cooperation are reached for most games. In missing-hero dilemmas and in snowdrift games with $r < 0.5$, the population remains completely non-assortative and cooperation does not emerge.

The full-or-null feature of our assortativity results is driven by the absence of frictions for moving in-between (or the absence of 'costs' for upholding) assortativity regimes. However, there are also real-world systems displaying intermediate levels of assortativity. An important avenue for future work is therefore to extend our analysis to systems where assortativity is associated with costs, in which case intermediate levels of assortativity could be explained as a tradeoff between upholding cooperation and costly assortativity. Our results are likely to apply also to the context of n-player social dilemmas games such as public goods games, where assortativity/assortment are known explanations for the evolution of cooperation (Fletcher and Doebeli, 2009, 2010; Nax et al., 2015).

Another avenue for generalization of our model is to consider negative assortativity. Indeed, in situations where asymmetric outcomes lead to higher payoffs, there may be matching processes that favor the creation of mixed pairs rather than homogeneous ones and have a "dissociative" rather than an associative/assortative character. Amongst the games considered in this paper, the missing-hero dilemmas and snowdrift games are such games, where the efficient outcome is achieved by matching cooperators to defectors. Indeed, our dynamics could result in negative assortativities, and Jensen and Rigos (2014) discuss such rules with "(almost) constant indices of dissociation". In our setting, the dynamics resulting from negative assortativities can be very different from the ones under positive assortativities, because all individuals that follow one of the two behaviors are matched in mixed pairs if either $x < -\alpha/(1-\alpha)$ or $x > (1-\alpha)^{-1}$, in which case the matching outcome no longer changes as α increases. Studying negative assortativity is therefore left for future research.

Finally, different ways of endogenizing the matching process as tailored – not to humans– but to evolving kin selection and greenbeard genes in animals

should be explored. For example, when interactions are spatial, then individuals' choices to relocate would determine evolving assortativity through dynamical networks. Importantly, we note that dynamical networks may lead to the re-scaling of payoffs, as reported by Pacheco et al. (2006a,b), to instability, as reported by Cavaliere et al. (2012), as well as to the emergence of multilevel selection and strong heterogeneities, as reported by Szolnoki and Perc (2009) and Szolnoki et al. (2008). This surely has implications for local assortativity and voting procedures (which may no longer be global) and is an interesting aspect that deserves attention in future research.

Acknowledgments

We thank Chris Wallace, Martin Kaae Jensen, Peyton Young, Jörgen Weibull, Matthias Leiss, Richard Mann, Stefano Duca, Caleb Koch and Jonathan Newton, as well as seminar participants of the Norms Actions Games Conference at King's College, the COSS Game Theory Workshop at ETH Zürich and the EECS Lecture Seminar at Queen Mary University of London. We also thank one associate editor and two anonymous referees. All remaining errors are ours. Furthermore, Nax acknowledges support by the European Commission through the ERC Advanced Investigator Grant 'Momentum' (Grant 324247).

References

Abdellaoui, A., Verweij, K. J. H., Zietsch, B. P., 2014. No evidence for genetic assortative mating beyond that due to population stratification. Proceedings of the National Academy of Sciences 111 (40), E4137.

Alger, I., Weibull, J. W., 2012. A generalization of Hamilton's rule - Love others how much. Journal of Theoretical Biology 299, 42–54.

Alger, I., Weibull, J. W., 2013. Homo Moralis - Preference Evolution Under Incomplete Information and Assortative Matching. Econometrica 81, 2269–2302.

Alger, I., Weibull, J. W., 2014. Evolutionarily stable strategies, preferences and moral values, in n-player Interactions. Tech. Rep. 14-10.

Alger, I., Weibull, J. W., 2015. Evolution leads to Kantian morality. Tech. rep.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books.

Axelrod, R., Hamilton, W. D., 1981. The evolution of cooperation. Science 211 (4489), 1390–1396.

Bearhop, S., Fiedler, W., Furness, R. W., Votier, S. C., Waldron, S., Newton, J., Bowen, G. J., Berthold, P., Farnsworth, K., 2005. Assortative mating as a mechanism for rapid evolution of a migratory divide. Science 310 (5747), 502–504.

Bergstrom, T. C., 2003. The Algebra of Assortative Encounters and the Evolution of Cooperation. International Game Theory Review 5 (3), 211–228.

Bergstrom, T. C., 2013. Measures of Assortativity. Biological Theory 8 (2), 133–141.

Biernaskie, J. M., West, S. A., Gardner, A., 2011. Are greenbeards intragenomic outlaws? Evolution 65 (10), 2729–42.

Blume, L. E., 1993. The Statistical Mechanics of Strategic Interactions. Games Econ. Behav. 5, 387–424.

Brown, S. P., Buckling, A., 2008. A social life for discerning microbes. Cell 135(4), 600–603.

Cavaliere, M., Sedwards, S., Tarnita, C., Nowak, M., Csiká-Nagy, A., 2012. Prosperity is associated with instability in dynamical networks. J. Theor. Biol. 299, 126–138.

Chatterjee, K., Reiter, J. G., Nowak, M. A., 2012. Evolutionary dynamics of biological auctions. Theoretical Population Biology 81 (1), 69–80.

Clutton-Brock, T., 2010. Cooperation between non-kin in animal societies. Nature 462 (7269).

Couzin, I. D., Ioannou, C. C., Demirel, G., Gross, T., Torney, C. J., Hartnett, A., Conradt, L., Levin, S. A., Leonard, N. E., 2011. Uninformed Individuals Promote Democratic Consensus in Animal Groups. Science 334 (6062), 1578–1580.

Darwin, C., 1871. The Descent of Man and Selection in Relation to Sex. John Murray.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford.

Dieckmann, U., Doebeli, M., 1999. On the origin of species by sympatric speciation. Nature 400 (6742), 354–357.

Diekmann, A., 1985. Volunteer's Dilemma. Journal of Conflict Resolution 29, 605–610.

Diekmann, A., Przepiorka, W., 2015. Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. Scientific Reports 5, 17–52.

Doebeli, M., Hauert, C., 2005. Models of cooperation based on Prisoner's Dilemma and Snowdrift game. Ecol. Lett. 8, 748–766.

Domingue, B. W., Fletcher, J., Conley, D., Boardman, J. D., 2014. Genetic and educational assortative mating among US adults. Proceedings of the National Academy of Sciences 111 (22), 7996–8000.

Dyson-Hudson, R., Smith, E. A., 1978. Human Territoriality: An Ecological Reassessment. American Anthropologist 80 (1), pp. 21–41.

Eshel, I., 1983. Evolutionary and Continuous Stability. Journal of Theoretical Biology 103, 99–111.

Eshel, I., Motro, U., Sansone, E., 1997. Continuous Stability and Evolutionary Convergence. Journal of Theoretical Biology 185, 333–343.

Eshel, I., Samuelson, L., Shaked, A., 1998. Altruists, egoists and hooligans in a local interaction models. American Economic Review 88, 157–179.

Fletcher, J. A., Doebeli, M., 2009. A simple and general explanation for the evolution of altruism. Proc. R. Soc. B 276, 13–19.

Fletcher, J. A., Doebeli, M., 2010. Assortment is a more fundamental explanation for the evolution of altruism than inclusive fitness or multilevel selection: reply to Bijma and Aanen. Proc. R. Soc. B 277, 677–678.

Foster, D., Young, H. P., 1990. Stochastic evolutionary game dynamics. Theoretical Population Biology 38, 219–232.

Frank, S. A., 2010. Foundations of social evolution. Princeton University Press.

Franks, N. R., Pratt, S. C., Mallon, E. B., Britton, N. F., Sumpter, D. J. T., 2002. Information flow, opinion polling and collective intelligence in house–hunting social insects. Philosophical Transactions of the Royal Society of London B: Biological Sciences 357 (1427), 1567–1583.

Gardner, A., West, S. A., 2010. Greenbeards. Evolution 64 (1), 25–38.

Gunnthorsdottir, A., Vragov, R., Seifert, S., McCabe, K., 2010. Near-efficient equilibria in contribution-based competitive grouping. Journal of Public Economics 94, 987–994.

Hamilton, I. M., Taborsky, M., 2005a. Contingent movement and cooperation evolve under generalized reciprocity. Proceedings of the Royal Society B: Biological Sciences 272, 2259–2267.

Hamilton, I. M., Taborsky, M., 2005b. Unrelated helpers will not fully compensate for costs imposed on breeders when they pay to stay. Proceedings of the Royal Society B: Biological Sciences 272, 445–454.

Hamilton, W. D., 1963. The evolution of altruistic behavior. American Naturalist 97, 354–356.

Hamilton, W. D., 1964a. Genetical evolution of social behavior I. Journal of Theoretical Biology 7, 1–16.

Hamilton, W. D., 1964b. Genetical evolution of social behavior II. Journal of Theoretical Biology 7, 17–52.

Hardin, G., 1968. The Tragedy of the Commons. Science 162, 1243–1248.

Hauert, C., 2006. Spatial effects in social dilemmas. J. Theor. Biol 240, 627–636.

Helbing, D., 1992. Interrelations between stochastic equations for systems with pair interactions. Physica A 181, 29–52.

Jansen, V. A. A., Baalen, M., 2006. Altruism through beard chromodynamics. Nature 440, 663–666.

Jensen, M. K., Rigos, A., 2014. Evolutionary Games with Group Selection, working Paper No. 14/9, University of Leicester.

Jiang, Y., Bolnick, D. I., Kirkpatrick, M., 2013. Assortative Mating in Animals. The American Naturalist 181 (6), 125–138.

Lehmann, L., Keller, L., 2006. The evolution of cooperation and altruism âĂŞ a general framework and a classification of models. Journal of Evolutionary Biology 19 (5), 1365–1376.

Maynard Smith, J., 1987. The Theory of Games and the Evolution of Animal Conflicts. J. Theor. Biol. 47, 209–221.

Maynard Smith, J., Price, G. R., 1973. The logic of animal conflict. Nature 246, 15–18.

Myatt, D. P., Wallace, C., 2008. An evolutionary analysis of the volunteer's dilemma. Games and Economic Behavior 62 (1), 67–76.

Nash, J., 1950. Non-cooperative games. Ph.D. thesis, Princeton University.

Nash, J., 1951. Non-cooperative games. Ann. Math. 54, 286–295.

Nax, H. H., Balietti, S., Murphy, R. O., Helbing, D., 2015. Meritocratic Matching Can Dissolve the Efficiency-Equality Tradeoff: The Case of Voluntary Contributions Games. mimeo (SSRN 2604140).

Nax, H. H., Murphy, R. O., Helbing, D., 2014. Stability and welfare of 'merit-based' group-matching mechanisms in voluntary contribution games, Risk Center working paper, ETH Zurich.

Newton, J., 2014. Comment on homo moralis: When assortativity evolves. University of Sidney Economics WP 14.

Nowak, M. A., May, R. M., 1992. Evolutionary Games and Spatial Chaos. Nature 359, 826–829.

Olson, M., 1965. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge, MA.

Pacheco, J. M., Traulsen, A., Nowak, M. A., 2006a. Active linking in evolutionary games. J. Theor. Biol. 243, 437–443.

Pacheco, J. M., Traulsen, A., Nowak, M. A., 2006b. Coevolution of strategy and structure in complex networks with dynamical linking. Phys. Rev. Lett. 97, 258103.

Rabanal, J. P., Rabanal, O. A., 2014. Efficient Investment via Assortative Matching: A laboratory experiment. mimeo.

Raihani, N. J., Bshary, R., 2011. The evolution of punishment in n-player public goods games: a volunteer's dilemma. Evolution 65 (10), 2725–2728.

Rapoport, A., Chammah, A. M., 1965. Prisoner's Dilemma. University of Michigan Press, Michigan.

Sachs, J. L., Mueller, U. G., Wilcox, T. P., Bull, J. J., 2004. The evolution of cooperation. The Quarterly Review of Biology 79 (2), pp. 135–160.

Schelling, T., 1971. The ecology of micromotives. Public Interest 25, 61–98.

Seeley, T. D., Visscher, P. K., 2004. Quorum sensing during nest-site selection by honeybee swarms. Behavioral Ecology and Sociobiology 56 (6), 594–601.

Sinervo, B., Chaine, A., Clobert, J., Calsbeek, R., Hazard, L., Lancaster, L., McAdam, A. G., Alonzo, S., Corrigan, G., Hochberg, M. E., 2006. Self-recognition, color signals, and cycles of greenbeard mutualism and altruism. Proc. Natl. Acad. Sci. USA 103, 7372–7377.

Skyrms, B., 2004. Stag-Hunt Game and the Evolution of Social Structure. Cambridge University Press, Cambridge, U.K.

Staveley, E. S., 1972. Greek and Roman voting and elections. Thames and Hudson, London, UK.

Szolnoki, A., Perc, M., 2009. Emergence of multilevel selection in the prisoner's dilemma game on coevolving random networks. New J. Phys. 11, 093033.

Szolnoki, A., Perc, M., Danku, Z., 2008. Making new connections towards cooperation in the prisoner's dilemma game. EPL 84, 50007.

Taylor, P., 1979. Evolutionary stable strategies with two types of players. J. Appl. Prob. 16, 76–83.

Taylor, P., Jonker, L., 1978. Evolutionary stable strategies and game dynamics. Math. Biosci. 40, 145–156.

Trivers, R. L., 1971. The evolution of reciprocal altruism. Q. Rev. Biol. 46, 35–57.

von Neumann, J., Morgenstern, O., 1944. Theory of Games and Economic Behaviour. Princeton University Press.

Wang, J., Suri, S., Watts, D., 2012. Cooperation and assortativity with dynamic partner updating. Proc. Natl. Acad. Sci. USA 109, 14363–14368.

Weibull, J. W., 1995. Evolutionary Game Theory. MIT Press, Cambridge, MA.

West, S. A., Griffin, A. S., Gardner, A., 2007. Evolutionary explanations for cooperation. Current Biology 17 (16), R661 – R672.

West, S. A., Mouden, C. E., Gardner, A., 2011. Sixteen common misconceptions about the evolution of cooperation in humans. Evolution and Human Behavior 32 (4), 231–262.

Wright, S., 1921. Systems of mating. Genetics 6, 111–178.

Wright, S., 1922. Coefficients of inbreeding and relationship. American Naturalist 56, 330–338.

Wright, S., 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19, 395–420.

Xie, Y., Cheng, S., Zhou, X., 2015. Assortative mating without assortative preference. Proceedings of the National Academy of Sciences 112 (19), 5974–5978.

Young, H. P., January 2011. Commentary: John Nash and evolutionary game theory. Games and Economic Behavior 71 (1), 12–13.

Appendix

A Omitted Proofs

A.1 Proof of Lemma 1

In order for $x^* \in (0,1)$ to be an environment equilibrium, we need

$$\pi_C(x^*, \alpha) = \pi_D(x^*, \alpha). \tag{9}$$

When $a - r + 1 \neq 0$ the above condition yields

$$x^* = \frac{a - r + \frac{r}{1-\alpha}}{a - r + 1}. \tag{10}$$

Notice that the above condition is only necessary as we haven't verified that the stability condition is satisfied. Taking stability into account, we get:

$$\dot{x} > 0 \Rightarrow \pi_C(x, \alpha) > \pi_D(x, \alpha) \Rightarrow$$

$$x(a - r + 1) < a - r + \frac{r}{1-\alpha} \tag{11}$$

So $x^* \in (0,1)$ given by (10) will be an environment equilibrium for social dilemmas with $a - r + 1 \geq 0$ as in that case (11) yields $\dot{x} > 0 \Rightarrow x < x^*$.

Also, $x^* = 0$ will be an equilibrium if

$$\alpha \leq \frac{a}{a - r} \tag{12}$$

and $x^* = 1$ will be an equilibrium if

$$\alpha \geq 1 - r. \tag{13}$$

Following (Jensen and Rigos, 2014) we can separate the set of social dilemmas in three classes depending on their equilibrium behavior.

Case A: $a - r + 1 > 0$. This class contains most of the social dilemmas considered here and includes all MHDs, SDs, UDs and some PDs. All environments of social dilemmas of this class always have a unique equilibrium given by

$$x^* = \begin{cases} 0 & \text{if } \alpha \leq \frac{a}{a-r} \\ \frac{a - r + \frac{r}{1-\alpha}}{a - r + 1} & \text{if } \frac{a}{a-r} < \alpha < 1 - r \\ 1 & \text{if } \alpha \geq 1 - r \end{cases} \tag{14}$$

Case B: $a-r+1<0$. This class contains a subset of PDs. In this type of social dilemmas the only possible environment equilibria are at $x^*=0$ and $x^*=1$. More specifically, for $\alpha \leq 1-r$ we have a unique environment equilibrium at $x^*=0$. For $\alpha \in (1-r, \frac{a}{a-r})$ we have two environment equilibria: one at $x^*=0$ and one at $x^*=1$. Finally, for $\alpha \geq \frac{a}{a-r}$ we have a unique environment equilibrium at $x^*=1$.

Case C: $a-r+1=0$. This class contains a subset of PDs. In this type of social dilemmas we have a unique environment equilibrium at $x^*=0$ for $\alpha < 1-r(=-a)$, a unique environment equilibrium at $x^*=1$ for $\alpha > 1-r$. When $\alpha = 1-r$, any $x \in [0,1]$ is a stationary point (indifferent stability) and there are no environment equilibria. The set of these environments is "small" in the sense that it is of measure zero under any continuous probability measure over the space of environments. ■

A.2 Proof of Lemma 2

Consider a social dilemma $G=(a,r)$ and it's corresponding environments $E=(G,\alpha)$ with $\alpha \in [0,1]$.

If $x^* \in (0,1)$ can be an environment equilibrium only for social dilemmas with $a-r+1 \geq 0$ and it has to satisfy

$$x^* = \frac{a-r+\frac{r}{1-\alpha}}{a-r+1}$$

(see equation 10).

The uniform population without any cooperators ($x=0$) will be an environment equilibrium if

$$\alpha \leq -\frac{a}{r-a}$$

and the uniform population consisting solely of cooperators ($x=1$) will be an environment equilibrium if

$$\alpha \geq 1-r.$$

Now using the efficiency formula

$$W(x,\alpha) = r\,x + (1-\alpha)(1-x)x(1+a-r)$$

we can calculate efficiency at an interior environment equilibrium. After calculation, this gives

$$W_G^{\text{int}}(\alpha) = \frac{\alpha r + (1-\alpha)a}{1+a-r}. \tag{15}$$

It is clear that efficiency when the environment equilibrium is $x^*=1$ will be

$$W_G^1(\alpha) = r$$

and when the environment equilibrium is $x^*=0$ efficiency is

$$W_G^0(\alpha)=0.$$

Notice that $0\leq W_G^{\text{int}}(\alpha)\leq r$.

We look into efficiency for the three classes of social dilemmas mentioned in the proof of Lemma 2.

Case A: $a-r+1>0$. As described in (11), the interior point will be an environment equilibrium for some environments if $1+a-r>0$ in which case it is the unique environment equilibrium and it is clear from (15) that efficiency in this case is increasing in α. More specifically:

$$\max W^*(\alpha)=W^*(\alpha)=\begin{cases}\frac{a}{1+a-r} & \text{if } \alpha\leq\frac{a}{a-r}\\ \frac{\alpha r+(1-\alpha)a}{1+a-r} & \text{if } \frac{a}{a-r}<\alpha<1-r\\ r & \text{if } \alpha\geq 1-r\end{cases} \tag{16}$$

Case B: $a-r+1<0$. In such environments, for low values of α ($\alpha\leq -a/(r-a)$) the unique environment equilibrium is $x^*=0$ and equilibrium efficiency is 0. For high values of α ($\alpha>1-r$), the unique equilibrium is at $x^*=1$. For intermediate values of α there are two environment equilibria at $x^*=0$ and $x^*=1$. In these types of environments, maximum equilibrium efficiency is achieved for the environment equilibrium at $x^*=1$ and from the analysis here it is clear that in these social dilemmas maximum equilibrium efficiency is increasing in α.

Case C: $a-r+1=0$. In such environments, for low values of α ($\alpha\leq -a/(r-a)=1-r$) the unique environment equilibrium is $x^*=0$ and equilibrium efficiency is 0. For high values of α ($\alpha>1-r$), the unique equilibrium is at $x^*=1$. For $\alpha=1-r$ there are no environment equilibria. In these types of environments, from the analysis here it is clear that in these social dilemmas maximum equilibrium efficiency is increasing in α. More specifically:

$$\max W^*(\alpha)=\begin{cases}0 & \text{if } \alpha<1-r\\ r & \text{if } \alpha>1-r\end{cases}$$

As the set of environments with $a-r+1=0$ and $\alpha=1-r$ is of measure zero and as such an α is never encountered in a full equilibrium (see Obesrvation 1), not having a well-defined maximum equilibrium efficiency for these environments does not affect the results reported here.

A.3 Proof of Observation 1

We formally prove the statement of Observation 1 for cases where either $z(0)<0$ (SDs with $r<0.5$ and all MHDs) or $z(0)+z(1)+z(a)+z(r)>0$ (PDs with

$a-r+1<0$, and SDs and UDs with $a>1-r$). For the rest of the cases we provide computational results to support the statement.

All proofs are provided for relative speed $s=1$. Increasing or decreasing the value of s does not change the loci $\dot{x}=0$ and $\dot{\alpha}=0$ nor does it change the type of the stationary points (nodes are still nodes and saddle points are still saddle points). Therefore the result holds for any $s\in(0,\infty)$.

Clearly, in order for the pair (x^*,α^*) to be a full equilibrium (*i.e.* an asymptotically stable state of the full dynamics), we need x^* to be an environment equilibrium for $E=(G,\alpha^*)$, and α^* to be an evolutionarily stable state of the voting dynamics given x^*. We begin with the following observation.

Observation 2 *Consider a social dilemma G. If (x^*,α^*) with $\alpha^*\in(0,1)$ is a full equilibrium, then $x^*\in(0,1)$.*

Proof.

By way of contradiction, say $(1,\alpha^*)$ with $\alpha^*\in(0,1)$ is a full equilibrium. Then the α dynamic (8) for $x=1$ yields

$$\dot{\alpha}=\alpha(1-\alpha)z(r).$$

So, for all points $(1,\alpha)$ with $\alpha\in(0,1)$, $\dot{\alpha}$ retains its sign. That is, if $z(r)>0$ the only possible full equilibrium with $x^*=1$ will be the one at $(x^*,\alpha^*)=(1,1)$ and if $z(r)<0$ the only possible full equilibrium with $x^*=1$ will be the one at $(x^*,\alpha^*)=(1,0)$. Finally, if $z(r)=0$, there exists no full equilibrium with $x^*=1$ as $\dot{\alpha}=0$ for all (x,α) with $x=1$. This is a contradiction.

Similarly for $x^*=0$. Say $(0,\alpha^*)$ is a full equilibrium. Then the α dynamic for $x=0$ yields

$$\dot{\alpha}=\alpha(1-\alpha)z(0).$$

As $z(0)<0$ for all social dilemmas, $\dot{\alpha}<0$ for all $\alpha\in(0,1)$ and thus, the only possible full equilibrium with $x^*=0$ is $(x^*,\alpha^*)=(0,0)$. This is a contradiction.

In light of Observation 2, any potential full equilibrium (x^*,α^*) that contradicts the statement of Observation 1 needs to have $x^*\in(0,1)$ and $\alpha^*\in(0,1)$.

For the rest of the proof, we define the quantity $Z=z(1)+z(a)+z(r)+z(0)$ which turns out to be crucial for the behavior of the dynamic system.

Case 1: $Z>0$ For $\alpha\in(0,1)$, the α dynamics equation (8) gives:

$$\dot{\alpha}>0\Rightarrow\alpha x(1-x)Z>x(1-x)(z(a)+z(1))+x^2z(r)+(1-x)^2z(0)$$

So, for social dilemmas with $Z>0$ any interior steady state (x^*,α^*) (with $\dot{\alpha}=\dot{x}=0$) will not be stable as a small deviation to $(x^*,\alpha^*+\varepsilon)$ will drive the system away from (x^*,α^*).

Case 2: $z(r) < 0$ So, we now focus on social dilemmas with $Z < 0$. In this case, the sign of $z(r)$ is important.

We proceed with another observation.

Observation 3 *For any social dilemma $G = (r, a)$ with $Z < 0$ and $z(r) < 0$, and for all $\alpha \in (0,1)$ and $x \in (0,1)$, $\dot{\alpha} < 0$.*

Proof. Define $V = v^+ - v^-$. For $(x, \alpha) \in (0,1)^2$, we have the following:

$$\dot{\alpha} > 0 \Leftrightarrow V > 0$$

$$\frac{\partial V}{\partial \alpha} = x(1-x)Z < 0$$

So, V would obtain its maximum value for $\alpha = 0$. This value is

$$V_0 = (x^2 z(r) + (1-x)^2 z(0)) + (-x(1-x)(z(a) + z(1)))$$

As z is increasing, the maximum value the first term of the right-hand side of the above equation is $z(r)$, which is negative in the social dilemmas under consideration. Also, notice that as $z(1) = 1$ for all social dilemmas and $z(a) > -1$, the second term in the equation will also be negative.

So, any full equilibrium would need to have $\alpha^* = 0$.

Case 3: $Z < 0$ **and** $z(r) > 0$ For the rest of the cases, we note that there is always at most one interior rest point of the full dynamics.

We numerically calculate the Jacobian matrix of the dynamical system and take the real parts of its two eigenvalues at the the interior rest point. Figure 4 provides the plot of the maximum real part of the aforementioned eigenvalues. Since there is always at least one eigenvalue with a positive real part, we can conclude that the interior rest point cannot be stable and thus cannot be an environment equilibrium.

Furthermore, as the interior rest point (if it exists) is unique and a saddle point, there can be no closed trajectories in the interior of the state space. From this, we conclude that there must exist stable points (sinks) at the boundary of the state space which are the full equilibria.

B Analysis for different relative speeds

Consider a social dilemma G with $a < r - 1$ and take the limit where $s \to 0$. In that case, the α-adjustment is so slow that, for initial conditions (x_0, α_0), the

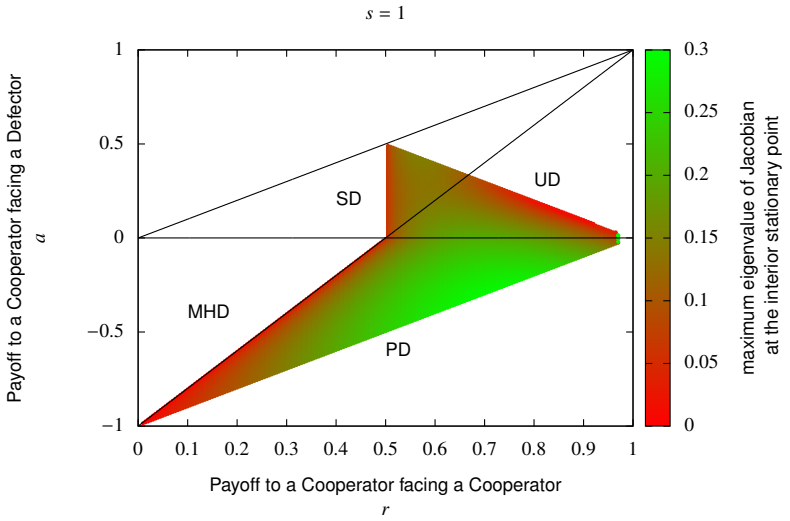

Figure 4: Maximum eigenvalue of the Jacobian matrix calculated at the interior rest point for social dilemmas with $z(r) > 0$

level of cooperation fully adjusts to one of the two environment equilibria (at either $x = 0$ or at $x = 1$, see the discussion in subsection 2.2) before assortativity gets a chance to evolve. If the environment equilibrium reached is $x = 0$, then the α-dynamics will slowly bring the population to the full equilibrium $\mathbf{x}_0$. On the other hand, if the environment equilibrium reached is $x = 1$, the population is led to the full equilibrium at $\mathbf{x}_1$. So, assortativity robustness will be the highest λ for which $x = 1-\lambda$ is in the basin of attraction of the equilibrium at $x = 1$ under the environment $(G, 1-\lambda)$. At this limit, the social dilemmas where full assortativity is more robust than null assortativity ($\varrho_G > 1/2$) are the ones with $r > (1-\alpha)/3$. Our numerical results reflect this (see the diagram for $s = 0.01$ in Figure 3).

Increasing speed s leads some of the initial conditions (those below the $\dot{x} = 0$ curve and above the $\dot{\alpha} = 0$ curve) to higher levels of α (see top row of Figure 5). This, in turn, decreases the threshold for x to start being increasing in time. These initial conditions would be driven to $x = 0$ by the dynamics if s was low and, therefore, eventually led to the full equilibrium $\mathbf{x}_0$.

The opposite is true for social dilemmas with $a > r-1$. Let (x^*, α^*) be the interior rest point of the full dynamics. When $s \to 0$, initial conditions with $\alpha_0 < \alpha^*$ are first attracted to the $\dot{x} = 0$ locus and then, slowly, to the $\mathbf{x}_0$ equilibrium. As s increases, a larger set of initial conditions along the $\lambda\mathbf{x}_0 + (1-\lambda)\mathbf{x}_1$ line are being attracted to $\mathbf{x}_0$ and therefore full assortativity robustness decreases (see bottom row of Figure 5).

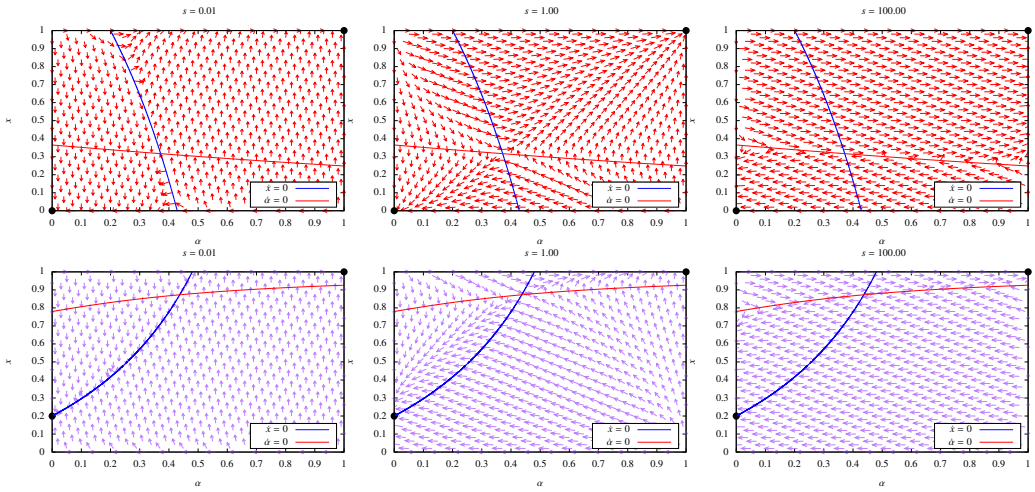

Figure 5: Top: Full dynamics for a Prisoners' Dilemma with $r = 0.80$, $a = -0.60$ for different values of speed s. Higher speeds of the α-dynamics lead to a larger basin of attraction of the full equilibrium $\mathbf{x}_1$ and to higher robustness of full assortativity.
Bottom: Full dynamics for a Snowdrift game with $r = 0.52$, $a = 0.12$ for different values of speed s. Higher speeds of the α-dynamics lead to a higher basin of attraction of the full equilibrium $\mathbf{x}_0$ and to loweer robustness of full assortativity.