Evolutionarily stable altruism*

Ingela Alger† Jörgen W. Weibull‡

September 13, 2010

Abstract

We define a concept of evolutionary stability for degrees of altruism and provide an operational criterion for such stability. The concept is first defined for two-player simultaneous move games, with assortative and uniform random matching as special cases. We apply this criterion to a variety of public-goods provision games. The concept is then generalized to n-player simultaneous-move games, to repeated two-player games, and to social preferences other than pure altruism.

Keywords: Evolutionary stability, altruism, preference evolution, assortative matching.

JEL codes: D64

1 Introduction

It is by now well-known that altruistic behavior, whereby individuals forego material benefits for the benefit of others, is common in non-market interactions. In this paper, we ask how

*Preliminary and incomplete. Please do not quote. Both authors thank the Knut and Alice Wallenberg Research Foundation for financial support. Ingela Alger thanks Carleton University for financial support, as well as the Stockholm School of Economics for its hospitality.

†Department of Economics, Carleton University.

‡Department of Economics, Stockholm School of Economics; Department of Economics, Ecole Polytechnique, Paris; and Department of Mathematics, Royal Institute of Technology, Stockholm.

altruistic we should expect people to be, from first principles. In pursuing this endeavour, the efforts made in this direction by biologists are notable. Indeed, ever since Charles Darwin's pioneering work (Darwin, 1859, 1871), altruistic behavior has been perceived to be potentially problematic for his theory of evolution, and evolutionary biologists have sought to explain such behavior. One of the most successful ideas, formalized by Hamilton (1964a,b), is that of kinship altruism. Hamilton's rule states that a fitness-decreasing action will be undertaken if and only if $c < rb$, where b is the *fitness benefit* to the recipient and c the *fitness cost* to the donor, and r is is a measure of "genetic closeness" between kin called Wright's (1922) *coefficient of relatedness*. For a pair of siblings, this coefficient is $r = 1/2$, between first cousins, it is $r = 1/8$ etc. Hamilton argues that evolution will select behaviors that maximize the weighted sum of own individual fitness, given weight one, and that of one's sibling, given the weight equals the coefficient of relatedness, r. It is as if individuals have *altruistic preferences* towards their kin, with an altruistic inclination of strength r. This result, which we will use as a benchmark, is known to hold in evolutionary models where behaviors, rather than preferences, are inherited (Grafen, 1979, Hines and Maynard Smith, 1979, Bergstrom, 1995, Day and Taylor, 1998, Grafen, 2006, and Rowthorn, 2006).

It turns out that the r in Hamilton's rule may be given a much broader interpretation than the degree of relatedness among kin, and that this interpretation significantly widens the rule's applicability to economics. Indeed, in a population consisting of similar individuals (incumbents), and where a vanishingly small share of the population differs from the rest (mutants), r is the probability that a mutant be matched with another mutant: it is an index of assortativity. With this broad interpretation in mind, we ask whether a result similar to Hamilton's rule would hold if evolution operated on preferences rather than on behavior.

We propose a model of the evolution of altruistic preferences in interactions within small groups of individuals who know each other well. We define a concept of local evolutionary stability for degrees of altruism and provide an operational criterion for such stability. The evolutionary process operates at a population level, where group formation is stochastic, with assortative and uniform random matching as special cases. We define a degree of altruism to be *locally evolutionarily stable* if, in a population consisting of individuals with that disposition, a small population share of mutants with a slightly higher or lower degree of altruism would fare less well, in terms of their (material) payoff. This definition is first applied to groups of two individuals involved in a one-shot interaction. The concept is then generalized to groups of arbitrary finite size, and to infinitely repeated two-player games.

We take the matching process as given, and use an index of assortativity to measure the probability that a mutant would be matched with another mutant. In games played between relatives, this index equals Wright's coefficient of relatedness. Our model thus gives precise predictions regarding Hamilton's rule; this rule holds if and only if the stable degree of altruism equals the index of assortativity. In the two-person, one-shot game setting, we derive two principles.

The first principle maps the specifics of the interaction at hand (for any given index of assortativity) to deviations from Hamilton's rule. This rule holds if the interaction is such that each player's best reply is independent of the other's action. By contrast, when strategic externalities are present, individuals' strategic adaptation of their behaviors will lead them away from Hamilton's rule. In particular, if the actions are strategic complements, the stable degree of altruism will exceed the index of assortativity while in games with strategic substitutes the stable degree of altruism will fall below it. As a special case we obtain that if the index of assortativity is zero (uniform random matching) — as it will be in anonymous trading in large markets — then spitefulness is evolutionarily stable in the case of strategic substitutes while a positive degree of altruism is stable in the case of strategic complements.

The second principle maps the index of assortativity (for given specifics of the interaction at hand) to locally stable degrees of altruism. The principle states that this degree is increasing in the index of assortativity. The higher is this index (for instance, the closer relatives the individuals are), the higher is the likelihood that a more (less) altruistic mutant will be paired with another more (less) altruistic mutant. As a result, more (less) altruistic mutants will do better (worse) than the incumbents.

We also show that these principles carry over to n-player one-shot games. Interestingly, increasing the size of the group does not affect the stable degree of altruism dramatically, and a larger group size does not necessarily lead to a lower degree of altruism. In our example of a game with strategic substitutes, the stable degree of altruism is increasing in the group size, while the opposite is true in our example of a game with strategic complements.

Although our evolutionary methodology is developed first for one-shot two-player games, and then for one-shot n-player games, games between individuals who may be more or less altruistic or spiteful, we show that the methodology can be applied to infinitely repeated two-player games, as well as to games between individuals with other types of social preferences, such as inequity aversion.

The greatest challenge is to treat infinitely repeated games, since these may have a

plethora of (subgame perfect) equilibria. Our evolutionary approach requires that one can attach a value function to the interaction in question. In the case of an infinitely repeated game, we need a value function that, for a given stage game and (common) discount factor, maps pairs of degrees of altruism to expected material utility outcomes. In order to achieve such a value function, we adopt two alternative methods. The first method is to let the value function be defined by a fixed convex combination of two subgame perfect equilibrium outcomes: (a) the repeated play of the stage-game's unique Nash equilibrium and (b) play of a "best possible" subgame-perfect equilibrium (given the discount factor and the players' altruistic or spiteful preferences), defined as the Nash bargaining solution over the set of subgame-perfect equilibrium outcomes, with the (unique) stage-game Nash equilibrium as the default point. By way of numerical simulations, we show that the stable degree of altruism is decreasing in the (common) discount factor, from the stable degree of altruism for the one-shot interaction towards zero or even negative values. It is as if altruism becomes a negative asset when players coordinate on paths of play in this way; a spiteful or selfish individual has a stronger bargaining position than an altruist. The second method we use is to let the value function be defined in terms of the Nash bargaining solution over the set of individually rational and feasible stage-game payoff pairs (defined in terms of the individual's degrees of altruism), with endogenous threat points, as defined in Nash (1953). For the case of selfish players, Abreu and Pearce (2007) provide microfoundations for this equilibrium selection.[1]

In order to demonstrate how our evolutionary approach can be applied to other social preferences than altruism/spite, we analyze two games. First, we study a sharing game between two individuals who may be more or less inequity averse in the sense of Fehr and Schmidt (1999). We derive the evolutionary stable degree of inequity aversion, and show that it is higher the higher is the index of assortativity. We also show that the resulting equilibrium transfers, at the evolutionarily stable degree of inequity aversion, are identical with the those for individuals who instead of being inequity averse are altruistic at the evolutionarily stable degree of altruism. In other words, at least in this simple application, the resulting evolutionarily stable behavior is independent of how we formalize social preferences. Second, we analyze a game between individuals who exhibit conditional altruism in the sense of Levine (1998) and Sethi and Somanathan (2001).

The present analysis builds on Alger and Weibull (2010), where we define and analyze

[1]See also Abreu and Pearce (2002) for another approach leading to the same selection.

stable degrees of altruism in a setting where two individuals have the option to pool risks by way of voluntary transfers *ex post*. It uses the indirect evolutionary approach, pioneered by Güth and Yaari (1992) and Güth (1995), to endogenously determine preferences. Some contributions in this literature have aimed at providing general conditions under which payoff-maximizing preferences exhibit some form of evolutionary stability, and much of the focus has been on the role of the observability of players' preferences (Ok and Vega-Redondo, 2001, Dekel, Ely and Yilankaya, 2007, and Heifetz, Shannon and Spiegel, 2007). Other authors have applied the indirect evolutionary approach to analyze the evolution of a specific kind of non-payoff-maximizing preferences, such as overconfidence (Kyle and Wang, 1997), altruism (Bester and Güth, 1998), social status (Fershtman and Weiss, 1998), a concern for relative payoffs (Koçkesen, Ok and Sethi, 2000), and reciprocal preferences (Sethi and Somanathan, 2001).[2] The papers that are closest to ours are by Heifetz, Shannon and Spiegel (2006, 2007). They consider symmetric two-player games with uniform random matching. To the best of our knowledge, all contributions to this literature share the assumption that matching is purely random.

By contrast, the (usually positive) effects of assortative matching on the evolution of altruistic behaviors have been studied extensively in biology, see, e.g., Haldane (1955), Williams and Williams (1957), Wilson (1977), Robson (1990), Bergstrom (1995, 2003), Nakamaru and Levin (2004), Durrett and Levin (2005), Grafen (2006), Nowak (2006), Fletcher and Doebeli (2009), and Tarnita et al. (2009). However, these studies consider strategy evolution, not preference evolution, as under the indirect evolutionary approach and as we do here. Thus, our model provides a bridge between the biology literature focussed on the role of assortative matching and strategy evolution of altruistic behaviors, and the economics literature on preference evolution under uniform random matching.

The remainder is organized as follows. The basic setting and our definition of local evolutionary stability in two-player interactions is given in Section 2. Section 3 provides general results for this setting, and applies the methodology to some canonical interactions. Section 4 analyzes symmetric n-player interactions. We extend the domain of analysis to the case of infinitely repeated interactions in Section 5, and to studies of evolutionarily stable degrees of inequity aversion of the form suggested in Fehr and Schmidt (1999), and of reciprocal altruism in Section 6. Section 7 discusses alternative matching processes, biological and

[2]Recent analyses of preferences in non-strategic environments include Robson and Samuelson (2009), and Samuelson and Swinkels (2006).

socioeconomic, and points at potential further applications, and Section 8 concludes. Mathematical proofs are given in Appendix A, while Appendix B provides background calculations for infinitely repeated interactions.

2 The model

2.1 The pairwise interaction

We here consider a class of symmetric two-player games in which both players simultaneously choose a non-negative real number, x and y, respectively and where the *material payoff* from taking *action* $x \geq 0$ against *action* $y \geq 0$ is denoted $\pi(x, y)$. The payoff function $\pi : \mathbb{R}^2_+ \to \mathbb{R}$ is continuous, and it is twice continuously differentiable on the interior of its domain. We write Π for this class of (material) payoff functions.

The two players may be more or less altruistic or spiteful towards each other. Hence, the *utility* from taking action x against action y, for a player with degree α of altruism towards the other, is

$$u(x, y, \alpha) = \pi(x, y) + \alpha \cdot \pi(y, x) \tag{1}$$

For any pair of player "types," $\alpha, \beta \in (-1, 1)$, this defines a normal-form game $G = \langle \pi, \alpha, \beta \rangle$. We focus on such games that have a unique Nash equilibrium, that, moreover, is pure, interior and regular.

A necessary first-order condition for a pair $(x, y) > 0$ to be a Nash equilibrium is

$$\begin{cases} \pi_1(x, y) + \alpha \cdot \pi_2(y, x) = 0 \\ \pi_1(y, x) + \beta \cdot \pi_2(x, y) = 0. \end{cases} \tag{2}$$

For fixed α and β, this pair of equations can be written on the form $\Phi(x, y) = 0$, where $\Phi : \mathbb{R}^2_{++} \to \mathbb{R}^2$ is the continuously differentiable function defined by the left-hand side of (2). The Nash equilibrium is *regular* if the Jacobian of Φ has non-zero determinant at the solution point (x, y). In this case, there exists a neighborhood B of (α, β) and a continuously differentiable function $x^* : B \to \mathbb{R}^2$ that to each parameter pair (α', β') in B assigns the associated (unique) solution $(x^*(\alpha', \beta'), x^*(\beta', \alpha'))$ to (2) when α is replaced by α' and β by β'. The action pair $(x^*(\alpha', \beta'), x^*(\beta', \alpha'))$ is the unique Nash equilibrium of the game

$G = \langle \pi, \alpha', \beta' \rangle$. We thus have

$$\begin{cases} \pi_1 \left[x^* \left(\alpha', \beta' \right), x^* \left(\beta', \alpha' \right) \right] + \alpha' \cdot \pi_2 \left[x^* \left(\beta', \alpha' \right), x^* \left(\alpha', \beta' \right) \right] = 0 \\ \pi_1 \left[x^* \left(\beta', \alpha' \right), x^* \left(\alpha', \beta' \right) \right] + \beta' \cdot \pi_2 \left[x^* \left(\alpha', \beta' \right), x^* \left(\beta', \alpha' \right) \right] = 0 \end{cases} \tag{3}$$

for all $(\alpha', \beta') \in B$. For all $(\alpha', \beta') \in B$, let $V(\alpha', \beta')$ be the material payoff to the α-altruist in the unique Nash equilibrium of G:

$$V(\alpha', \beta') = \pi \left[x^* \left(\alpha', \beta' \right), x^* \left(\beta', \alpha' \right) \right]. \tag{4}$$

2.2 The matching into pairs

Individuals in a large population are matched into pairs to play the above game. When more than one degree of altruism is present in the population, the expected payoff to an individual with a given degree of altruism will depend on the population distribution and the matching process. We allow for non-uniform random matching of individuals and follow the "algebra of assortative encounters" developed by Bergstrom (2003).

Suppose that two degrees of altruism are present in the population. Let $1 - \varepsilon$ be the proportion of α-altruists and ε that of β-altruists, for some $\varepsilon \in (0, 1)$ and $\alpha, \beta \in A = (-1, 1)$ with $\alpha \neq \beta$. Let $\sigma_\varepsilon \in [0, 1]$ be the difference between the conditional probabilities for an individual to be matched with an α-altruist, given that the individual is either also an α-altruist, or, alternatively, a β-altruist:

$$\sigma_\varepsilon = \Pr[\alpha | \alpha, \varepsilon] - \Pr[\alpha | \beta, \varepsilon]. \tag{5}$$

The following equation is a necessary balancing condition for the number of matches between α- and β-altruists:

$$(1 - \varepsilon) \cdot [1 - \Pr[\alpha | \alpha, \varepsilon]] = \varepsilon \cdot \Pr[\alpha | \beta, \varepsilon]. \tag{6}$$

Equations (5) and (6) together allow the two conditional probabilities to be expressed explicitly in terms of the index of assortativity and population mix:

$$\begin{cases} \Pr[\alpha | \alpha, \varepsilon] = \sigma_\varepsilon + (1 - \varepsilon)(1 - \sigma_\varepsilon) \\ \Pr[\alpha | \beta, \varepsilon] = (1 - \varepsilon)(1 - \sigma_\varepsilon) \end{cases} \tag{7}$$

Hence, the expected material payoff to an α-individual (in population share $1 - \varepsilon$) is

$$W^\alpha(\alpha, \beta, \varepsilon) = [\sigma_\varepsilon + (1 - \varepsilon)(1 - \sigma_\varepsilon)] \cdot V(\alpha, \alpha) + \varepsilon (1 - \sigma_\varepsilon) \cdot V(\alpha, \beta) \tag{8}$$

and that to an β-individual (in population share ε) is

$$W^{\beta}(\alpha,\beta,\varepsilon) = (1-\varepsilon)(1-\sigma_{\varepsilon})\cdot V(\beta,\alpha) + [\sigma_{\varepsilon} + \varepsilon(1-\sigma_{\varepsilon})]\cdot V(\beta,\beta) \tag{9}$$

We assume that the matching is such that σ_{ε} is continuous in ε with limit $\sigma_0 \geq 0$ as $\varepsilon \to 0$. In the literature, the limit value σ_0 is sometimes called the *index of assortativity* (Bergstrom, 2003). We will henceforth refer to the case $\sigma_0 = 0$ as *uniform* random matching, while cases with $\sigma_0 > 0$ will be called *assortative* matching. A canonical example of the latter is sibling interaction: then $\sigma_0 = 1/2$ (see Alger and Weibull, 2010); for other examples, see Section 7. For ease of notation we will henceforth write σ for σ_0.

2.3 Evolutionary stability

The evolutionary success of a degree of altruism is assumed to depend on the average (material) payoff to its carriers. Each matched pair plays a game of the type described in section 2.1 once, and the two individuals obtain their equilibrium (material) payoffs, defined in equation (4). Consider a population consisting mostly of α-altruists, for some $\alpha \in (-1,1)$ and a few β-altruists, for some $\beta \neq \alpha$. Let the population shares be $1-\varepsilon$ and ε, respectively, for some small $\varepsilon \in (0,1)$. The "incumbent" degree of altruism, α, is *evolutionarily stable* against the "mutant" degree, β, if α-altruists on average earn higher equilibrium (material) payoffs than β-altruists, in their respective matches, for all sufficiently small population shares $\varepsilon > 0$ of β-altruists. The degree of altruism α is *evolutionarily stable* if this is true for every $\beta \neq \alpha$.

By continuity of payoffs, a necessary condition for a degree of altruism α to be *evolutionarily stable against a degree* $\beta \neq \alpha$ is that $\lim_{\varepsilon\to 0} W^{\alpha}(\alpha,\beta,\varepsilon) \geq \lim_{\varepsilon\to 0} W^{\beta}(\alpha,\beta,\varepsilon)$, and a sufficient condition is that this inequality holds strictly, or, equivalently, that

$$V(\alpha,\alpha) > (1-\sigma)V(\beta,\alpha) + \sigma V(\beta,\beta). \tag{10}$$

The left-hand side is the average material payoff earned by an α-altruist and the right-hand side that of a β-altruist, in the limit as the population share of β-individuals tends to zero. The following formal definition is thus a slight strengthening of the verbal definition above (see also Alger and Weibull, 2010):

Definition 1 *A degree of altruism α is **evolutionarily stable** if inequality (10) holds for all $\beta \neq \alpha$, and α is **locally evolutionarily stable** if (10) holds for all β in some neighborhood of α.*

3 Analysis

Given the degree of altruism, α, and the index of assortativity, σ, the right-hand side of (10) is a function of β. At $\beta = \alpha$, this function obtains the value $V(\alpha, \alpha)$. Hence, a degree of altruism α is locally evolutionarily stable if and only if the right-hand side of (10) has a strict local maximum at $\beta = \alpha$. This simple observation turns out to provide analytical power. For any given index of assortativity $\sigma \in [0, 1]$, Let $H : A^2 \to R$ be defined by

$$H(\alpha, \beta) = V(\beta, \alpha) - V(\alpha, \alpha) + \sigma \cdot [V(\beta, \beta) - V(\beta, \alpha)]. \tag{11}$$

This is the *material payoff advantage* of the "mutant" degree of altruism β, when this appears in an infinitesimal population share. Evidently $H(\alpha, \alpha) = 0$ for all $\alpha \in A$, and α is evolutionarily stable if and only if $H(\alpha, \beta) < 0$ for all $\beta \neq \alpha$. Likewise, α is locally evolutionarily stable if and only if $H(\alpha, \beta) < 0$ for all $\beta \neq \alpha$ in some neighborhood of α.

Using subscripts for partial derivatives, the derivative of H with respect to its second argument, the mutant degree of altruism β, is

$$D(\alpha, \beta) = (1 - \sigma) \cdot V_1(\beta, \alpha) + \sigma \cdot [V_1(\beta, \beta) + V_2(\beta, \beta)]. \tag{12}$$

Slightly generalizing Alger and Weibull (2010), we call $D : A \to \mathbb{R}^2$ the *drift function*.

Proposition 1 *Condition (i) below is necessary for a degree of altruism $\alpha \in A$ to be locally evolutionarily stable (and hence also for being evolutionarily stable). Conditions (i) and (iii) are together sufficient for α to be evolutionarily stable (and hence also locally evolutionarily stable). Conditions (i) and (ii) are together sufficient for α to be locally evolutionarily stable.*

(i) $D(\alpha, \alpha) = 0$

(ii) $D(\alpha, \beta) > 0$ *for all* $\beta < \alpha$ *and* $D(\alpha, \beta) < 0$ *for all* $\beta > \alpha$

(iii) $D_2(\alpha, \alpha) < 0$

For the class of games specified in Section 2, the drift function can be expressed in terms of equilibrium behavior as follows:

Lemma 1 *Let $\sigma \in [0, 1]$ and $\alpha \in A$. Then*

$$D(\alpha, \alpha) = [(\sigma - \alpha) \cdot x_1^*(\alpha, \alpha) + (1 - \sigma\alpha) \cdot x_2^*(\alpha, \alpha)] \cdot \pi_2[x^*(\alpha, \alpha), x^*(\alpha, \alpha)]. \tag{13}$$

The terms that involve α reflect the changes in the mutant's own behavior brought about by a change in the mutant degree of altruism, because the mutant carries it himself (the term $x_1^*(\alpha,\alpha)$), and because with probability σ the other individual carries it, too (the term $x_2^*(\alpha,\alpha)$). The factor $\alpha\pi_2[x^*(\alpha,\alpha), x^*(\alpha,\alpha)]$ shows how the behavior of the mutant under consideration differs from the behavior that would maximize material payoff, as long as he is not selfish ($\alpha \neq 0$). The other terms show how the other individual adjusts behavior in response to a change in the mutant degree of altruism. First, whether a mutant or not, the other individual adapts behavior by $x_2^*(\alpha,\alpha)$ to the presence of the mutant degree of altruism in the mutant under consideration. Second, with probability σ the other individual also carries the mutant degree of altruism, and hence adjusts behavior by $x_1^*(\alpha,\alpha)$.

We conclude by noting that for games with $\pi_2 \neq 0$, the necessary condition (i) for local evolutionary stability boils down to

$$(\sigma-\alpha)\cdot x_1^*(\alpha,\alpha)+(1-\sigma\alpha)\cdot x_2^*(\alpha,\alpha)=0. \tag{14}$$

A mutation in the degree of altruism typically involves both a material cost and a material benefit, and the relative magnitude of these will determine the value of the stable degree of altruism. Since the said cost and benefit depend on how the two individuals' behavior is affected by their degrees of altruism, which we presume are known by both, evolutionary stability rests on how individuals adapt their behavior to their own and the other's degrees of altruism — the terms $x_1^*(\alpha,\alpha)$ and $x_2^*(\alpha,\alpha)$ in condition (13). The nature of these adaptations is in turn determined by the properties of the (material) payoff function π. More precisely, the adaptations depend on how the marginal material return changes when the other individual changes his or her behavior.

Definition 2 *Let $\pi \in \Pi$. The actions are* strategically independent *if $\pi_{12}(x,y)=0$ for all $x,y>0$. The actions are* strategic substitutes *if $\pi_{12}(x,y)<0$ for all $x,y>0$ and* strategic complements *if $\pi_{12}(x,y)>0$ for all $x,y>0$.*

Lemma 2 *Let $\pi \in \Pi$ and suppose that the game $G=\langle \pi,\alpha,\alpha\rangle$ has a unique Nash equilibrium (x^*,x^*) and that this is regular. Then*

$$x_1^*(\alpha,\alpha)\cdot\pi_2(x,y)\mid_{x=y=x^*(\alpha,\alpha)}>0$$

and

1. *$x_1^*(\alpha,\alpha)\neq 0$ and $x_2^*(\alpha,\alpha)=0$ if the actions are strategically independent.*

2. $x_1^*(\alpha,\alpha)\cdot x_2^*(\alpha,\alpha)<0$ *if the actions are strategic substitutes.*

3. $x_1^*(\alpha,\alpha)\cdot x_2^*(\alpha,\alpha)>0$ *if the actions are strategic complements.*

A change in own altruism always affects own behavior. Interpreting actions as "efforts": a more altruistic individual will exert more (less) effort if this benefits (hurts) the other. The other individual will react by either reciprocating, if actions are strategic complements, or reducing her effort, if actions are strategic substitutes.

Combining the two lemmas, we obtain:

Proposition 2 *For any* $\sigma\in[0,1)$ *and any locally evolutionarily stable degree of altruism* $\alpha\in A$*:*

1. $\alpha=\sigma$ if the actions are strategically independent.

2. $\alpha<\sigma$ if the actions are strategic substitutes.

3. $\alpha>\sigma$ if the actions are strategic complements.

To enhance one's intuition for this result, it is useful to consider the special case of uniform random matching, that is, $\sigma=0$. Is selfishness ($\alpha=0$) stable, as Hamilton's rule would suggest? Imagine a population with selfish incumbents ($\alpha=0$). A mutant who is slightly altruistic makes a higher effort than an incumbent, while a mutant who is slightly spiteful makes a lower effort than an incumbent. If actions are strategically independent, then both types of mutant obtain a lower material payoff than a selfish incumbent. Hence, selfishness is stable. By contrast, if actions are strategic substitutes, a selfish incumbent will adapt his or her effort to the anticipated low effort of a spiteful mutant by way of increasing his or her own effort. This benefits the mutant. Hence, spitefulness can "invade" a selfish population, and evolutionary stability requires a degree of spitefulness ($\alpha<0$). If actions instead are strategic complements, a selfish incumbent will adapt his or her effort to the anticipated higher effort of an altruistic mutant by increasing his or her own effort. This benefits the mutant, so evolutionary stability requires some altruism ($\alpha>0$).[3]

[3]Rotemberg (1994) draws a similar conclusion in a model where co-workers choose their own levels of altruism, towards each other, before they simultaneously choose their individual effort levels according to their altruistic preferences.

To see that $\alpha = \sigma$ whenever actions are strategically independent, note that, because $x_1^*(\alpha, \alpha) \neq 0$ and $x_2^*(\alpha, \alpha) = 0$, (13) becomes

$$D(\alpha, \alpha) = \sigma \cdot x_1^*(\alpha, \alpha) - \alpha \cdot x_1^*(\alpha, \alpha).$$

In this expression, the term $\sigma \cdot x_1^*(\alpha, \alpha)$ represents the mutant's expected "material benefit" from the opponent's effort adaptation, which occurs only if the opponent is also a mutant (thus the factor σ). The second term, $-\alpha \cdot x_1^*(\alpha, \alpha)$ represents the mutant's "material cost" of his or her own altruism; an altruist's equilibrium action differs from the one that would maximize own material payoff.

More generally, suppose that $\alpha = \sigma$. Then equation (13) gives

$$D(\sigma, \sigma) = \left(1 - \sigma^2\right) \cdot x_2^*(\alpha, \alpha) \cdot \pi_2\left[x^*(\alpha, \alpha), x^*(\alpha, \alpha)\right].$$

This expression shows that the strength of the drift away from the "Hamiltonian" degree of altruism $\alpha = \sigma$ depends on how strongly individuals adjust their behavior to their opponent's degree of altruism/spitefulness and on how sensitive their material payoff is to the other's action.

The observation in Lemma 2 — that a more altruistic individual will exert more effort if this benefits the other, and less effort if it hurts the other — implies that it is always beneficial to meet a more altruistic opponent; $V_2(\alpha, \beta) > 0$. As a result, an increase in the index of assortativity raises the benefit of mutating towards higher degrees of altruism (see (12)), and, hence:

Proposition 3 *If α is the unique evolutionarily stable degree of altruism for some degree of assortativity σ, and α' is evolutionarily stable for $\sigma' > \sigma$, then $\alpha' > \alpha$.*

We proceed to illustrate the above results in a number of applications.

3.1 Applications

3.1.1 CES production of a public good

We will consider some applications to situations in which the two individuals simultaneously choose how much *effort* to exert in the production of a public good,

$$\pi(x, y) = B(x, y) - C(x)$$

for some *production* function B and *cost* function C. The production functions B to be considered are symmetric in the sense that $B(x,y) = B(y,x)$ for all $x, y \geq 0$. Both B and C are twice differentiable and strictly increasing, with B concave and C convex. An alternative interpretation is that the output is a private good, with total output always split in equal shares.[4]

Let us first consider production functions with constant elasticity of substitution,

$$B(x,y) = \left(x^{\rho} + y^{\rho}\right)^{1/\rho}.$$

for $\rho \leq 1$. Assuming the cost function to be a power function, we write material payoffs in the form

$$\pi(x,y) = \left(x^{\rho} + y^{\rho}\right)^{1/\rho} - cx^{1+\nu}$$

for some $c, \nu > 0$. Since

$$\pi_{12}(x,y) = (1-\rho)(xy)^{\rho-1}\left(x^{\rho} + y^{\rho}\right)^{\frac{1-2\rho}{\rho}},$$

it follows that actions are strategic complements when $\rho < 1$, and strategically independent when $\rho = 1$.

Proposition 4 *If α^* is locally evolutionarily stable, then*

$$\alpha^* = \frac{2\sigma\nu + (1+\sigma)(1-\rho)}{2\nu + (1+\sigma)(1-\rho)}.$$

Since

$$\frac{d\alpha^*}{d\rho} = \frac{2\nu(1+\sigma)(\sigma-1)}{\left[2\nu + (1+\sigma)(1-\rho)\right]^2} < 0$$

and

$$\frac{d\alpha^*}{d\nu} = \frac{2(1-\rho)(1+\sigma)(\sigma-1)}{\left[2\nu + (1+\sigma)(1-\rho)\right]^2} < 0$$

we conclude that a stable degree of altruism is decreasing in the degree of substitutability between the two individuals' efforts.

Figure 1 shows α^* as a function of the index of assortativity σ, for $\nu = 1$, when $\rho = 0.5$ (thick curve) and when $\rho = 1$ (thin curve).

[4] To see this, write $B(x,y) = \frac{1}{2}B^*(x,y)$, where $B^*(x,y)$ is total output of the private good.

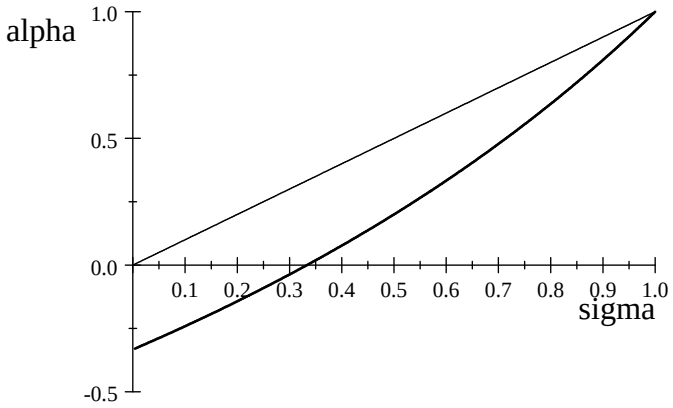

Figure 1: The stable degree of altruism as a function of σ.

3.2 Rent-seeking

We turn now to rent-seeking, that is, situations in which each of the two individuals undertakes some effort, $x \geq 0$ and $y \geq 0$, respectively, and a resource of given size is divided proportionately according to efforts.[5] We focus on the following classical representation (see Tullock, 1980):

$$\pi(x, y) = \frac{x}{x+y} - cx,$$

where $c > 0$. We note the discontinuity at zero and that this game does not fall into any of the three interaction classes studied in general above; the other's effort may increase or decrease the marginal benefit of one's own action, depending on whose effort is biggest,

$$\pi_{12}(x, y) = \frac{x-y}{(x+y)^3}.$$

The necessary first-order condition for interior Nash equilibrium (2) is

$$(1-\alpha)\, y = c\,(x+y)^2 \text{ and } (1-\beta)\, x = c\,(x+y)^2, \tag{15}$$

which gives

$$x^*(\alpha, \beta) = \frac{(1-\alpha)^2(1-\beta)}{c\,(2-\alpha-\beta)^2} \text{ and } x^*(\beta, \alpha) = \frac{(1-\beta)^2(1-\alpha)}{c\,(2-\alpha-\beta)^2}.$$

It is easily verified that this is indeed the unique Nash equilibrium. We note that $x_1^*(\alpha, \alpha) \neq 0$ and $x_2^*(\alpha, \alpha) = 0$:

$$x_1^*(\alpha, \beta) = -\frac{2(1-\alpha)(1-\beta)^2}{(2-\alpha-\beta)^3}$$

[5] With the convention that they each obtain half the resource if $x = y = 0$.

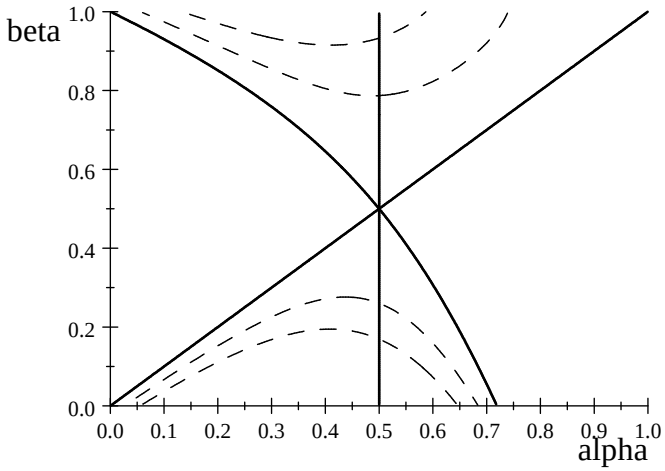

Figure 2: Level curves of H, the difference in expected fitness of an incumbent with altruism α and a mutant with altruism β in a rent-seeking game.

$$x_2^*(\alpha,\beta) = \frac{(1-\alpha)^2(\alpha-\beta)}{(2-\alpha-\beta)^3}.$$

From Lemma 1 we immediately obtain:

$$D(\alpha,\alpha) = 0 \Leftrightarrow \alpha = \sigma.$$

Is this locally evolutionarily stable degree of altruism also globally stable? Figure 2 shows level curves of H (see (11)). The solid curves correspond to those (α,β) where H vanishes. $H(\alpha,\beta)$ is negative for (α,β) in the north and south regions, and $H(\alpha,\beta)$ is positive for (α,β) in the east and west regions. The dashed curves are negative level curves for the function H. We see that H has a saddle point at $\alpha = \beta = 1/2$.

We have thus found that the unique stable degree of sibling altruism is $1/2$. How can this result be explained, given the non-linear form of strategic interaction? Recall that

$$D(\alpha,\alpha) = \left[(1/2-\alpha)\,x_1^*(\alpha,\alpha) + (1-\alpha/2)\,x_2^*(\alpha,\alpha)\right]\cdot\pi_2\left(x^*(\alpha,\alpha), x^*(\alpha,\alpha)\right).$$

Hence, $D(1/2,1/2) = 0$, because $x_2^*(\alpha,\alpha) = 0$. This is illustrated in Figure 3, where we plot the best-response functions defined in (15). The solid upward-bending curve is the best-response of the α-altruist (x as a function of y) and the solid downward-bending curve is the best response of a β-altruist (y as a function of x). These curves are drawn for the case of equal altruism, $\alpha = \beta = 1/2$. The Nash equilibrium is the intersection of these two curves.

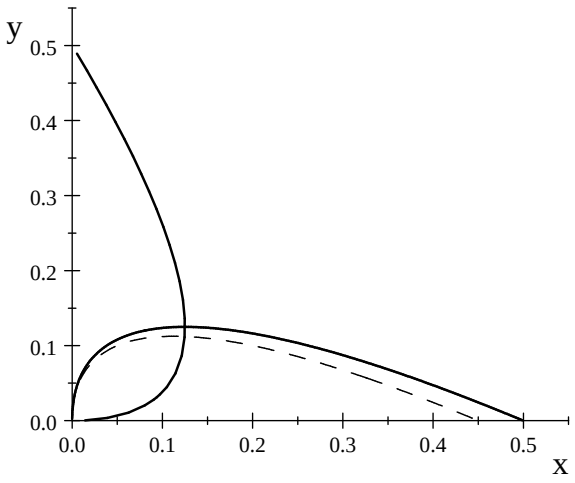

Figure 3: Best response functions in the rent-seeking game, for $c = 1$, $\alpha = \beta = 1/2$ (solid curves) and $\beta = 0.55$ (dashed curve).

An increase in β would cause a downward shift in the best response function; the dashed curve is the best response y to x for $\beta = 0.55$. Such an increase in β causes the β-altruist to decrease her effort, y. However, small changes in β has no first-order effect on the α-altruist; the latter does not alter his or her effort x. This is seen in the diagram: the best-response curve for x against y is horizontal at the intersection $(x^*(\alpha,\alpha), x^*(\alpha,\alpha))$. Hence, $x_2^*(\alpha,\alpha) = 0$, and so *the strategic externality vanishes at equilibrium*!

The result that $\alpha = \sigma$ is the unique locally evolutionarily stable degree of altruism holds for any convex cost function. Suppose, thus, that

$$\pi(x,y) = \frac{x}{x+y} - C(x), \tag{16}$$

where $C' > 0$, $C'' \geq 0$, and $C'(0) = 0$.

Proposition 5 *For material payoff functions of the form (16), if α^* is locally evolutionarily stable, then $\alpha = \sigma$.*

4 n-player interactions

Many interactions involve more than two individuals. Can the definitions and results derived above be applied in populations where individuals interact in n-person groups?

4.1 The class of games

Consider an n-player game, for $n \geq 2$, in which each player has the continuum set $S = \mathbb{R}_+$ of pure strategies, or *actions*, and where the material *payoff* from using strategy $x_i \in S$ against the strategies $x_j \in S$, $j \neq i$, is $\pi(x_i, \mathbf{x}_{-i})$, where $\pi : \mathbb{R}_+^n \to \mathbb{R}$ is twice differentiable and symmetric in $\mathbf{x}_{-i}$.[6] We assume $\pi_1 > 0$, $\pi_{11} < 0$, $\pi_j \neq 0$ and $\pi_j \cdot \pi_{jj} \leq 0$ for all $j > 1$. Actions are strategic substitutes if $\pi_{1j} < 0$ for all $j > 1$, strategic complements if $\pi_{1j} > 0$ for all $j > 1$, and strategically independent if $\pi_{1j} = 0$ for all $j > 1$.

Individuals may be more or less altruistic or spiteful towards each other, so the *utility,* for an individual i with degree till $\alpha_i \in (-1, 1)$ of altruism towards all the others, from playing strategy $x_i \in S$ against $\mathbf{x}_{-i} \in S^{n-1}$, is

$$u(x_i, \mathbf{x}_{-i}, \alpha) = \pi(x_i, \mathbf{x}_{-i}) + \alpha_i \sum_{j \neq i} \pi(x_j, \mathbf{x}_{-j}).$$

The players choose their strategies simultaneously. For any two degrees of altruism, α and β, and any number $k \in \{0, 1, 2, ..., n\}$ of individuals with degree α, and remaining number, $n-k$, of individuals with degree β, this defines a normal-form game $G = \langle \pi, \alpha, \beta, n, k \rangle$. This game is symmetric if $\alpha = \beta$ or $k \in \{0, n\}$. Just as in the special case $n = 2$, we focus here on games $G = \langle \pi, \alpha, \beta, n, k \rangle$ that have a unique Nash equilibrium, and where this equilibrium is interior and regular.

This uniqueness assumption implies that all individuals with the same degree of altruism take the same equilibrium actions. To see this, let $\mathbf{x}^*$ be the unique Nash equilibrium and suppose that $\alpha_i = \alpha_j$ and $x_i^* \neq x_j^*$. Let $\mathbf{x}^o$ be defined by $x_i^o = x_j^*$, $x_j^o = x_i^*$ and $x_k^o = x_k^*$ for all $k \notin \{i, j\}$. Then also $\mathbf{x}^o$ is a Nash equilibrium, since by hypothesis π is symmetric and thus $\arg\max_{x_i} u\left(x_i, \mathbf{x}_{-i}^*, \alpha_i\right) = \arg\max_{x_j} u\left(x_j, \mathbf{x}_{-j}^o, \alpha_j\right)$, and likewise with i and j exchanged. Moreover,

$$\arg\max_{x_k} u\left(x_k, \mathbf{x}_{-k}^*, \alpha_i\right) = \arg\max_{x_k} u\left(x_k, \mathbf{x}_{-k}^o, \alpha_k\right) \ \forall k \notin \{i, j\}.$$

Henceforth, let $x^*(\alpha_i, \boldsymbol{\alpha}_{-i}, k)$ denote the equilibrium strategy of an individual with altruism $\alpha_i \in \{\alpha, \beta\}$ playing a game $G = \langle \pi, \alpha, \beta, n, k \rangle$ against players whose altruism levels are represented by the vector $\boldsymbol{\alpha}_{-i} \in \{\alpha, \beta\}^{n-1}$, and let $V(\alpha_i, \boldsymbol{\alpha}_{-i}, k)$ denote this individual's equilibrium material payoff. The group size n will be kept fixed and notationally suppressed.

[6] By *symmetry* we mean invariance under permutation of the arguments: for any $x_i \in \mathbb{R}_+$ and $x_{-i} \in \mathbb{R}_+^{n-1}$, and any bijection $h : \{1, 2, ..., n-1\} \to \{1, 2, ..., n-1\}$: $\pi\left(x_i, x_{h(1)}, x_{h(2)}, ..., x_{h(n-1)}\right) = \pi(x_i, x_{-i})$.

For each player i (and with altruism α_i), let $\mathbf{x}^*_{-i}$ be the vector of the equilibrium strategies of the other players, in game $G = \langle \pi, \alpha, \beta, n, k \rangle$. Thus

$$V(\alpha_i, \boldsymbol{\alpha}_{-i}, k) = \pi\left(x^*(\alpha_i, \boldsymbol{\alpha}_{-i}, k), \mathbf{x}^*{}_{-i}\right).$$

Let $\hat{\boldsymbol{\alpha}}_{-i}$ denote the $(n-1)$-vector, and $\hat{\boldsymbol{\alpha}}$ the n-vector, whose elements all equal α. Focusing on games where x^* is differentiable at the point $(\alpha, \hat{\boldsymbol{\alpha}}_{-i})$, let $x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, k)$ denote the partial derivative with respect to the individual's own altruism (α), and let $x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}, k)$ the partial derivative with respect to any other individual's altruism (also α). Similarly, let $\pi_{own}(\mathbf{x}^*(\hat{\boldsymbol{\alpha}}))$ denote the partial derivative of the material payoff with respect to the individual's own action, and $\pi_{other}(\mathbf{x}^*(\hat{\boldsymbol{\alpha}}))$ the partial derivative with respect to any other individual's action, at the point $\hat{\boldsymbol{\alpha}}$.

4.2 Local evolutionary stability

Consider a large population consisting of a large share of α-altruists and a small share of β-altruists, for some $\alpha, \beta \in (-1, 1)$ with $\alpha \neq \beta$. Let $\varepsilon \in (0, 1)$ be the population share of β-altruists. Individuals are randomly matched into groups of size n, where the random draws to a group need not be statistically independent. We define the index of assortativity, $\sigma_\varepsilon \in [0, 1]$, to be the difference between two conditional probabilities. The first probability is the probability that, for any given α-altruist in the group, a (uniformly) randomly drawn other group member is also an α-altruist. The second probability is the probability for the same event, for any given β-altruist in the group. Suppose that σ_ε converges to $\sigma \in [0, 1]$ as $\varepsilon \to 0$.[7] Our previous condition for evolutionary stability of a degree of altruism α, for $n = 2$, generalizes to

$$V(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) > \sum_{m=0}^{n-1} \binom{n-1}{m} \sigma^m (1-\sigma)^{n-1-m} V(\beta, \boldsymbol{\alpha}_{-i}, m+1) \quad \forall \beta \neq \alpha. \tag{17}$$

While the left-hand side is the material equilibrium payoff to an incumbent (α-altruist) when all group members are α-altruists, the right-hand side is the expected material payoff to a mutant, with altruism $\beta \neq \alpha$, according to the binomial distribution of all possible group compositions with at least one β-altruist (m is the number of other mutants in the group, so that the total number of mutants in the group is $m + 1$). As in the special case

[7] For instance, under sexual reproduction, haploid genetics, and random pairwise matching between parents, $\sigma = 1/2$ for siblings.

of pairwise interactions, we note that the right-hand side equals the left-hand side when $\beta = \alpha$. Accordingly, the material payoff advantage of a mutant with degree of altruism β in a population consisting mostly of α-altruists is

$$H(\alpha, \beta) = \sum_{m=0}^{n-1} \binom{n-1}{m} \sigma^m (1-\sigma)^{n-1-m} V(\beta, \boldsymbol{\alpha}_{-i}, m+1) - V(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n),$$

and the drift function D, which is the derivative of H with respect to the mutant degree of altruism β is:

$$D(\alpha, \beta) = \sum_{m=0}^{n-1} \binom{n-1}{m} \sigma^m (1-\sigma)^{n-1-m} \left[\frac{\partial V(\beta, \boldsymbol{\alpha}_{-i}, m+1)}{\partial \beta} \right]_{\beta=\alpha}. \tag{18}$$

Remark 1 *In the special case $n = 2$, we obtain*

$$\begin{aligned} D(\alpha, \beta) &= (1-\sigma) \cdot \frac{\partial V(\beta, \alpha, 1)}{\partial \beta} + \sigma \cdot \frac{\partial V(\beta, \beta, 2)}{\partial \beta} \\ &= (1-\sigma) \cdot V_1(\beta, \alpha) + \sigma \cdot [V_1(\beta, \beta) + V_2(\beta, \beta)]. \end{aligned}$$

In the two-player case, we have an expression for the drift that relies directly on how equilibrium actions depend on marginal changes in own altruism and in the other's altruism, see (13). The following lemma gives the corresponding expression for the case with $n \geq 2$ players.

Lemma 3 *Let $\alpha \in \mathcal{A}$. Then*

$$\begin{aligned} D(\alpha, \alpha) &= (n-1) \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] \cdot \\ &\quad [(\sigma - \alpha) \cdot x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) + [1 - \sigma(n-1)\alpha + \sigma(n-2)] \cdot x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)]. \end{aligned} \tag{19}$$

As in the two-player case, this equation shows that the stability of a degree of altruism in general depends on how a marginal increase in an individual's own altruism would affect his own behavior, $x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)$, as well on how it would affect other individuals' behavior, $x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)$. The factors by which these effects are multiplied reflect the fact that such a marginal increase in own altruism comes about by a marginal change in the mutant degree of altruism and hence affect other members of the group, depending on the index of assortativity, σ, and group size, n.[8] Finally, the common factor, $\pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})]$, accounts for how these marginal changes in individual behaviors affect the level of material payoff.

[8] These other mutants in the group adjust behavior to the presence of the mutant degree of altruism in themselves (the term $(n-1)\sigma x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)\pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})]$), and in each other (the term $(n-2)(n-1)\sigma x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)\pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})]$).

4.3 The two principles

For pairwise interactions, we identified two principles: first, that a higher index of assortativity implies a higher stable degree of altruism, and second, that whether the stable degrees of altruism are lower or higher than the index of assortativity depends on whether actions are strategic substitutes or complements. Here we show that these principles carry over to (symmetric) interactions in groups of arbitrary (finite) size.

Proposition 6 *For any $\sigma \in [0, 1)$ and any locally evolutionarily stable degree of altruism $\alpha \in \mathcal{A}$:*

1. $\alpha = \sigma$ if the actions are strategically independent.
2. $\alpha < \sigma$ if the actions are strategic substitutes.
3. $\alpha > \sigma$ if the actions are strategic complements.

Proposition 7 *Let α and α' be the unique evolutionarily stable degrees of altruism, for degrees of assortativity $\sigma < \sigma'$, respectively. Then $\alpha < \alpha'$ if the actions in G are strategically independent or strategic complements, or if they are strategic substitutes and $\alpha > 0$.*

Since free-riding in public goods games played by selfish individuals usually depends heavily on the size of the group, one might believe that the same might hold for the stable degree of altruism in such interactions, and, more generally, that the stable degree of altruism might be quite sensitive to group size. As the following examples show, this is not at all the case. The reason for the relative insensitivity to group size has to do with the fact that the expected average degree of altruism does not change much with group size.

4.4 Stable public-goods provision

Here we consider applications to situations in which the two individuals simultaneously choose how much *effort* to exert in the production of a public good.

4.4.1 When efforts are additive

Suppose first that the individual contributions/efforts, $x_i \in \mathbb{R}_+$, are perfect substitutes, output is a power function of total effort, and the cost is quadratic in effort:

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = X^\tau - \frac{c}{2}x_i^2$$

where $X = \sum_{i=1}^n x_i$, for some $\tau \in (0,1]$ and $c > 0$. Here efforts are strategic substitutes if $\tau < 1$ and strategically independent if $\tau = 1$. The *utility* for an individual i with degree $\alpha \in (-1,1)$ of altruism towards all the others, from playing strategy $x_i \in S$ against $x_j \in S$, $\forall j \neq i$, is

$$u\left(x_i, x_{-i}, \alpha\right) = \left[1 + (n-1)\,\alpha\right] X^\tau - \frac{c}{2}x_i^2 - \alpha \cdot \frac{c}{2}\sum_{j \neq i} x_j^2.$$

Proposition 8 *For $\sigma \in [0,1]$ and $n \geq 2$, if α^* is locally evolutionarily stable, then*

$$\alpha^* = \frac{(\tau-1)\left[1+\sigma\left(n-1\right)\right] + \sigma n\left(2-\tau\right)}{(\tau-1)\left[1+\sigma\left(n-1\right)\right] + n\left(2-\tau\right)}.$$

In particular, Hamilton's rule holds in the linear case $\tau = 1$:

Corollary 1 *For $\tau = 1$ and any n, if α^* is locally evolutionarily stable, then $\alpha^* = \sigma$.*

By contrast, α^* always falls short of σ in the strategic substitutes case ($\tau < 1$). Moreover, α^* is increasing in σ, n, and τ.[9] Figure 4 shows how α^* depends on τ, for $\sigma = 1/2$, and $n = 2, 3, 6, 12$ (full solid, dashed solid, full thin and dashed thin).

We also see that changes in n do not affect the stable degree very much. In the limit, as $n \to \infty$, we obtain

$$\alpha^* \to \frac{\sigma}{(1-\sigma)\left(1-\tau\right)+1},$$

see Figure 5 (solid curve, with thin dashed curve for $n = 12$).

4.4.2 When efforts are multiplicative

Assume now that production is multiplicative in efforts:

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = \left(\prod_{i=1}^n x_i^{1/n}\right)^\theta - \frac{c}{2}x_i^2$$

[9] $\frac{d\alpha^*}{d\sigma} = \frac{n^2(2-\tau)}{[(\tau-1)[1+\sigma(n-1)]+n(2-\tau)]^2}$, $\frac{d\alpha^*}{dn} = \frac{(1-\tau)(2-\tau)(1-\sigma)^2}{[(\tau-1)[1+\sigma(n-1)]+n(2-\tau)]^2}$, and $\frac{d\alpha^*}{d\tau} = \frac{n[1+\sigma(n-1)](1-\sigma)}{[(\tau-1)[1+\sigma(n-1)]+n(2-\tau)]^2}$.

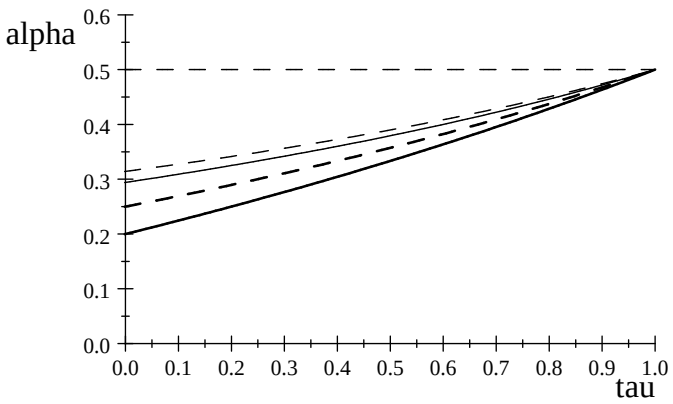

Figure 4: Stable degree of altruism in the public goods game with strategic substitutes, for $n = 2, 3, 6, 12$ (full solid, dashed solid, full thin and dashed thin).

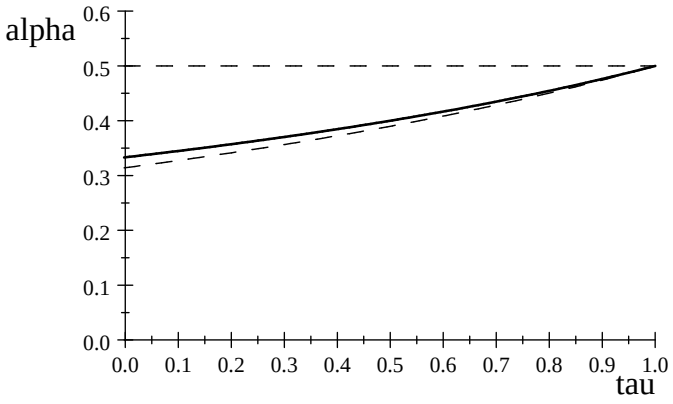

Figure 5: Asymptotic value of α^* in the public goods game with strategic substitutes (solid curve), and for $n = 12$ (thin dashed curve).

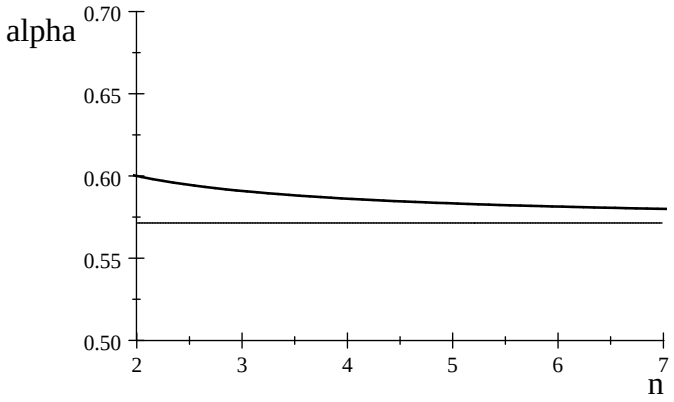

Figure 6: α^* as a function of n, for $\theta = 1/2$.

for some $\theta \in (0,1]$ and $c > 0$**.** Here efforts are strategic complements. The total *utility,* for an individual i with degree $\alpha \in (-1,1]$ of altruism towards the others, from playing strategy $x_i \in S$ against $x_j \in S$, $\forall j \neq i$,is

$$u\left(x_i, \mathbf{x}_{-i}, \alpha\right) = \left[1 + (n-1)\,\alpha\right] \left(\prod_{i=1}^{n} x_i^{1/n}\right)^{\theta} - \frac{c}{2} x_i^2 - \alpha \cdot \frac{c}{2} \sum_{j \neq i} x_j^2.$$

Proposition 9 *For $\sigma \in [0,1]$ and $n \geq 2$, if α^* is locally evolutionarily stable, then*

$$\alpha^* = \frac{\theta\left[1 + \sigma\left(n-1\right)\right] + \sigma n\left(2-\theta\right)}{\theta\left[1 + \sigma\left(n-1\right)\right] + n\left(2-\theta\right)}.$$

Here a stable degree of altruism always exceeds σ, is decreasing in n and increasing in σ and in θ.[10] Moreover, for a given value of θ, a stable degree of altruism is bounded from below by $2\sigma/\left(2 - \left(1-\sigma\right)\theta\right)$. We also note that, for θ close to zero, the game is almost linear, and then $\alpha^* \simeq \sigma$, independently of the number n of individuals. Figure 6 shows α^* as a function of n, for $\sigma = 1/2$, and $\theta = 1/2$, along with its asymptotic value, $\alpha = 1/\left(2 - \theta/2\right) \simeq 0.57$.

5 Repeated interactions

So far, we have assumed that the underlying interaction, as expressed in terms of material payoffs, is a symmetric simultaneous-move game with a unique equilibrium. However, in

[10] $\frac{d\alpha^*}{dn} = -\frac{\theta(2-\theta)(1-\sigma)^2}{[\theta[1+\sigma(n-1)]+n(2-\theta)]^2}$, $\frac{d\alpha^*}{d\sigma} = \frac{2n^2(2-\theta)}{[\theta[1+\sigma(n-1)]+n(2-\theta)]^2}$, and $\frac{d\alpha^*}{d\theta} = \frac{2n[1+\sigma(n-1)](1-\sigma)}{[\theta[1+\sigma(n-1)]+n(2-\theta)]^2}$.

many relevant applications, the interaction is more naturally modelled as a repeated game (and hence may have many equilibria). Repeated games between altruists have not been much analyzed, a notable exception being Bernheim and Stark (1988). As they point out, altruism weakens both the incentive to deviate and the incentive to punish deviations. They find that moderate degrees of altruism may make cooperation more difficult to sustain. We here consider the infinitely repeated play of a symmetric two-player game of the type analyzed above. Hence, the utility to an α-altruist from taking action x against action y is $u(x, y, \alpha)$, as defined in (1). For any pair of player "types," $\alpha, \beta \in (-1, 1)$, this defines the normal form $G = \langle \pi, \alpha, \beta \rangle$ of a simultaneous-move game. As before, we focus on such games that have a unique Nash equilibrium, $(x^*(\alpha, \beta), x^*(\beta, \alpha))$ and where this is interior and regular. Let $u^*(\alpha, \beta)$ and $u^*(\beta, \alpha)$ be the associated equilibrium utilities, and let $V(\alpha, \beta)$ and $V(\beta, \alpha)$ be the corresponding material payoffs. Suppose that this game is repeated indefinitely, with perfect monitoring between periods, and suppose that both players evaluate the outcomes in terms of the average discounted value of the stream of stage-game (altruistic) utilities, where both players discount future utilities by the same factor $\delta \in (0, 1)$. Evolution is supposed to operate on the associated average discounted value of material payoffs. We will first follow Bernheim and Stark (1988) and focus on stationary action profiles supported by the threat of reversion forever to play of the stage-game Nash equilibrium.[11] In a later subsection, we instead use an approach based upon results in Abreu and Pearce (2002, 2007).

5.1 Nash reversion threats

Perpetual play of an action pair (x, y), under threat of Nash reversion, is a subgame perfect equilibrium outcome in the so-defined repeated game if and only if

$$\begin{cases} (1-\delta)\, u^+(\alpha, y) + \delta u^*(\alpha, \beta) \leq u(x, y, \alpha) \\ (1-\delta)\, u^+(\beta, x) + \delta u^*(\beta, \alpha) \leq u(y, x, \beta) \end{cases} . \tag{20}$$

where $u^+(\alpha, y)$ is the maximal deviation utility to the α-altruist, when her opponent plays y:

$$u^+(\alpha, y) = \max_{x \geq 0} u(x, y, \alpha)$$

[11] In general, more severe punishments are available, see Abreu (1986). Note also that in public-provision games among altruists, a player may severely punish an altruistic player by acting "martyr," that is, by making some very costly contribution to the public good.

and likewise for the β-altruist when his opponent plays x:

$$u^{+}(\beta, x) = \max_{y \geq 0} u(y, x, \beta)$$

Action pairs (x, y) satisfying condition (20) will be called *sustainable* in the repeated game, and we denote the set of such action pairs $S(\alpha, \beta, \delta)$. Clearly the stage-game Nash equilibrium always belongs to this set: $(x^{*}(\alpha, \beta), x^{*}(\beta, \alpha)) \in S(\alpha, \beta, \delta)$ for all α, β and all $\delta \leq 1$, and the set is increasing in δ: if $(x, y) \in S(\alpha, \beta, \delta)$ and $\delta' > \delta$ then $(x, y) \in S(\alpha, \beta, \delta')$.[12] In order to apply our evolutionary stability criterion, we need to assign expected material utilities, $V(\alpha, \beta, \delta)$ and $V(\beta, \alpha, \delta)$ to the two players when playing this repeated game with discount factor $\delta \in (0, 1)$. In the absence of a selection theory for outcomes in repeated games, three approaches comes to mind.

An *agnostic* or *Laplacian* approach is to assign a uniform probability distribution over all action pairs in $S(\alpha, \beta, \delta)$. Thus $V(\alpha, \beta, \delta)$ and $V(\beta, \alpha, \delta)$ are the associated expected values of $\pi(x, y)$ and $\pi(y, x)$. This is a mathematically well-defined, but analytically cumbersome, approach. An *optimistic* or *Panglossian* approach would be to instead assign unit probability to part (or all) of the Pareto frontier (as defined in terms of the players' altruistic preferences) of the set $S(\alpha, \beta, \delta)$. A prominent selection from this set is the Nash bargaining solution, relative to the stage-game Nash equilibrium,

$$(\hat{x}, \hat{y}) \in \arg \max_{(x,y) \in S(\alpha,\beta,\delta)} \quad [u(x, y, \alpha) - u^{*}(\alpha, \beta)] \cdot [u(y, x, \beta) - u^{*}(\beta, \alpha)] \tag{21}$$

In this case, it is as if the two individuals, before playing the repeated game, bargain over what sustainable action pair to play in all future periods, with the stage-game Nash equilibrium as the default option (that is, if no agreement is reached). Let $V(\alpha, \beta, \delta)$ be the associated material payoff to the α-altruist.

We note that $V(\alpha, \beta, 0) = V(\alpha, \beta)$ is the material payoff to the α-altruist when both parties are "infinitely impatient," $\delta = 0$. This is the stage-game Nash equilibrium payoff to the α-altruist. At the opposite extreme, $V(\alpha, \beta, 1)$ is the material payoff to the α-altruist in the limit when both parties are "infinitely patient," $\delta \to 1$. This is the same as the material payoff obtained by the α-altruist in the "unconstrained" Nash bargaining solution,

$$(\hat{x}, \hat{y}) \in \arg \max_{x,y \geq 0} \quad [u(x, y, \alpha) - u^{*}(\alpha, \beta)] \cdot [u(y, x, \beta) - u^{*}(\beta, \alpha)] \tag{22}$$

[12] To see this, note that the left-hand side in the first row of (20) is the convex combination of $u^{*}(\alpha, \beta)$ and a number, $u^{+}(\alpha, y)$, that is no smaller than the right-hand side. As δ increases, the weight placed on $u^{+}(\alpha, y)$ diminishes, and likewise in the second row.

In the symmetric case, $\alpha = \beta$, the unconstrained Nash bargaining solution is the action x that maximizes $u(x, x, \alpha) = (1 + \alpha)\pi(x, x)$, i.e., it maximizes $\pi(x, x)$. Hence, for all δ close enough to 1,

$$V(\alpha, \alpha, \delta) = V(\alpha, \alpha, 1) = V(0, 0, 1)$$

for all $\alpha \in (-1, 1)$.

More generally, in the $\boldsymbol{\lambda}$-*optimistic* approach, and for arbitrary $\delta \in (0, 1)$, one assigns probability $\lambda \in [0, 1]$ to the Nash bargaining solution, $(\hat{x}(\alpha, \beta, \delta), \hat{x}(\beta, \alpha, \delta))$, and probability $1-\lambda$ to the stage-game Nash equilibrium, $(x^*(\alpha, \beta), x^*(\beta, \alpha))$. In this approach, it is as if the two individuals bargain only with probability λ. The optimistic or Panglossian approach is the special case $\lambda = 1$ while $\lambda = 0$ may be called the *pessimistic* approach. The case $\lambda = 1/2$ defines an intermediate approach between the optimistic and pessimistic ones, and is close to the agnostic approach — being an average of a favorable and an unfavorable outcome in the sustainable set. For $\delta < 1$ close to one, this approach is amenable to analysis.

Let us elaborate somewhat on the $\boldsymbol{\lambda}$-optimistic approach, for arbitrary $\lambda \in (0, 1)$. Under this approach, we will apply our evolutionary stability criterion to the value function

$$V^{\lambda}(\alpha, \beta, \delta) = \lambda V(\alpha, \beta, \delta) + (1 - \lambda) V(\alpha, \beta).$$

For a given index of assortativity, $\sigma \in [0, 1]$, the condition for a degree of altruism α to be evolutionarily stable against a degree $\beta \neq \alpha$ writes

$$\begin{aligned} \lambda V(\alpha, \alpha, \delta) + (1 - \lambda) V(\alpha, \alpha) \;>\; & (1 - \sigma)\left[\lambda V(\beta, \alpha, \delta) + (1 - \lambda) V(\beta, \alpha)\right] \\ & + \sigma\left[\lambda V(\beta, \beta, \delta) + (1 - \lambda) V(\beta, \beta)\right] \end{aligned} \tag{23}$$

As in the analysis of the stage game, a degree of altruism α is locally evolutionarily stable if and only if the right-hand side of (23) has a strict local maximum at $\beta = \alpha$**.** Using subscripts for partial derivatives, the drift function is defined by

$$\begin{aligned} D(\alpha, \beta) &= (1 - \sigma) V_1^{\lambda}(\beta, \alpha, \delta) + \sigma\left[V_1^{\lambda}(\beta, \beta, \delta) + V_2^{\lambda}(\beta, \beta, \delta)\right] \\ &= (1 - \sigma)\left[\lambda V_1(\beta, \alpha, \delta) + (1 - \lambda) V_1(\beta, \alpha)\right] \\ &\quad + \sigma\left[\lambda V_1(\beta, \beta, \delta) + (1 - \lambda) V_1(\beta, \beta)\right] \\ &\quad + \sigma\left[\lambda V_2(\beta, \beta, \delta) + (1 - \lambda) V_2(\beta, \beta)\right], \end{aligned} \tag{24}$$

which, evaluated at the incumbent degree of altruism α, is

$$\begin{aligned} D(\alpha, \alpha) &= (1 - \lambda)\left[V_1(\alpha, \alpha) + \sigma V_2(\alpha, \alpha)\right] \\ &\quad + \lambda\left[V_1(\alpha, \alpha, \delta) + \sigma V_2(\alpha, \alpha, \delta)\right]. \end{aligned}$$

We are already familiar with the expression in the first square bracket; this is the drift function for the stage game. When individuals are very impatient, that is, when δ is close to zero, then $V(\alpha, \alpha, \delta)$ is close to $V(\alpha, \alpha)$, and hence the evolutionarily stable degree of altruism is close to that for the stage game. By contrast, when individuals are very patient, that is, when δ is sufficiently close to one, $V(\alpha, \alpha, \delta)$ not only is close to $V(\alpha, \alpha, 1)$; it is actually identical with $V(\alpha, \alpha, 1)$, and $V(\alpha, \alpha, 1)$, as noted above, is independent of the common degree α of altruism. Formally:

Proposition 10 *There exists a* $\bar{\delta} < 1$ *such that*

$$D(\alpha, \alpha) = \frac{1}{2}\left[V_1(\alpha, \alpha) + \sigma V_2(\alpha, \alpha)\right] + \frac{1-\sigma}{2} V_1(\alpha, \alpha, 1)$$

for all $\delta \in \left(\bar{\delta}, 1\right)$ *and all* $\alpha \in (-1, 1)$.

In other words: when individuals are equally altruistic and sufficiently patient, they will both take the efficient action, and hence a decrease in both players' degrees of altruism has no effect: the term $V(\beta, \beta, \delta)$ in (23) is a constant. Hence, the possibility of reaching an agreement to play the Nash bargaining solution affects the evolutionarily stable degree of altruism solely through the term $V(\beta, \alpha, \delta)$. Hence, if a decrease in own altruism raises own material payoff (if $V_1(\alpha, \alpha, 1) < 0$), then drift is smaller, and consequently the stable degree of altruism will be smaller in the repeated game than in the stage game.

5.2 Endogenous threats

Abreu and Pearce (2002, 2007) provide two distinct motivations for the Nash bargaining solution under endogenous threats (Nash, 1953) as the relevant prediction of infinitely repeated play when $\delta \to 1$. We here investigate what this selection gives in our evolutionary framework. First, for arbitrary $x_0, y_0 \geq 0$, let

$$(\hat{x}, \hat{y}) \in \arg\max_{x,y \geq 0} \quad [u(x, y, \alpha) - u(x_0, y_0, \alpha)] \cdot [u(y, x, \beta) - u(y_0, x_0, \beta)]$$

This defines $\hat{x}$ as a function φ of $(x_0, y_0, \alpha, \beta)$, and we have $\hat{y} = (y_0, x_0, \beta, \alpha)$. Consider now the simultaneous-move game which the α-altruist chooses $x_0 \geq 0$, the β-altruist chooses $y_0 \geq 0$ and the payoff to the first is

$$v(x_0, y_0, \alpha, \beta) = u\left[\varphi(x_0, y_0, \alpha, \beta), \varphi(y_0, x_0, \beta, \alpha), \alpha\right]$$

and to the second it is

$$v\left(y_0, x_0, \beta, \alpha\right) = u\left[\varphi\left(y_0, x_0, \beta, \alpha\right), \varphi\left(x_0, y_0, \alpha, \beta\right), \beta\right].$$

The Nash bargaining with endogenous threat is the (by hypothesis unique) Nash equilibrium of this two-player game.

We note that, given (x_0, y_0), the first-order conditions for $(\hat{x}, \hat{y})$ are

$$\begin{cases} u_1\left(x, y, \alpha\right) \cdot \left[u\left(y, x, \beta\right) - u\left(y_0, x_0, \beta\right)\right] + u_2\left(y, x, \beta\right) \cdot \left[u\left(x, y, \alpha\right) - u\left(x_0, y_0, \alpha\right)\right] = 0 \\ u_2\left(x, y, \alpha\right) \cdot \left[u\left(y, x, \beta\right) - u\left(y_0, x_0, \beta\right)\right] + u_1\left(y, x, \beta\right) \cdot \left[u\left(x, y, \alpha\right) - u\left(x_0, y_0, \alpha\right)\right] = 0. \end{cases}$$

In the symmetric case, when $\alpha = \beta$, and $x_0 = y_0$, the Nash bargaining solution is efficient in terms of the players' altruistic preferences (and hence also in terms of their material preferences):

$$u_1\left(x, x, \alpha\right) + u_2\left(x, x, \alpha\right) = 0.$$

Unfortunately, this approach seems to be algebraically so cumbersome that we have to leave its exploration for future studies.

6 Other social preferences

6.1 Inequity aversion

Here we generalize the analytical machinery developed above to more general social preferences than true altruism. Imagine, thus, that individuals both care about their own material utility and also about "welfare," "equity" or "fairness" of the resulting allocation, so that the total utility is of the form

$$u\left(x, y, \gamma\right) = \pi\left(x, y\right) + \gamma \cdot \phi\left(x, y\right)$$

for some function ϕ and $\gamma \in \mathbb{R}$. In this general setting, γ represents the *intensity* of the player's social preference. If π and ϕ are such that the resulting simultaneous-move game, with payoff functions $u\left(x, y, \gamma\right)$ and $u\left(y, x, \delta\right)$, for some and $\gamma, \delta \in \mathbb{R}$, has a unique, interior and regular Nash equilibrium, then one can, just as in the case of pure altruism studied above, define $V\left(\gamma, \delta\right)$ as the material equilibrium payoff to an γ-player when playing a δ-player. One can then use the drift function to study evolutionarily stable intensities of diverse forms of social preferences.

Clearly, if welfare is represented as the sum of the two players' material payoffs,

$$\phi(x,y) = \pi(x,y) + \pi(y,x),$$

then we have already studied this class of social preferences, since then we have

$$u(x,y,\alpha) \equiv (1-\gamma)\,\pi(x,y) + \gamma\pi(y,x)$$

so for $\gamma \neq 1$ this is equivalent with true altruism of degree $\alpha = \gamma/(1-\gamma)$. This is a variation of the model in Charness and Rabin (2002).

By contrast, social preferences in terms of "equity" or "fairness" cannot be represented directly in terms of altruism. We proceed to briefly analyze stable sharing rules, when individuals are not motivated by altruism but by inequity aversion of the form studied in Fehr and Schmidt (1999).

Consider two individuals who have different amounts of initial wealth, and who may transfer any part of their own wealth to the other individual. With initial wealth levels w_A and w_B, let x denote the amount transferred from A to B, y the amount transferred from B to A, and let $\gamma \in \mathbb{R}$ and $\delta \in \mathbb{R}$ be A's and B's degrees of inequity aversion, respectively, formalized as follows:

$$\begin{cases} u(x,y,\gamma) = \pi(w_A - x + y) - \gamma \cdot \max\{0, w_A - w_B - 2(x-y)\} \\ u(y,x,\delta) = \pi(w_B + x - y) - \delta \cdot \max\{0, w_B - w_A - 2(y-x)\} \end{cases}$$

In other words, each individual gets some disutility from the other's relative poverty.[13] We assume that with probability 1/2 individual A's initial wealth is w^H and B's is w^L, and with probability 1/2 A's initial wealth is w^L and B's w^H. In the event when A is rich, A will give the minimal transfer $x \geq 0$ that satisfies

$$\pi'\left(w^H - x\right) \geq 2\gamma. \tag{25}$$

Likewise, in the event when B is rich, B will give the minimal transfer $y \geq 0$ that satisfies

$$\pi'\left(w^H - y\right) \geq 2\delta.$$

This determines uniquely two transfers, which we denote $\tilde{\tau}(\gamma)$ and $\tilde{\tau}(\delta)$, respectively. The so-defined function $\tilde{\tau} : \mathbb{R}_+ \to \left[0, w^H\right)$ is continuous and non-decreasing. Moreover, there

[13] Fehr and Schmidt (1999) also allow for inequity aversion in the opposite direction. However, since we only allow for non-negative transfers, such aversion is here immaterial.

exists a $\gamma_0 \in (0,1)$, such that $\tilde{\tau}(\gamma) = 0$ for all $\gamma \leq \gamma_0$, while $\tilde{\tau}(\gamma)$ is positive and satisfies (25) with equality for all $\gamma > \gamma_0$. Indeed, $\tilde{\tau}$ is differentiable on $(\gamma_0, +\infty)$ with

$$\tilde{\tau}'(\gamma) = -\frac{2}{\pi''\left[w^H - \tilde{\tau}(\gamma)\right]} > 0.$$

The resulting expected utility to the γ-individual is

$$V(\gamma, \delta) = \frac{1}{2}\pi\left[w^H - \tilde{\tau}(\gamma)\right] + \frac{1}{2}\pi\left[w^L + \tilde{\tau}(\delta)\right].$$

Let γ denote the incumbent degree of inequity aversion. For any index of assortativity $\sigma \in [0,1]$, the corresponding drift function, evaluated at γ**,** thus satisfies

$$D(\gamma, \gamma) = \begin{cases} 0 & \forall\ \gamma < \gamma_0 \\ \left(\sigma \cdot \pi'\left[w^L + \tilde{\tau}(\gamma)\right] - \pi'\left[w^H - \tilde{\tau}(\gamma)\right]\right) \cdot \tilde{\tau}'(\gamma) & \forall\ \gamma > \gamma_0 \end{cases} \tag{26}$$

In particular, no degree of inequity aversion below $\gamma_0 > 0$ is locally evolutionarily stable. For degrees of inequity aversion $\gamma > \gamma_0$ we have $\pi'\left[w^H - \tilde{\tau}(\gamma)\right] = 2\gamma$, so a necessary condition for local evolutionary stability of such a degree of inequity aversion is

$$\sigma \cdot \pi'\left[w^L + \tilde{\tau}(\gamma)\right] = 2\gamma$$

where

$$\tilde{\tau}(\gamma) = w^H - (\pi')^{-1}(2\gamma).$$

Hence, an evolutionarily stable value of γ has to satisfy the fixed-point equation

$$\gamma = \frac{\sigma}{2} \cdot \pi'\left[w^L + w^H - (\pi')^{-1}(2\gamma)\right]$$

In the special case of logarithmic consumption utility, $\pi(c) \equiv \ln c$:

$$\gamma = \frac{\sigma}{2(w^L + w^H) - 1/\gamma}.$$

The unique solution to this equation is

$$\gamma^* = \frac{1}{2} \cdot \frac{1+\sigma}{w^L + w^H} > 0. \tag{27}$$

In this special case, we have $\gamma_0 = 1/\left(2w^H\right)$, so

Proposition 11 *For logarithmic consumption utility, no degree of inequity aversion is locally evolutionarily stable if $\sigma w^H \leq w^L$. If $\sigma w^H > w^L$, then (27) is the unique locally evolutionarily stable degree of inequity aversion.*

We note that, for $\sigma w^H > w^L$, the stable degree of inequity aversion is positive and increasing in the degree of assortativity (in line with our finding that the stable degree of altruism is increasing in the degree of assortativity). In particular, under uniform random matching ($\sigma = 0$), we predict selfish behavior.

We conclude this analysis by comparing the equilibrium transfer, at the stable degree of inequity aversion, with the equilibrium transfer at the stable degree of altruism in this sharing game. For this comparison, assume that $\sigma w^H > w^L$. At the evolutionarily stable degree of inequity aversion, the equilibrium transfer from the rich to the poor is

$$\tilde{\tau}(\gamma^*) = w^H - \frac{w^L + w^H}{1+\sigma} = \frac{\sigma w^H - w^L}{1+\sigma}.$$

We now derive stable degrees of pure altruism in this sharing game. Let

$$\begin{cases} u(x, y, \alpha) = \pi(w_A - x + y) + \alpha \cdot \pi(w_B + x - y) \\ u(y, x, \beta) = \pi(w_B + x - y) + \beta \cdot \pi(w_A - x + y), \end{cases}$$

where $\alpha \in (-1, 1)$ is A's degree of altruism towards B, and $\beta \in (-1, 1)$ B's degree of altruism towards A. In the event when A is rich, B will give nothing to A, but A will give the minimal transfer $x \geq 0$ that satisfies

$$\alpha \cdot \pi'\left(w^L + x\right) \leq \pi'\left(w^H - x\right). \tag{28}$$

Likewise, in the event when B is rich, A will give nothing to B, but B will give the minimal transfer $y \geq 0$ that satisfies

$$\beta \cdot \pi'\left(w^L + y\right) \leq \pi'\left(w^H - y\right).$$

This determines uniquely two transfers, which we denote $x^{NE}(\alpha, \beta) = \tau(\alpha)$ and $y = x^{NE}(\beta, \alpha) = \tau(\beta)$, respectively. We note that the so-defined function $\tau : (-1, 1) \to \left[0, w^H\right)$ is continuous and non-decreasing. Let $\alpha_0 = \pi'\left(w^H\right)/\pi'\left(w^L\right)$, and note that $\alpha_0 \in (0, 1)$. It follows that $\tau(\alpha) = 0$ for all $\alpha \leq \alpha_0$, while $\tau(\alpha)$ is positive and satisfies (28) with equality for all $\alpha \in (\alpha_0, 1)$. Indeed, τ is differentiable on $(\alpha_0, 1)$, and we have

$$\tau'(\alpha) = \frac{\pi'\left[w^H - \tau(\alpha)\right]}{\pi''\left[w^H - \tau(\alpha)\right] - \alpha\pi''\left[w^L + \tau(\alpha)\right]} > 0.$$

Conditional upon the union of the two events, the resulting expected personal fitness to the α-altruist is

$$V(\alpha, \beta) = \frac{1}{2}\pi\left[w^H - \tau(\alpha)\right] + \frac{1}{2}\pi\left[w^L + \tau(\beta)\right].$$

For any $\sigma \in [0,1]$,

$$D(\alpha,\alpha) = \begin{cases} 0 & \forall\ \alpha < \alpha_0 \\ \left(\sigma \cdot \pi'\left[w^L + \tau(\alpha)\right] - \pi'\left[w^H - \tau(\alpha)\right]\right) \cdot \tau'(\alpha) & \forall\ \alpha > \alpha_0. \end{cases} \tag{29}$$

Then, from Proposition 1:

Proposition 12 *No degree of altruism is locally evolutionarily stable if $\sigma \leq \alpha_0$. For any $\sigma \in (\alpha_0, 1)$, the only degree of altruism that may be locally evolutionarily stable is $\alpha = \sigma$.*

Proof: Since $\tau'(\alpha) = 0$ for all $\alpha < \alpha_0$, also $D(\alpha,\alpha) = 0$ for all $\alpha < \alpha_0$, and thus such α-values fail conditions (ii) and (iii) in Proposition 1. By contrast, for all $\alpha > \alpha_0$: $\tau'(\alpha) > 0$ and $D(\alpha,\alpha) = 0$ iff $\sigma \cdot \pi'\left[w^L + \tau(\alpha)\right] = \pi'\left[w^H - \tau(\alpha)\right]$. Hence, $\alpha = \sigma$ meets condition (i), and no other α does. Moreover, from the assumed concavity of π, conditions (ii) and (iii) in Proposition 1 are easily verified to hold at $\alpha = \sigma$. **End of proof**.

Under logarithmic consumption utility, $\pi(c) \equiv \ln c$, a positive transfer from a rich α-altruist satisfies

$$\frac{\alpha}{w^L + \tau(\alpha)} = \frac{1}{w^H - \tau(\alpha)}$$

or

$$\tau(\alpha) = \max\left\{0, \frac{\alpha w^H - w^L}{1+\alpha}\right\}.$$

It is easily verified that $\tau(\alpha^*) = \tilde{\tau}(\gamma^*)$. Evolutionarily stable inequity aversion results in the same *behavior* as evolutionarily stable altruism!

Indeed, the observation that the evolutionarily stable behavior in the two distinct formalizations of social preferences can be seen to hold generally: the condition for stable altruism (29) boils down to $\sigma \cdot \pi'\left[w^L + \tau(\alpha^*)\right] = \pi'\left[w^H - \tau(\alpha^*)\right]$ for $\alpha^* > \alpha_0$, and the condition for stable inequity aversion (26) is $\sigma \cdot \pi'\left[w^L + \tilde{\tau}(\gamma^*)\right] = \pi'\left[w^H - \tilde{\tau}(\gamma^*)\right]$ for $\gamma^* > \gamma_0$. Hence, $\tau(\alpha^*) = \tilde{\tau}(\gamma^*)$.

6.2 Reciprocal altruism

In another strand of the economics literature, the focus is on conditional altruism, that is, altruism that in part depends on the other individual's "type" (or on the first individual's belief about this type). Levine (1998) postulates and analyzes a form of such conditional altruism. Sethi and Somanathan (2001) analyze different forms of "reciprocal altruism."

They find that, under uniform random matching, "reciprocators," i.e., individuals who are altruistic towards other reciprocators but spiteful against selfish individuals, can invade a population of selfish individuals. They also show that a monomorphic population of reciprocators would resist the invasion by selfish individuals. However, Sethi and Somanathan (2001) do not determine evolutionarily stable preferences. Here, we explore how our approach may be adapted to determine evolutionarily stable preferences in a simple example of their model.

Consider a two-player normal form game, with strategy spaces and material payoffs as in our baseline model. Each individual's preferences are defined by two parameters: one *altruism* parameter, $\alpha \in (-1, 1)$, and one *sensitivity* parameter, $\lambda \geq 0$, that represents the extent to which the individual cares about the opponent's degree of altruism. Letting (α, λ) be the "type" of the individual undertaking action x, and (β, μ) that of the individual undertaking action y, their preferences are represented by the utility functions

$$u(x, y, \alpha, \lambda, \beta) = \pi(x, y) + \frac{\alpha + \lambda(\beta - \alpha)}{1 + \lambda} \cdot \pi(y, x),$$

and

$$u(y, x, \beta, \mu, \alpha) = \pi(y, x) + \frac{\beta + \mu(\alpha - \beta)}{1 + \mu} \cdot \pi(x, y),$$

respectively.

A necessary first-order condition for a pair $(x, y) > 0$ to be a Nash equilibrium is

$$\begin{cases} \pi_1(x, y) + \frac{\alpha+\lambda(\beta-\alpha)}{1+\lambda} \cdot \pi_2(y, x) = 0 \\ \pi_1(y, x) + \frac{\beta+\mu(\alpha-\beta)}{1+\mu} \cdot \pi_2(x, y) = 0. \end{cases}$$

Focusing on material payoff functions π and parameters such that Nash equilibrium is unique and interior, let $x^*(\alpha, \lambda, \beta, \mu)$ denote the Nash equilibrium action of the (α, λ)-individual and let $V(\alpha, \lambda, \beta, \mu)$ be the individual's material payoff in equilibrium. An index $k = 1, 2, 3, 4$ will denote the partial derivative of V with respect to its k-th argument.

A sufficient condition for a pair of preference parameters (α, λ) to be evolutionarily stable against a pair $(\beta, \mu) \neq (\alpha, \lambda)$ is that

$$V(\alpha, \lambda, \alpha, \lambda) > (1 - \sigma) V(\beta, \mu, \alpha, \lambda) + \sigma V(\beta, \mu, \beta, \mu).$$

Given the degrees of altruism, α, and of sensitivity, λ, and the index of assortativity, σ, the right-hand side can be viewed as a function of β and μ. At $(\beta, \mu) = (\alpha, \lambda)$, this function obtains the value $V(\alpha, \lambda, \alpha, \lambda)$. Hence, the pair of preference parameters (α, λ) is locally

evolutionarily stable if and only if the right-hand side has a strict local maximum at $(\beta, \mu) = (\alpha, \lambda)$, and drift now occurs along two dimensions, so that:

$$D_\alpha(\alpha, \lambda, \alpha, \lambda) = V_1(\alpha, \lambda, \alpha, \lambda) + \sigma V_3(\alpha, \lambda, \alpha, \lambda)$$

$$D_\lambda(\alpha, \lambda, \alpha, \lambda) = V_2(\alpha, \lambda, \alpha, \lambda) + \sigma + V_4(\alpha, \lambda, \alpha, \lambda),$$

A necessary condition for (α, λ) to be locally evolutionarily stable is

$$\begin{cases} D_\alpha(\alpha, \lambda, \alpha, \lambda) = 0 \\ D_\lambda(\alpha, \lambda, \alpha, \lambda) = 0. \end{cases}$$

In a similar fashion as in the model with pure altruism, it can be shown that this condition boils down to

$$\begin{cases} \left[\sigma - \frac{\alpha}{1+\lambda}\right] \cdot x_1^*(\alpha, \lambda, \alpha, \lambda) + \left[1 - \frac{\alpha}{1+\lambda} \cdot \sigma\right] \cdot x_3^*(\alpha, \lambda, \alpha, \lambda) = 0 \\ \left[\sigma - \frac{\alpha}{1+\lambda}\right] \cdot x_2^*(\alpha, \lambda, \alpha, \lambda) + \left[1 - \frac{\alpha}{1+\lambda} \cdot \sigma\right] \cdot x_4^*(\alpha, \lambda, \alpha, \lambda) = 0. \end{cases}$$

Applying this to interactions with the material payoff function

$$\pi(x, y) = (x + y)^\tau - \frac{1}{2}x^2$$

we obtain:

Proposition 13 *In the two-player public-goods game with additive efforts, and for any $\sigma \in [0, 1)$, if (α, λ) is an evolutionarily stable reciprocal type, then $(\alpha, \lambda) = \left(\frac{13-5\sigma}{15-8\sigma+\sigma^2}, \frac{3(1-\sigma)}{2(5-\sigma)}\right)$.*

7 Matching processes

We here discuss a few alternative matching processes, biological and socio-economic, and point at some potential further applications of the present analytical machinery.

7.1 Biological processes

The first interpretation is biological: that altruism is transmitted genetically from one generation to the next, and that the n group members are relatives, with common ancestors, and equal degree of relatedness between each pair of players. For instance, in the case of siblings, if parents match randomly (independently of their degrees of sibling altruism), each

couple gets exactly n children, and each child is equally likely to inherit each parent's degree of altruism, then $\sigma_\varepsilon \to \sigma_0 = 1/2$. To see this, consider a population where a proportion ε of the adult population carries degree of altruism β, and a proportion carries degree of altruism α. An α-altruistic child has at least one α-altruistic parent. With probability $1-\varepsilon$ the other parent is also α-altruistic, in which case each of the child's siblings must be α-altruistic. If the other parent is β-altruistic, which happens with probability ε, any of the child's siblings is an α-altruist with probability $1/2$, and a β-altruist with probability $1/2$. Hence, the probability that an α-altruist's sibling also is an α-altruist is $1-\varepsilon/2$. Similarly, a β-altruistic child has at least one β-altruistic parent. With probability $1-\varepsilon$ the other parent is α-altruistic, in which case the probability that a sibling is α-altruistic is $1/2$. Hence, the probability that a β-altruist's sibling is a α-altruist is $(1-\varepsilon)/2$. Thus, the index of assortativity is $\sigma_\varepsilon \to \sigma_0 = 1/2$, the coefficient of relationship between siblings (Wright, 1922). Note that, in this case, the index of assortativity does not depend on the share ε of β-altruists.

Our analysis does not require monogamous relationships. Consider interactions between two siblings. Assume first that the two siblings always have different fathers**.** Let α be the incumbent degree of half-sibling altruism and let β be a mutant degree, appearing in population share ε. Then the probability that α-altruistic child's mother is a α-altruist is $1/(1+\varepsilon)$ and the probability that the mother instead is a β-altruist is $\varepsilon/(1+\varepsilon)$. Conditional on the mother being an α-altruist, with probability $1-\varepsilon$ the father of the half-sibling is also an α-altruist, in which case the half-sibling must be an α-altruist. With probability ε the father of the half-sibling is instead a β-altruist, in which case the half-sibling is an α-altruist with probability $1/2$. Conditional on the child's mother being a β-altruist, the half-sibling can be an α-altruist only if his or her father is a α-altruist, which happens with probability $1-\varepsilon$, in which case the half-sibling is an α-altruist with probability $1/2$. Thus, the probability that an α-altruistic child's half-sibling is an α-altruist as well is:

$$\Pr[\alpha|\alpha,\varepsilon] = \frac{1}{1+\varepsilon}\left[(1-\varepsilon)+\frac{1}{2}\varepsilon\right]+\frac{1}{2}\cdot\frac{\varepsilon}{1+\varepsilon}(1-\varepsilon) = \frac{1-\varepsilon^2/2}{1+\varepsilon}.$$

Consider now a β-altruistic child. This child must have at least one β-altruistic parent. The probability that the mother is an α-altruist is $(1-\varepsilon)/(2-\varepsilon)$ and the probability that she is a β-altruist is $1/(2-\varepsilon)$. Thus, the probability that a β-altruistic child's half-sibling is an α-altruist is:

$$\Pr[\alpha|\beta,\varepsilon] = \frac{1-\varepsilon}{2-\varepsilon}\left[(1-\varepsilon)+\frac{1}{2}\varepsilon\right]+\frac{1}{2}\cdot\frac{1}{2-\varepsilon}(1-\varepsilon) = \frac{(1-\varepsilon)(3-\varepsilon)}{2(2-\varepsilon)}.$$

In the limit as ε tends to zero:

$$\sigma = \Pr[\alpha|\alpha,\varepsilon] - \Pr[\alpha|\beta,\varepsilon] \to 1 - \frac{3}{4} = \frac{1}{4}.$$

More generally, if a fraction $\kappa \in [0,1]$ of mothers have children with two different fathers, and the remaining fraction have children with the same father, then $\sigma = (1-\kappa)/2 + \kappa/4$.

Similarly, assume that a fraction η of couples divorce after their first child, and rematch before they have their second child, but otherwise things are as described above, then we have $\sigma = \frac{1}{2}(1-\eta) + \frac{1}{4}\eta$. Draw a child at random in this population; with probability $(1-\eta)$ this child's parents did not divorce, in which case the difference in the conditional probability that a α-altruist meets a α-altruist, and the conditional probability that a β-altruist meets a α-altruist, is $1/2$, as in the analysis above; with probability η, this child's parents did divorce, in which case the difference in the conditional probability that a α-altruist meets a α-altruist, and the conditional probability that a β-altruist meets a α-altruist, is $1/4$ (with probability $1/2$ the child remained with the parent from whom he inherited his degree of altruism, in which case the likelihood that the half-sibling inherits the same altruism is $1/2$). Our analysis shows how such alternative social patterns lead to different degrees of stable sibling altruism.

Our model also allows for assortative mating among parents. Suppose that with probability $1-\mu \in [0,1]$ a given mutant grown-up has a random match, as in the example above, but that with probability μ the mutant settles only for a match with another mutant. Then, $\sigma = (1-\mu)/2 + \mu$.

7.2 Socio-economic processes

The second interpretation is socioeconomic, namely, that altruism is transmitted by way of imitation or education from one generation to the next. Then our model applies to any situation where individuals interact according to the general class of games described in Section 2.

It should be pointed out that, while it is natural to think of altruism as reflecting true care for another individual in the case of relatives, an individual's degree of altruism may simply be interpreted as the extent to which the individual takes into account the effect of his or her actions on the others. Hence, for interactions between non-relatives, an arguably reasonable interpretation of α is the degree of team spirit.

Examples abound. For instance, suppose the n group members are employees in a firm. Assume that each employee acquired a certain degree of team spirit in school (and that all alumni from any given school acquire the same team spirit). Employees live for one period, and every period each firm hires n new employees. If each firm has a favorite school from which it hires a fraction ζ of the n employees, while the others are recruited randomly from all the other schools, the index of assortativity is $\sigma = \zeta$.

More generally, suppose the n group members are partners in a business or members of a team or a club. Members are recruited either completely at random, in which case the index of assortativity is zero, $\sigma = 0$, or else there is some heterogeneity across groups in terms of their recruitment process. A positive index of assortativity means that a group that has recruited one mutant member is more likely to also have recruited another mutant member, as compared with a group that has not recruited any mutant member. Such deviations from pure random recruitment is arguably very plausible in most partnerships, teams or clubs. For instance, the stability of a certain degree of team spirit may depend on the socioeconomic environment and recruitment methods. Our analysis shows how the recruitment process affects the stable degree of altruism within each group.

Consider now a pre-historic society, where individuals live in villages, and assume that each village has a hunting team consisting of n members. Suppose that each man acquires team spirit from the elder in his native village. If, prior to hunting team formation, a fraction ξ of all young men migrate from their native village to a randomly chosen village, while the others remain in their native village, the index of assortativity is $\sigma = 1 - \xi$. In traditional societies, such migration was often linked to marriage, and typically tradition would dictate whether it is the man or the woman who migrates. Thus, our theory suggests that differences in the nature of interactions among females and among males, as well as in male vs. female migration patterns, may have affected the evolution of altruism.

Assortative matching may also arise if players can choose their opponents (Wilson and Dugatkin, 1997, Bergstrom, 2003). Suppose that individuals are randomly matched into groups of N individuals (where N is even), and that each individual's degree of altruism is public information within the group. Suppose further that pairs form within each group (that is, the strategic interaction takes place in groups of size $n = 2$), and that each individual's partner choice is voluntary. If there are two degrees of altruism in the population, α and $\beta > \alpha$, then every individual would prefer a β-individual as a partner. Hence, β-individuals will be paired together whenever there is more than one β-altruist in a group. Here, the

index of assortativity depends on the fraction ε of β-altruists, and on the group size N.

It could also be that other aspects of preferences affect the probability of interacting assortatively (Eshel and Cavalli-Sforza, 1982). For instance, suppose that the habitat preferences of mutants differ from those of incumbents. Then, mutants end up being more likely to live in the same area and interact with each other, than if their habitat preferences coincided with those of the incumbents.

Finally, note that this broader interpretation may also be applied to interactions between kin, should altruism be culturally rather genetically transmitted between generations. In such a setting it is natural to think that children are not only influenced by their parents. Our model easily encompasses such a possibility. Consider again an interaction between siblings. Suppose that parents mate randomly, that with probability $1-\psi$ a child inherits one of its parents' degree of family altruism as in the baseline example above, but that with probability $\psi \in [0,1]$, the child imitates a randomly drawn grown-up from the population at large. Then $\sigma = (1-\psi)/2$.

8 Conclusion

In this paper we have built a bridge between the economics literature on the evolution of preferences, and the biology literature on the evolution of altruistic behavior in populations with assortative matching. We have proposed a definition of evolutionary stability for altruistic preferences in a population where individuals are matched, assortatively or not, to play symmetric games with perfect information. Hence, individuals adapt their behavior to their preferences and to the situation at hand, including the preferences of their opponent. We have also provided an operational method to calculate locally stable altruistic preferences in this context; the drift function.

Our results suggest that evolution selects for altruistic preferences that depend on the specifics of the strategic interaction, and on the extent of assortative matching. We determine the stable degree of altruism in a variety of games, and we obtain empirically testable predictions. For populations where individuals tend to interact in games in which their behaviors are strategic substitutes, the stable degree of altruism falls short of the index of assortativity σ (which is equal to the coefficient of relatedness r when individuals interact with kin). By contrast, for populations where individuals tend to interact in games in which their behaviors are strategic complements, the stable degree of altruism exceeds this coeffi-

cient. In all types of games, however, the index of assortativity impacts the stable degree of altruism positively.

The present model relies on several restrictive assumptions. We conjecture, however, that, it should be fairly straightforward to relax some of these assumptions. The key requirement for our approach to work is a value function that maps the players' preference parameters to expected material utility outcomes. It should be feasible to fulfill this requirement in a reasonable manner in situations where individuals observe their opponents' preferences with some noise, or with some probability only. In the latter case one could assume that when players do not observe each others' preferences, they attach the population average value to the preference parameter of their opponent(s). Other kinds of incomplete (but symmetric) information, for instance about the payoffs, may also be encompassed by incorporating the relevant expected values. For games with asymmetric information, one could, to begin with, impose some exogenously given contracting or signaling framework, and thus obtain a value function mapping preference parameters to expected material utility outcomes. Recall, however, that our evolutionary approach relies on symmetric games. To allow for games with asymmetric information, one could circumvent the challenge of the asymmetry in the strategic interaction by assuming that, ex ante, each individual is equally likely to end up being the player with private information (like in our simple sharing game, where each individual is equally likely to be the richest or the poorest from an ex ante perspective).

It will likely be much more challenging to analyze situations that are inherently asymmetric, such as interactions between parents and offspring, between males and females, or between the leaders and the subordinates in a group.

9 Appendix A: Proofs

9.1 Lemma 1

In any game $G = \langle \pi, \alpha, \beta \rangle$ with $\pi \in \Pi$ and $\alpha, \beta \in (-1, 1)$:

$$\begin{aligned} V_1(\beta, \alpha) &= \pi_1 \left[x^*(\beta, \alpha), x^*(\alpha, \beta) \right] \cdot x_1^*(\beta, \alpha) \\ &\quad + \pi_2 \left[x^*(\beta, \alpha), x^*(\alpha, \beta) \right] \cdot x_2^*(\alpha, \beta), \end{aligned} \tag{30}$$

and

$$\begin{aligned} V_2(\beta,\alpha) &= \pi_1\left[x^*(\beta,\alpha), x^*(\alpha,\beta)\right]\cdot x_2^*(\beta,\alpha) \\ &\quad +\pi_2\left[x^*(\beta,\alpha), x^*(\alpha,\beta)\right]\cdot x_1^*(\alpha,\beta). \end{aligned} \tag{31}$$

Using the β-altruist's first-order condition,

$$\pi_1\left[x^*(\beta,\alpha), x^*(\alpha,\beta)\right] + \beta\cdot\pi_2\left[x^*(\alpha,\beta), x^*(\beta,\alpha)\right] = 0, \tag{32}$$

(30) and (31) may be re-written

$$\begin{aligned} V_1(\beta,\alpha) &= -\beta\cdot\pi_2\left[x^*(\alpha,\beta), x^*(\beta,\alpha)\right]\cdot x_1^*(\beta,\alpha) \\ &\quad +\pi_2\left[x^*(\beta,\alpha), x^*(\alpha,\beta)\right]\cdot x_2^*(\alpha,\beta) \end{aligned}$$

and

$$\begin{aligned} V_2(\beta,\alpha) &= -\beta\cdot\pi_2\left[x^*(\alpha,\beta), x^*(\beta,\alpha)\right]\cdot x_2^*(\beta,\alpha) \\ &\quad +\pi_2\left[x^*(\beta,\alpha), x^*(\alpha,\beta)\right]\cdot x_1^*(\alpha,\beta), \end{aligned}$$

so that

$$V_1(\beta,\alpha)_{|\beta=\alpha} = \left[x_2^*(\alpha,\alpha) - \alpha\cdot x_1^*(\alpha,\alpha)\right]\cdot\pi_2\left[x^*(\alpha,\alpha), x^*(\alpha,\alpha)\right] \tag{33}$$

and

$$V_2(\beta,\alpha)_{|\beta=\alpha} = \left[x_1^*(\alpha,\alpha) - \alpha\cdot x_2^*(\alpha,\alpha)\right]\cdot\pi_2\left[x^*(\alpha,\alpha), x^*(\alpha,\alpha)\right] \tag{34}$$

which, together with (12) and Proposition 1, imply (13).

9.2 Lemma 2

Consider a game $G = \langle\pi,\alpha,\beta\rangle$ with $\pi\in\Pi$ and $\alpha,\beta\in(-1,1)$. Applying the Implicit Function Theorem to the equation

$$\pi_1(x,y) + \alpha\cdot\pi_2(y,x) = 0, \tag{35}$$

which implicitly defines an α-altruist's best response x to the other's action/effort y, yields:

$$\frac{dx}{d\alpha} = -\frac{\pi_2(y,x)}{\pi_{11}(x,y) + \alpha\pi_{22}(y,x)}. \tag{36}$$

The denominator is negative under our hypothesis of a unique Nash equilibrium, which requires that (35) uniquely define x, or

$$\pi_{11}(x,y) + \alpha\cdot\pi_{22}(y,x) < 0 \qquad \forall x,y>0. \tag{37}$$

Hence, (36) implies that $dx/d\alpha$ has the same sign as π_2. Moreover, $dy/d\alpha = 0$, since a β-altruist's best response y to the other's action/effort x, is defined by:

$$\pi_1(y, x) + \beta \cdot \pi_2(x, y) = 0.$$

Consider the Nash equilibrium $(x^*(\alpha, \beta), x^*(\beta, \alpha))$. The previous observations imply $x_1^*(\alpha, \beta) > 0 \Leftrightarrow \pi_2 > 0$ and $x_1^*(\alpha, \beta) < 0 \Leftrightarrow \pi_2 < 0$.

Applying again the Implicit Function Theorem to equation (35), we obtain

$$\frac{dx}{dy} = -\frac{\pi_{12}(x, y) + \alpha\pi_{21}(y, x)}{\pi_{11}(x, y) + \alpha\pi_{22}(y, x)}. \tag{38}$$

Since $\pi_{11}(x, y) + \alpha \cdot \pi_{22}(y, x) < 0$, and recalling that $\pi_{12} = \pi_{21}$, this implies:

1. $\pi_{12}(x, y) = 0 \Rightarrow dx/dy = 0 \Rightarrow x_2^*(\beta, \alpha) = 0$.
2. $\pi_{12}(x, y) < 0 \Rightarrow dx/dy < 0$, so that $x_2^*(\beta, \alpha) \neq 0$ and has the opposite sign as $x_1^*(\alpha, \beta)$.
3. $\pi_{12}(x, y) > 0 \Rightarrow dx/dy > 0$, so that $x_2^*(\beta, \alpha) \neq 0$ and has the same sign as $x_1^*(\alpha, \beta)$.

9.3 Proposition 2

Equation (13) and the assumption $\pi_2 \neq 0$ imply

$$D(\alpha, \alpha) = 0 \Leftrightarrow (\alpha - \sigma) \cdot x_1^*(\alpha, \alpha) = (1 - \sigma\alpha) \cdot x_2^*(\alpha, \alpha).$$

Since $1 - \sigma\alpha > 0$, it follows that:

1. when actions are strategically independent (so that $x_2^*(\alpha, \alpha) = 0$ and $x_1^*(\alpha, \alpha) \neq 0$): $D(\alpha, \alpha) = 0 \Leftrightarrow \alpha - \sigma = 0$;
2. when actions are strategic substitutes (so that $x_1^*(\alpha, \alpha) \cdot x_2^*(\alpha, \alpha) < 0$): $D(\alpha, \alpha) = 0 \Leftrightarrow \alpha - \sigma < 0$;
3. when actions are strategic complements (so that $x_1^*(\alpha, \alpha) \cdot x_2^*(\alpha, \alpha) > 0$): $D(\alpha, \alpha) = 0 \Leftrightarrow \alpha - \sigma > 0$.

9.4 Proposition 3

From equation (12) and Proposition 1, a necessary and sufficient condition for the result in the proposition to obtain is $V_2(\alpha,\alpha) > 0$, which we now show.

From equation (34), we have

$$V_2(\alpha,\alpha) = \left[x_1^*(\alpha,\alpha) - \alpha x_2^*(\alpha,\alpha)\right] \cdot \pi_2\left[x^*(\alpha,\alpha), x^*(\alpha,\alpha)\right].$$

Under our hypothesis of unique Nash equilibrium, the slope of an individual's reaction function must be strictly smaller than 1, so that $|x_1^*(\alpha,\alpha)| > |x_2^*(\alpha,\alpha)|$. Hence, the term in square brackets has the same sign as $x_1^*(\alpha,\alpha)$, and

$$V_2(\alpha,\alpha) > 0 \Leftrightarrow x_1^*(\alpha,\alpha) \cdot \pi_2\left[x^*(\alpha,\alpha), x^*(\alpha,\alpha)\right] > 0,$$

which is true by Lemma 2.

9.5 Proposition 4

The necessary first-order conditions for Nash equilibrium (2) are

$$\begin{cases} (1+\alpha)\, x^{\rho-1}\left(x^\rho + y^\rho\right)^{\frac{1-\rho}{\rho}} = c(1+\nu)\, x^\nu \\ (1+\beta)\, y^{\rho-1}\left(x^\rho + y^\rho\right)^{\frac{1-\rho}{\rho}} = c(1+\nu)\, y^\nu. \end{cases}$$

Solving for x and y gives

$$x^*(\alpha,\beta) = \left[\frac{(1+\alpha)(1+\mu)^{\frac{1-\rho}{\rho}}}{c(1+\nu)}\right]^{\frac{1}{\nu}},$$

so that

$$x_1^*(\alpha,\beta) = \frac{1}{\nu}\left[\frac{(1+\alpha)^{1-\nu}(1+\mu)^{\frac{1-\rho}{\rho}}}{c(1+\nu)}\right]^{\frac{1}{\nu}}\left[1 - \frac{1-\rho}{(1+\mu)(1+\nu-\rho)}\left(\frac{1+\beta}{1+\alpha}\right)^{\frac{\rho}{1+\nu-\rho}}\right],$$

and

$$x_2^*(\alpha,\beta) = \frac{1}{\nu}\left[\frac{(1+\alpha)^{1-\nu}(1+\mu)^{\frac{1-\rho}{\rho}}}{c(1+\nu)}\right]^{\frac{1}{\nu}}\frac{1-\rho}{(1+\mu)(1+\nu-\rho)}\left(\frac{1+\beta}{1+\alpha}\right)^{\frac{2\rho-1-\nu}{1+\nu-\rho}}$$

Since

$$x_1^*(\alpha,\alpha) = \frac{1}{\nu}\left[\frac{(1+\alpha)^{1-\nu}\, 2^{\frac{1-\rho}{\rho}}}{c(1+\nu)}\right]^{\frac{1}{\nu}}\frac{1+2\nu-\rho}{2(1+\nu-\rho)},$$

and

$$x_2^*(\alpha,\alpha)=\frac{1}{\nu}\left[\frac{(1+\alpha)^{1-\nu}\,2^{\frac{1-\rho}{\rho}}}{c(1+\nu)}\right]^{\frac{1}{\nu}}\frac{1-\rho}{2(1+\nu-\rho)},$$

$D(\alpha,\alpha)=0$ reduces to

$$(\sigma-\alpha)(1+2\nu-\rho)+(1-\sigma\alpha)(1-\rho)=0,$$

or

$$\alpha^*=\frac{\sigma 2\nu+(1+\sigma)(1-\rho)}{2\nu+(1+\sigma)(1-\rho)}.$$

9.6 Lemma 3

The relevant situations are those when the group has at least one mutant. Hence, consider a group with n individuals of which $m+1$ are mutants, for $m=0,1,2,...,n-1$. Without loss of generality, index the individual mutants by the numbers $j=1,...,m+1$ and the incumbents by the numbers $j=m+2,...,n$. Focus on one of the mutants, and give this individual index $i=1$. This mutant's equilibrium payoff is affected as follows by a marginal change in the mutant degree of altruism, β:

$$\begin{aligned}\frac{\partial V(\beta,\boldsymbol{\alpha},m+1)}{\partial\beta} &= \frac{\partial\pi\left(x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1),\mathbf{x}^*_{-1}\right)}{\partial x_1}\cdot\frac{\partial x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1)}{\partial\alpha_1}\\ &+\sum_{j=2}^{n}\frac{\partial\pi\left(x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1),\mathbf{x}^*_{-1}\right)}{\partial x_j}\cdot\frac{\partial x^*(\alpha_j,\boldsymbol{\alpha}_{-j},m+1)}{\partial\alpha_1}\\ &+\sum_{j=2}^{m+1}\frac{\partial\pi\left(x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1),\mathbf{x}_{-1}\right)}{\partial x_1}\cdot\frac{\partial x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1)}{\partial\alpha_j}\\ &+\sum_{j=2}^{m+1}\frac{\partial\pi\left(x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1),\mathbf{x}^*_{-1}\right)}{\partial x_j}\cdot\frac{\partial x^*(\alpha_j,\boldsymbol{\alpha}_{-j},m+1)}{\partial\alpha_j}\\ &+\sum_{k=2}^{n}\sum_{j=2}^{m+1}\frac{\partial\pi\left(x^*(\alpha_1,\boldsymbol{\alpha}_{-1},m+1),\mathbf{x}^*_{-1}\right)}{\partial x_k}\cdot\frac{\partial x^*(\alpha_k,\boldsymbol{\alpha}_{-k},m+1)}{\partial\alpha_j}.\end{aligned}\tag{39}$$

The first two terms show how a change in the mutant's own degree of altruism affects the mutant's own behavior, and the behavior of the $n-1$ other individuals, respectively. The term on the third line is the effect of the change in the other m mutants' degree of altruism on the behavior of the mutant under consideration. The term on the fourth line is the effect of the change in the other m mutants' degrees of altruism on their own behavior. The last line represents the effect of the change in the other m mutants' degrees of altruism on other

individuals (except the mutant under consideration, whose corresponding behavioral change is already accounted for in the second line).

Using the first-order condition for mutant $i = 1$,

$$\frac{\partial \pi \left(x^* \left(\alpha_1, \boldsymbol{\alpha}_{-1}, m+1 \right), \mathbf{x}^*_{-1} \right)}{\partial x_1} + \beta \sum_{\ell=2}^{n} \frac{\partial \pi \left(x^* \left(\alpha_\ell, \boldsymbol{\alpha}_{-\ell}, m+1 \right), \mathbf{x}^*_{-\ell} \right)}{\partial x_1} = 0,$$

the expression in (39) becomes

$$\begin{aligned} \frac{\partial V \left(\beta, \boldsymbol{\alpha}, m+1 \right)}{\partial \beta} &= -\beta \sum_{\ell=2}^{n} \frac{\partial \pi \left(x^* \left(\alpha_\ell, \boldsymbol{\alpha}_{-\ell}, m+1 \right), \mathbf{x}^*_{-\ell} \right)}{\partial x_1} \cdot \frac{\partial x^* \left(\alpha_1, \boldsymbol{\alpha}_{-1}, m+1 \right)}{\partial \alpha_1} \\ &+ \sum_{j=2}^{n} \frac{\partial \pi \left(x^* \left(\alpha_1, \boldsymbol{\alpha}_{-1}, m+1 \right), \mathbf{x}^*_{-1} \right)}{\partial x_j} \cdot \frac{\partial x^* \left(\alpha_j, \boldsymbol{\alpha}_{-j}, m+1 \right)}{\partial \alpha_1} \\ &- \beta \left(\sum_{\ell=2}^{n} \frac{\partial \pi \left(x^* \left(\alpha_\ell, \boldsymbol{\alpha}_{-\ell}, m+1 \right), \mathbf{x}^*_{-\ell} \right)}{\partial x_1} \right) \sum_{j=2}^{m+1} \cdot \frac{\partial x^* \left(\alpha_1, \boldsymbol{\alpha}_{-1}, m+1 \right)}{\partial \alpha_j} \\ &+ \sum_{j=2}^{m+1} \frac{\partial \pi \left(x^* \left(\alpha_1, \boldsymbol{\alpha}_{-1}, m+1 \right), \mathbf{x}^*_{-1} \right)}{\partial x_j} \cdot \frac{\partial x^* \left(\alpha_j, \boldsymbol{\alpha}_{-j}, m+1 \right)}{\partial \alpha_j} \\ &+ \sum_{k=2}^{n} \sum_{j=2}^{m+1} \frac{\partial \pi \left(x^* \left(\alpha_1, \boldsymbol{\alpha}_{-1}, m+1 \right), \mathbf{x}^*_{-1} \right)}{\partial x_k} \cdot \frac{\partial x^* \left(\alpha_k, \boldsymbol{\alpha}_{-k}, m+1 \right)}{\partial \alpha_j}. \end{aligned} \tag{40}$$

Evaluated at the point $\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right)$, this expression becomes

$$\begin{aligned} \frac{\partial V \left(\beta, \alpha, m+1 \right)}{\partial \beta} &= -\alpha \cdot (n-1) \cdot \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot x^*_{own} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right) \\ &+ (n-1) \cdot \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot x^*_{other} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right) \\ &- \alpha \cdot (n-1) \cdot \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot m \cdot x^*_{other} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right) \\ &+ m \cdot \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot x^*_{own} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right) \\ &+ m (n-2) \cdot \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot x^*_{other} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right). \end{aligned}$$

Inserting this in equation (18), we obtain

$$\begin{aligned} D \left(\boldsymbol{\alpha} \right) &= \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot \\ &\cdot \sum_{m=0}^{n-1} \left(\begin{array}{c} n-1 \\ m \end{array} \right) \sigma^m (1-\sigma)^{n-1-m} \left[m - \alpha (n-1) \right] x^*_{own} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right) \\ &+ \pi_{other} \left[\mathbf{x}^* \left(\hat{\boldsymbol{\alpha}} \right) \right] \cdot \\ &\cdot \sum_{m=0}^{n-1} \left(\begin{array}{c} n-1 \\ m \end{array} \right) \sigma^m (1-\sigma)^{n-1-m} \left[(n-1)(1-\alpha m) + m (n-2) \right] x^*_{other} \left(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n \right) \end{aligned} \tag{41}$$

Since

$$\sum_{m=0}^{n-1}\left(\begin{array}{c} n-1 \\ m \end{array}\right)\sigma^{m}\left(1-\sigma\right)^{n-1-m}=1$$

and

$$\sum_{m=0}^{n-1}\left(\begin{array}{c} n-1 \\ m \end{array}\right)\sigma^{m}\left(1-\sigma\right)^{n-1-m}m=\left(n-1\right)\sigma,$$

(41) becomes

$$\begin{aligned} D\left(\boldsymbol{\alpha}\right) &= \left(n-1\right)\pi_{other}\left[\mathbf{x}^{*}\left(\hat{\boldsymbol{\alpha}}\right)\right]\cdot \\ &\quad \left[\left(\sigma-\alpha\right)x_{own}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)+\left[1-\sigma\left(n-1\right)\alpha+\sigma\left(n-2\right)\right]x_{other}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\right]. \end{aligned}$$

9.7 Proposition 6

Consider a game $G=\langle\pi,\alpha,\beta,n,k\rangle$ with $\pi\in\Pi$ and $\alpha,\beta\in(-1,1)$. We first establish

Lemma 4 *Consider a game $G=\langle\pi,\alpha,\alpha,n,n\rangle$ in which $\pi\in\Pi$ and $\alpha\in(-1,1)$. Let $\mathbf{x}^{*}\left(\hat{\boldsymbol{\alpha}},n\right)$ be its unique Nash equilibrium and suppose that this is regular. Then*

$$x_{own}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\cdot\pi_{other}\left[\mathbf{x}^{*}\left(\hat{\boldsymbol{\alpha}}\right)\right]>0$$

and:

1. $x_{own}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\neq0$ *and* $x_{other}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)=0$ *if the actions are strategically independent.*
2. $x_{own}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\cdot x_{other}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)<0$ *if the actions are strategic substitutes.*
3. $x_{own}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\cdot x_{other}^{*}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)>0$ *if the actions are strategic substitutes.*

Proof. Applying the Implicit Function Theorem to the equation

$$\frac{\partial\pi\left(x_{i},\mathbf{x}_{-i}\right)}{\partial x_{i}}+\alpha_{i}\sum_{j\neq i}\frac{\partial\pi\left(x_{j},\mathbf{x}_{-j}\right)}{\partial x_{i}}=0, \tag{42}$$

which implicitly defines the best response x_i of individual i to the vector of the other individuals' actions $\mathbf{x}_{-i}$, yields:

$$\frac{dx_{i}}{d\alpha_{i}}=-\frac{\sum_{j\neq i}\frac{\partial\pi(x_{j},\mathbf{x}_{-j})}{\partial x_{i}}}{\frac{\partial^{2}\pi(x_{i},\mathbf{x}_{-i})}{\partial x_{i}^{2}}+\alpha_{i}\sum_{j\neq i}\frac{\partial^{2}\pi(x_{j},\mathbf{x}_{-j})}{\partial x_{i}^{2}}}. \tag{43}$$

The denominator is negative under our hypothesis of a unique Nash equilibrium, which requires that (42) uniquely define x_i. Hence, (43) implies that $dx_i/d\alpha_i$ has the same sign as $\sum_{j\neq i}\partial\pi\left(x_j,\mathbf{x}_{-j}\right)/\partial x_i$, which by symmetry of π is the same as the sign of $\partial\pi\left(x_i,\mathbf{x}_{-i}\right)/\partial x_j$. Moreover, $dx_j/d\alpha_i=0$, for all $j\neq i$, since another individual's best response x_j is defined by:

$$\pi_j\left(x_j,\mathbf{x}_{-j}\right)+\alpha_j\sum_{k\neq j}\frac{\partial\pi\left(x_k,\mathbf{x}_{-k}\right)}{\partial x_j}=0.$$

Consider the Nash equilibrium $\mathbf{x}^*\left(\hat{\boldsymbol{\alpha}},n\right)$. The previous observations imply $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)>0\Leftrightarrow\pi_j>0$ and $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)<0\Leftrightarrow\pi_j<0$.

Applying again the Implicit Function Theorem to equation (42), we obtain

$$\frac{dx_i}{dx_k}=-\frac{\frac{\partial^2\pi(x_i,\mathbf{x}_{-i})}{\partial x_i\partial x_k}+\alpha_i\sum_{j\neq i}\frac{\partial^2\pi(x_j,\mathbf{x}_{-j})}{\partial x_i\partial x_k}}{\frac{\partial^2\pi(x_i,\mathbf{x}_{-i})}{\partial x_i^2}+\alpha_i\sum_{j\neq i}\frac{\partial^2\pi(x_j,\mathbf{x}_{-j})}{\partial x_i^2}}. \tag{44}$$

Since the denominator is negative, the previous observations imply the following: First, if actions are strategically independent ($\pi_{ij}=0$ for all $i\neq j$), (44) gives $\frac{dx_i}{dx_k}=0$ and hence $x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)=0$. Second, if actions are strategic substitutes ($\pi_{ij}<0$ for all $i\neq j$), (44) gives $\frac{dx_i}{dx_k}<0$. Hence, $x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)$ is non-zero and has the opposite sign of $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)$. Finally, if actions are strategic complements ($\pi_{ij}>0$ for all $i\neq j$), (44) gives $\frac{dx_i}{dx_k}>0$. Hence, $x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)$ is non-zero and has the same sign as $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)$.

■

Equation (19) and the assumptions $n\geq 2$ and $\pi_2\neq 0$ together imply

$$D\left(\boldsymbol{\alpha}\right)=0\Leftrightarrow\left(\alpha-\sigma\right)\cdot x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)=\left[1-\sigma\left(n-1\right)\alpha+\sigma\left(n-2\right)\right]\cdot x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right).$$

Since $1-\sigma\left(n-1\right)\alpha+\sigma\left(n-2\right)=1-\sigma\alpha+\sigma\left(n-2\right)\left(1-\alpha\right)>0$ for any $\sigma\in[0,1)$, $n\geq 2$, and $\alpha<1$, it follows that:

1. when actions are strategically independent (so that $x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)=0$ and $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\neq 0$): $D\left(\boldsymbol{\alpha}\right)=0\Leftrightarrow\alpha-\sigma=0$;

2. when actions are strategic substitutes (so that $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\cdot x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)<0$): $D\left(\boldsymbol{\alpha}\right)=0\Leftrightarrow\alpha-\sigma<0$;

3. when actions are strategic complements (so that $x^*_{own}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)\cdot x^*_{other}\left(\alpha,\hat{\boldsymbol{\alpha}}_{-i},n\right)>0$): $D\left(\boldsymbol{\alpha}\right)=0\Leftrightarrow\alpha-\sigma>0$.

9.8 Proposition 7

The result in the proposition obtains if $D(\boldsymbol{\alpha})$ is strictly increasing in σ. But according to (19), this follows if

$$[x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) + [(n-2) - (n-1)\alpha] x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)] \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0$$

or, equivalently, if

$$[x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) - x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) + (n-1)(1-\alpha) x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)] \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0. \tag{45}$$

Under our hypothesis of a unique Nash equilibrium, $|x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)| > |x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)|$. Hence, $x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) - x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)$ has the same sign as $x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)$, and this implies $[x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) - x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n)] \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0$ (since $x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0$ — see the Proof of Proposition 6). Hence, a sufficient condition for (45) is that $x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] \geq 0$. This inequality holds when actions are strategic complements or strategically independent.

When actions are strategic substitutes, we use the fact that we only need to verify inequality (45) for α such that $D(\boldsymbol{\alpha}) = 0$, i.e., such that

$$x^*_{own}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) = \frac{1 - \sigma(n-1)\alpha + \sigma(n-2)}{\alpha - \sigma} \cdot x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n).$$

Using this expression, inequality (45) becomes

$$\frac{1 - \sigma(n-1)\alpha + \sigma(n-2) - (\alpha - \sigma) + (\alpha - \sigma)(n-1)(1-\alpha)}{\alpha - \sigma} \cdot x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0,$$

or, equivalently,

$$\frac{(1-\alpha)(1-\alpha+n\alpha)}{\alpha - \sigma} \cdot x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0. \tag{46}$$

Since $\alpha \in [0, 1)$ implies $1 - \alpha + n\alpha > 0$, (46) holds if

$$(\alpha - \sigma) \cdot x^*_{other}(\alpha, \hat{\boldsymbol{\alpha}}_{-i}, n) \cdot \pi_{other}[\mathbf{x}^*(\hat{\boldsymbol{\alpha}})] > 0.$$

It is easily verified that this is true by Proposition 6.

9.9 Proposition 8

For the purpose of this proof, let an index i denote individual i, and let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$. The first-order condition for individual i writes:

$$cx_i = [1 + (n-1)\alpha_i] \tau \left(\sum_{j=1}^{n} x_j \right)^{\tau - 1}. \tag{47}$$

Adding the n players' first-order conditions, and rearranging the terms gives

$$\sum_{j=1}^{n} x_j = \left[\frac{\tau}{c}\left(n + (n-1)\sum_{j=1}^{n}\alpha_j\right)\right]^{\frac{1}{2-\tau}}.$$

Using this in equation (47) gives the following expression for i's equilibrium action:

$$x_i(\boldsymbol{\alpha}) = [1 + (n-1)\alpha_i]\left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}\left[n + (n-1)\sum_{j=1}^{n}\alpha_j\right]^{\frac{\tau-1}{2-\tau}},$$

so that

$$\begin{aligned}\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_i} &= (n-1)\left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}\left[n + (n-1)\sum_{j=1}^{n}\alpha_j\right]^{\frac{\tau-1}{2-\tau}} \cdot \\ &\quad \left[1 + \frac{\tau-1}{2-\tau}\cdot\frac{1 + (n-1)\alpha_i}{n + (n-1)\sum_{j=1}^{n}\alpha_j}\right]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_k} &= (n-1)\left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}\left[n + (n-1)\sum_{j=1}^{n}\alpha_j\right]^{\frac{\tau-1}{2-\tau}} \cdot \\ &\quad \left[\frac{\tau-1}{2-\tau}\cdot\frac{1 + (n-1)\alpha_k}{n + (n-1)\sum_{j=1}^{n}\alpha_j}\right].\end{aligned}$$

For $\alpha_1 = ... = \alpha_n = \alpha$, these expressions become

$$\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_i} = n(n-1)\left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}[1 + (n-1)\alpha]^{\frac{\tau-1}{2-\tau}} \cdot \left[1 + \frac{\tau-1}{n(2-\tau)}\right]$$

and

$$\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_k} = n(n-1)\left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}[1 + (n-1)\alpha]^{\frac{\tau-1}{2-\tau}} \cdot \frac{\tau-1}{n(2-\tau)}.$$

The condition for evolutionarily stable altruism $D(\boldsymbol{\alpha}) = 0$ is then equivalent to

$$\begin{aligned}0 &= (\sigma-\alpha)\left[1 + \frac{\tau-1}{n(2-\tau)}\right] + [1 - \sigma(n-1)\alpha + \sigma(n-2)]\frac{\tau-1}{n(2-\tau)} \\ &\Leftrightarrow\ 0 = (\sigma-\alpha)[n(2-\tau) + \tau - 1] + [1 - \sigma(n-1)\alpha + \sigma(n-2)](\tau-1) \\ &\Leftrightarrow\ \alpha = \frac{(\tau-1)[1 + \sigma(n-1)] + \sigma n(2-\tau)}{(\tau-1)[1 + \sigma(n-1)] + n(2-\tau)}.\end{aligned}$$

9.10 Proposition 9

For the purpose of this proof, let an index i denote individual i, and let $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)$. The first-order condition for individual i writes:

$$cx_i = [1 + (n-1)\,\alpha_i]\,\frac{\theta}{n} x_i^{\frac{\theta}{n}-1} \left(\prod_{j \neq i} x_j\right)^{\frac{\theta}{n}},$$

or

$$x_i = \left[[1 + (n-1)\,\alpha_i]\,\frac{\theta}{nc}\left(\prod_{j=i}^{n} x_j\right)^{\frac{\theta}{n}}\right]^{1/2}. \tag{48}$$

Multiplying the n players' first-order conditions, and rearranging the terms gives

$$\left(\prod_{j=i}^{n} x_j\right)^{\frac{\theta}{n}} = \left(\frac{\theta}{nc}\right)^{\frac{\theta}{2-\theta}} \prod_{j=i}^{n} [1 + (n-1)\,\alpha_j]^{\frac{\theta}{n(2-\theta)}}.$$

Using this in equation (48) gives the following expression for i's equilibrium action:

$$x_i(\boldsymbol{\alpha}) = [1 + (n-1)\,\alpha_i]^{1/2} \left(\frac{\theta}{nc}\right)^{\frac{1}{2-\theta}} \prod_{j=i}^{n} [1 + (n-1)\,\alpha_j]^{\frac{\theta}{2n(2-\theta)}},$$

so that

$$\begin{aligned}\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_i} &= \left(\frac{\theta}{nc}\right)^{\frac{1}{2-\theta}} \cdot \frac{n-1}{2} \cdot \prod_{j=i}^{n} [1 + (n-1)\,\alpha_j]^{\frac{\theta}{2n(2-\theta)}} \\ &\quad \cdot [1 + (n-1)\,\alpha_i]^{-1/2} \cdot \left[1 + \frac{\theta}{n(2-\theta)}\right]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_k} &= \left(\frac{\theta}{nc}\right)^{\frac{1}{2-\theta}} \cdot \frac{n-1}{2} \cdot \prod_{j=i}^{n} [1 + (n-1)\,\alpha_j]^{\frac{\theta}{2n(2-\theta)}} \\ &\quad \cdot [1 + (n-1)\,\alpha_i]^{1/2} \cdot [1 + (n-1)\,\alpha_k]^{-1} \cdot \frac{\theta}{n(2-\theta)}.\end{aligned}$$

For $\alpha_1 = ... = \alpha_n = \alpha$, these expressions become

$$\begin{aligned}\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_i} &= \left(\frac{\theta}{nc}\right)^{\frac{1}{2-\theta}} \cdot \frac{n-1}{2} \cdot [1 + (n-1)\,\alpha]^{\frac{\theta}{2(2-\theta)}} \\ &\quad \cdot [1 + (n-1)\,\alpha]^{-1/2} \cdot \left[1 + \frac{\theta}{n(2-\theta)}\right]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial x_i(\boldsymbol{\alpha})}{\partial \alpha_k} &= \left(\frac{\theta}{nc}\right)^{\frac{1}{2-\theta}} \cdot \frac{n-1}{2} \cdot [1+(n-1)\alpha]^{\frac{\theta}{2(2-\theta)}} \\ &\quad \cdot [1+(n-1)\alpha]^{-1/2} \cdot \frac{\theta}{n(2-\theta)}.\end{aligned}$$

The condition for evolutionarily stable altruism $D(\boldsymbol{\alpha}) = 0$ is then equivalent to

$$\begin{aligned}0 &= (\sigma-\alpha)\left[1+\frac{\theta}{n(2-\theta)}\right] + [1-\sigma(n-1)\alpha+\sigma(n-2)]\frac{\theta}{n(2-\theta)} \\ &\Leftrightarrow\ 0 = (\sigma-\alpha)[n(2-\theta)+\theta] + [1-\sigma(n-1)\alpha+\sigma(n-2)]\theta \\ &\Leftrightarrow\ \alpha = \frac{\theta[1+\sigma(n-1)]+\sigma n(2-\theta)}{\theta[1+\sigma(n-1)]+n(2-\theta)}.\end{aligned}$$

9.11 Proposition 10

Let

$$x^1 \in \arg\max_{x\geq 0}\ \pi(x,x)$$

The unconstrained Nash bargaining solution in the symmetric case $\alpha = \beta$ is symmetric and efficient, so $\hat{x}(\alpha,\alpha,1) = x^1$ for all $\alpha \in (-1,1)$.[14] Moreover, $(\hat{x}(\alpha,\alpha,1), \hat{x}(\alpha,\alpha,1))$ is feasible and individually rational in the stage game. By the (Nash-threat version of the) Folk Theorem, there exists a $\bar{\delta} < 1$ such that $(x^1,x^1) \in S(\alpha,\alpha,\delta)$ for all $\delta \in \left(\bar{\delta},1\right)$ and all $\alpha \in (-1,1)$. Indeed, by continuity there exists a $\hat{\delta} \in \left(\bar{\delta},1\right)$ and $\varepsilon > 0$ such that $(\hat{x}(\alpha,\beta,\delta), \hat{x}(\beta,\alpha,\delta)) \in S(\alpha,\beta,\delta)$ for all $\delta \in \left(\hat{\delta},1\right)$ and all (α,β) with $|\alpha-\beta| < \varepsilon$. Moreover, $\hat{x}(\alpha,\beta,\delta) = \hat{x}(\alpha,\beta,1)$ and $\hat{x}(\beta,\alpha,\delta) = \hat{x}(\beta,\alpha,1)$ for all such α, β and δ. Hence, for such α, β and δ, $V(\alpha,\beta,1)$ is the material payoff to the α-altruist in the Nash bargaining solution in (22) for all $|\alpha-\beta| < \varepsilon$ and therefore $V_1(\alpha,\alpha,\delta) \equiv V_1(\alpha,\alpha,1)$ and $V_2(\alpha,\alpha,\delta) \equiv V_1(\alpha,\alpha,1)$ for all $\alpha \in (-1,1)$ and all $\delta \in \left(\hat{\delta},1\right)$. Finally, $V(\alpha,\alpha,1) = \pi(x^1,x^1)$ and thus $V_2(\alpha,\alpha,1) + V_1(\alpha,\alpha,1) = 0$ which gives the last term in the expression for the drift function.

9.12 Proposition 13

Since

$$\begin{aligned}V_1(\beta,\mu,\alpha,\lambda) &= \pi_1[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)] \cdot x_1^*(\beta,\mu,\alpha,\lambda) \\ &\quad +\pi_2[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)] \cdot x_3^*(\alpha,\lambda,\beta,\mu),\end{aligned}$$

[14] For such α, maximization of $(1+\alpha)\pi(x,x)$ is equivalent with maximization of $\pi(x,x)$.

$$\begin{aligned} V_2(\beta,\mu,\alpha,\lambda) &= \pi_1\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right]\cdot x_2^*(\beta,\mu,\alpha,\lambda) \\ &\quad +\pi_2\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right]\cdot x_4^*(\alpha,\lambda,\beta,\mu), \end{aligned}$$

$$\begin{aligned} V_3(\beta,\mu,\alpha,\lambda) &= \pi_1\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right]\cdot x_3^*(\beta,\mu,\alpha,\lambda) \\ &\quad +\pi_2\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right]\cdot x_1^*(\alpha,\lambda,\beta,\mu), \end{aligned}$$

$$\begin{aligned} V_4(\beta,\mu,\alpha,\lambda) &= \pi_1\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right]\cdot x_4^*(\beta,\mu,\alpha,\lambda) \\ &\quad +\pi_2\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right]\cdot x_2^*(\alpha,\lambda,\beta,\mu), \end{aligned}$$

and the first-order condition for the (β,μ)-type writes

$$\pi_1\left[x^*(\beta,\mu,\alpha,\lambda), x^*(\alpha,\lambda,\beta,\mu)\right] + \frac{\beta+\mu(\alpha-\beta)}{1+\mu}\cdot\pi_2\left[x^*(\alpha,\lambda,\beta,\mu), x^*(\beta,\mu,\alpha,\lambda)\right] = 0,$$

we obtain

$$\begin{aligned} V_1(\beta,\mu,\alpha,\lambda)+\sigma V_3(\beta,\mu,\alpha,\lambda) &= \left[\sigma - \frac{\beta+\mu(\alpha-\beta)}{1+\mu}\right]\cdot x_1^*(\beta,\mu,\alpha,\lambda) \\ &\quad \cdot\pi_2\left[x^*(\alpha,\lambda,\beta,\mu), x^*(\beta,\mu,\alpha,\lambda)\right] \\ &\quad + \left[1 - \frac{\beta+\mu(\alpha-\beta)}{1+\mu}\cdot\sigma\right]\cdot x_3^*(\beta,\mu,\alpha,\lambda) \\ &\quad \cdot\pi_2\left[x^*(\alpha,\lambda,\beta,\mu), x^*(\beta,\mu,\alpha,\lambda)\right], \end{aligned}$$

and

$$\begin{aligned} V_2(\beta,\mu,\alpha,\lambda)+\sigma V_4(\beta,\mu,\alpha,\lambda) &= \left[\sigma - \frac{\beta+\mu(\alpha-\beta)}{1+\mu}\right]\cdot x_2^*(\beta,\mu,\alpha,\lambda) \\ &\quad \cdot\pi_2\left[x^*(\alpha,\lambda,\beta,\mu), x^*(\beta,\mu,\alpha,\lambda)\right] \\ &\quad + \left[1 - \frac{\beta+\mu(\alpha-\beta)}{1+\mu}\cdot\sigma\right]\cdot x_4^*(\beta,\mu,\alpha,\lambda) \\ &\quad \cdot\pi_2\left[x^*(\alpha,\lambda,\beta,\mu), x^*(\beta,\mu,\alpha,\lambda)\right], \end{aligned}$$

so that

$$\begin{aligned} \left[V_1(\beta,\mu,\alpha,\lambda)+\sigma V_3(\beta,\mu,\alpha,\lambda)\right]_{|(\beta,\mu)=(\alpha,\lambda)} &= \left[\sigma - \frac{\alpha}{1+\lambda}\right]\cdot x_1^*(\alpha,\lambda,\alpha,\lambda) \\ &\quad \cdot\pi_2\left[x^*(\alpha,\lambda,\alpha,\lambda), x^*(\alpha,\lambda,\alpha,\lambda)\right] \\ &\quad + \left[1 - \frac{\alpha}{1+\lambda}\cdot\sigma\right]\cdot x_3^*(\alpha,\lambda,\alpha,\lambda) \\ &\quad \cdot\pi_2\left[x^*(\alpha,\lambda,\alpha,\lambda), x^*(\alpha,\lambda,\alpha,\lambda)\right] \end{aligned}$$

and

$$\begin{aligned}\left[V_2\left(\beta,\mu,\alpha,\lambda\right)+\sigma+V_4\left(\beta,\mu,\alpha,\lambda\right)\right]_{|(\beta,\mu)=(\alpha,\lambda)} &= \left[\sigma-\frac{\alpha}{1+\lambda}\right]\cdot x_2^*\left(\alpha,\lambda,\alpha,\lambda\right)\\ &\cdot\pi_2\left[x^*\left(\alpha,\lambda,\alpha,\lambda\right),x^*\left(\alpha,\lambda,\alpha,\lambda\right)\right]\\ &+\left[1-\frac{\alpha}{1+\lambda}\cdot\sigma\right]\cdot x_4^*\left(\alpha,\lambda,\alpha,\lambda\right)\\ &\cdot\pi_2\left[x^*\left(\alpha,\lambda,\alpha,\lambda\right),x^*\left(\alpha,\lambda,\alpha,\lambda\right)\right].\end{aligned}$$

Since $\pi_2\neq 0$,

$$\begin{cases}\left[V_1\left(\beta,\mu,\alpha,\lambda\right)+\sigma V_3\left(\beta,\mu,\alpha,\lambda\right)\right]_{|(\beta,\mu)=(\alpha,\lambda)}=0\\ \left[V_2\left(\beta,\mu,\alpha,\lambda\right)+\sigma+V_4\left(\beta,\mu,\alpha,\lambda\right)\right]_{|(\beta,\mu)=(\alpha,\lambda)}=0\end{cases}$$

if and only if

$$\begin{cases}\left[\sigma-\frac{\alpha}{1+\lambda}\right]\cdot x_1^*\left(\alpha,\lambda,\alpha,\lambda\right)+\left[1-\frac{\alpha}{1+\lambda}\cdot\sigma\right]\cdot x_3^*\left(\alpha,\lambda,\alpha,\lambda\right)=0\\ \left[\sigma-\frac{\alpha}{1+\lambda}\right]\cdot x_2^*\left(\alpha,\lambda,\alpha,\lambda\right)+\left[1-\frac{\alpha}{1+\lambda}\cdot\sigma\right]\cdot x_4^*\left(\alpha,\lambda,\alpha,\lambda\right)=0.\end{cases}\tag{49}$$

With the payoff function $\pi\left(x,y\right)=\left(x+y\right)^{\tau}-cx^2/2$, the necessary first-order conditions for Nash equilibrium (2) write

$$\begin{cases}\left[1+\frac{\alpha+\lambda(\beta-\alpha)}{1+\lambda}\right]\tau\left(x+y\right)^{\tau-1}=cx\\ \left[1+\frac{\beta+\mu(\alpha-\beta)}{1+\mu}\right]\tau\left(x+y\right)^{\tau-1}=cy.\end{cases}$$

Solving for x and y gives

$$\begin{aligned}x^*\left(\beta,\mu,\alpha,\lambda\right) &= \left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}\cdot\left[1+\frac{\beta+\mu\left(\alpha-\beta\right)}{1+\mu}\right]\cdot\\ &\left[\frac{2\left(1+\lambda\right)\left(1+\mu\right)+\left(1+2\mu-\lambda\right)\alpha+\left(1+2\lambda-\mu\right)\beta}{\left(1+\lambda\right)\left(1+\mu\right)}\right]^{\frac{\tau-1}{2-\tau}}.\end{aligned}$$

so that

$$\begin{aligned}x_1^*\left(\beta,\mu,\alpha,\lambda\right) &= \left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}\cdot\left[\frac{2\left(1+\lambda\right)\left(1+\mu\right)+\left(1+2\mu-\lambda\right)\alpha+\left(1+2\lambda-\mu\right)\beta}{\left(1+\lambda\right)\left(1+\mu\right)}\right]^{\frac{\tau-1}{2-\tau}}\cdot\\ &\frac{1}{1+\mu}\left[1-\mu+\frac{\left[1+\mu+\beta+\mu\left(\alpha-\beta\right)\right]\left(1+2\lambda-\mu\right)}{2\left(1+\lambda\right)\left(1+\mu\right)+\left(1+2\mu-\lambda\right)\alpha+\left(1+2\lambda-\mu\right)\beta}\right]\end{aligned}$$

$$\begin{aligned}x_2^*\left(\beta,\mu,\alpha,\lambda\right) &= \left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}}\cdot\left[\frac{2\left(1+\lambda\right)\left(1+\mu\right)+\left(1+2\mu-\lambda\right)\alpha+\left(1+2\lambda-\mu\right)\beta}{\left(1+\lambda\right)\left(1+\mu\right)}\right]^{\frac{\tau-1}{2-\tau}}\cdot\\ &\frac{1}{\left(1+\mu\right)^2}\left[\alpha-2\beta+\frac{\left[1+\mu+\beta+\mu\left(\alpha-\beta\right)\right]\left(\alpha-2\beta\right)\left(1+\lambda\right)}{2\left(1+\lambda\right)\left(1+\mu\right)+\left(1+2\mu-\lambda\right)\alpha+\left(1+2\lambda-\mu\right)\beta}\right]\end{aligned}$$

$$x_3^*(\beta,\mu,\alpha,\lambda) = \left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}} \cdot \left[\frac{2(1+\lambda)(1+\mu)+(1+2\mu-\lambda)\alpha+(1+2\lambda-\mu)\beta}{(1+\lambda)(1+\mu)}\right]^{\frac{\tau-1}{2-\tau}} \cdot \frac{1}{1+\mu}\left[\mu+\frac{[1+\mu+\beta+\mu(\alpha-\beta)](1+2\lambda-\mu)}{2(1+\lambda)(1+\mu)+(1+2\mu-\lambda)\alpha+(1+2\lambda-\mu)\beta}\right]$$

and

$$x_4^*(\beta,\mu,\alpha,\lambda) = \left(\frac{\tau}{c}\right)^{\frac{1}{2-\tau}} \cdot \left[\frac{2(1+\lambda)(1+\mu)+(1+2\mu-\lambda)\alpha+(1+2\lambda-\mu)\beta}{(1+\lambda)(1+\mu)}\right]^{\frac{\tau-1}{2-\tau}} \cdot \frac{1}{(1+\lambda)}\left[\frac{[1+\mu+\beta+\mu(\alpha-\beta)][\beta-2\alpha-2(1+\lambda)]}{2(1+\lambda)(1+\mu)+(1+2\mu-\lambda)\alpha+(1+2\lambda-\mu)\beta}\right].$$

Ignoring the irrelevant constants, (49) is equivalent to

$$\begin{cases} (3-2\lambda)[\sigma(1+\lambda)-\alpha]+(1+2\lambda)(1+\lambda-\alpha\sigma)=0 \\ -3\alpha[\sigma(1+\lambda)-\alpha]+[\alpha-2(1+\lambda)](1+\lambda-\alpha\sigma)=0. \end{cases}$$

The solution is $(\alpha,\lambda)=\left(\frac{13-5\sigma}{15-8\sigma+\sigma^2},\frac{3(1-\sigma)}{2(5-\sigma)}\right)$ if $\sigma<1$, and $\alpha=\lambda+1$ if $\sigma=1$.

10 Appendix B: Repeated interactions

10.0.1 Nash reversion

Let

$$\pi(x,y)=(x+y)^\tau-\frac{1}{2}x^2$$

for $x,y\geq 0$ and let

$$u(x,y,\alpha)=\pi(x,y)+\alpha\pi(y,x)=(1+\alpha)(x+y)^\tau-\frac{1}{2}\left(x^2+\alpha y^2\right)$$

$$u(y,x,\beta)=\pi(y,x)+\beta\pi(x,y)=(1+\beta)(x+y)^\tau-\frac{1}{2}\left(y^2+\beta x^2\right)$$

for $\alpha,\beta\in(-1,1)$.

In the unique Nash equilibrium of the stage game, an α-altruist plays

$$x^*(\alpha,\beta)=(1+\alpha)(2+\alpha+\beta)^{\frac{\tau-1}{2-\tau}}\tau^{\frac{1}{2-\tau}}.$$

against a β-altruist. Hence, the *material* payoff to the α-altruist, in the unique Nash equilibrium of the stage game, is

$$V^{NE}(\alpha,\beta) = (x^*(\alpha,\beta)+x^*(\beta,\alpha))^\tau-\frac{1}{2}[x^*(\alpha,\beta)]^2 = [(2+\alpha+\beta)\tau]^{\frac{1}{2-\tau}}-\frac{1}{2}(1+\alpha)^2(2+\alpha+\beta)^{\frac{2(\tau-1)}{2-\tau}}\tau^{\frac{2}{2-\tau}},$$

and the total equilibrium utility to the α-altruist is

$$\begin{aligned} u^*(\alpha,\beta) &= V^{NE}(\alpha,\beta)+\alpha V^{NE}(\beta,\alpha) \\ &= \left[(2+\alpha+\beta)\tau\right]^{\frac{1}{2-\tau}}-\frac{1}{2}(1+\alpha)^2(2+\alpha+\beta)^{\frac{2(\tau-1)}{2-\tau}}\tau^{\frac{2}{2-\tau}} \\ &\quad +\alpha\left[\left[(2+\alpha+\beta)\tau\right]^{\frac{1}{2-\tau}}-\frac{1}{2}(1+\beta)^2(2+\alpha+\beta)^{\frac{2(\tau-1)}{2-\tau}}\tau^{\frac{2}{2-\tau}}\right] \\ &= (1+\alpha)\left[(2+\alpha+\beta)\tau\right]^{\frac{1}{2-\tau}}-\frac{1}{2}\left[(1+\alpha)^2+\alpha(1+\beta)^2\right](2+\alpha+\beta)^{\frac{2(\tau-1)}{2-\tau}}\tau^{\frac{2}{2-\tau}} \end{aligned} \tag{50}$$

and likewise for the β-altruist.

Suppose that this stage game is repeated indefinitely, and that both players use the same discount factor $\delta\in(0,1)$. Repeated play of a NE-improving action pair (x,y) is a subgame perfect equilibrium outcome, supported by threat of reversion forever to the stage-game Nash equilibrium, if and only if

$$\begin{cases} (1-\delta)u^+(\alpha,y)+\delta u^*(\alpha,\beta)\leq u(x,y,\alpha) \\ (1-\delta)u^+(\beta,x)+\delta u^*(\beta,\alpha)\leq u(y,x,\beta), \end{cases} \tag{51}$$

where

$$\begin{cases} u^+(\alpha,y)=(1+\alpha)(x^+(y)+y)^\tau-\frac{1}{2}\left(x^+(y)^2+\alpha y^2\right) \\ u^+(\beta,x)=(1+\beta)(x+y^+(x))^\tau-\frac{1}{2}\left(y^+(x)^2+\beta x^2\right) \end{cases}$$

and

$$\begin{cases} x^+(y)=\arg\max_{x\geq 0}(1+\alpha)(x+y)^\tau-\frac{1}{2}(x^2+\alpha y^2) \\ y^+(x)=\arg\max_{y\geq 0}(1+\beta)(x+y)^\tau-\frac{1}{2}(y^2+\beta x^2). \end{cases}$$

The inequalities (51) define the set of sustainable actions $S(\alpha,\beta,\delta)$.

The stable degrees of altruism in Table 1 are based on the assumption that, with probability 1/2, in each period the players play the stage-game Nash equilibrium actions, and that with probability 1/2, in each period they play the pair of actions that solve the constrained Nash bargaining problem:

$$\begin{aligned} \max_{(x,y)\in S(\alpha,\beta,\delta)} &\left[(1+\alpha)(x+y)^\tau-\frac{1}{2}\left(x^2+\alpha y^2\right)-u^*(\alpha,\beta)\right] \\ &\cdot\left[(1+\beta)(x+y)^\tau-\frac{1}{2}\left(y^2+\beta x^2\right)-u^*(\beta,\alpha)\right]. \end{aligned}$$

10.0.2 Endogenous threats

For arbitrary $x_0, y_0 \geq 0$, consider the Nash bargaining problem

$$\max_{x,y\geq 0} \quad \left[(1+\alpha)(x-x_0+y-y_0) - \frac{1}{2}\left[x^2 - x_0^2 + \alpha\left(y^2 - y_0^2\right)\right]\right] \cdot \left[(1+\beta)(x-x_0+y-y_0) - \frac{1}{2}\left[y^2 - y_0^2 + \beta\left(x^2 - x_0^2\right)\right]\right].$$

Necessary first-order conditions defining a solution (x, y) are:

$$\begin{cases} (1+\alpha-x)\left[(1+\beta)(x-x_0+y-y_0) - \frac{1}{2}\left[y^2 - y_0^2 + \beta(x^2 - x_0^2)\right]\right] \\ +(1+\beta-\beta x)\left[(1+\alpha)(x-x_0+y-y_0) - \frac{1}{2}\left[x^2 - x_0^2 + \alpha(y^2 - y_0^2)\right]\right] = 0 \\ (1+\alpha-\alpha y)\left[(1+\beta)(x-x_0+y-y_0) - \frac{1}{2}\left[y^2 - y_0^2 + \beta(x^2 - x_0^2)\right]\right] \\ +(1+\beta-y)\left[(1+\alpha)(x-x_0+y-y_0) - \frac{1}{2}\left[x^2 - x_0^2 + \alpha(y^2 - y_0^2)\right]\right] = 0. \end{cases} \tag{52}$$

These equations implicitly define x and y as functions of x_0 and y_0: we denote these by $\hat{x}(x_0, y_0)$ and $\hat{y}(y_0, x_0)$.

In the simultaneous-move game in which the α-altruist chooses $x_0 \geq 0$, the β-altruist chooses $y_0 \geq 0$ and the payoff to the first is

$$v(x_0, y_0, \alpha, \beta) = (1+\alpha)\left(\hat{x}(x_0, y_0) + \hat{y}(y_0, x_0)\right) - \frac{1}{2}\left[\hat{x}(x_0, y_0)^2 + \alpha\hat{y}(y_0, x_0)^2\right]$$

and to the second it is

$$v(y_0, x_0, \beta, \alpha) = (1+\beta)\left(\hat{x}(x_0, y_0) + \hat{y}(y_0, x_0)\right) - \frac{1}{2}\left[\hat{y}(y_0, x_0)^2 + \beta\hat{x}(x_0, y_0)^2\right],$$

necessary first-order conditions for an interior Nash equilibrium, $x_0 > 0$ and $y_0 > 0$, are:

$$\begin{cases} \left(1+\alpha-\hat{x}(x_0, y_0)\right)\frac{\partial\hat{x}(x_0,y_0)}{\partial x_0} + \left(1+\alpha-\alpha\hat{y}(y_0, x_0)\right)\frac{\partial\hat{y}(y_0,x_0)}{\partial x_0} = 0 \\ \left(1+\beta-\hat{y}(y_0, x_0)\right)\frac{\partial\hat{y}(y_0,x_0)}{\partial y_0} + \left(1+\beta-\beta\hat{x}(x_0, y_0)\right)\frac{\partial\hat{x}(x_0,y_0)}{\partial y_0} = 0. \end{cases}$$

A necessary condition for $(0, 0)$ to be a Nash equilibrium is:

$$\begin{cases} \left(1+\alpha-\hat{x}\right)\frac{\partial\hat{x}(x_0,y_0)}{\partial x_0}_{|x_0=y_0=0} + \left(1+\alpha-\alpha\hat{y}\right)\frac{\partial\hat{y}(y_0,x_0)}{\partial x_0}_{|x_0=y_0=0} \leq 0 \\ \left(1+\beta-\hat{y}\right)\frac{\partial\hat{y}(y_0,x_0)}{\partial y_0}_{|x_0=y_0=0} + \left(1+\beta-\beta\hat{x}\right)\frac{\partial\hat{x}(x_0,y_0)}{\partial y_0}_{|x_0=y_0=0} \leq 0. \end{cases}$$

In (52), write the left-hand side of the first equation as $f\left(\hat{x}(x_0, y_0), \hat{y}(y_0, x_0), x_0, y_0\right)$, and the left-hand side of the second equation as $g\left(\hat{x}(x_0, y_0), \hat{y}(y_0, x_0), x_0, y_0\right)$. Also, let subscripts 1 to 4 denote the partial derivatives of f and g. Then, by the Implicit Function Theorem,

$$\frac{\partial\hat{x}(x_0, y_0)}{\partial x_0} = \frac{f_2 \cdot g_3 - f_3 \cdot g_2}{f_1 \cdot g_2 - f_2 \cdot g_1}$$

and

$$\frac{\partial \hat{y}\left(y_0, x_0\right)}{\partial y_0} = \frac{f_2 \cdot g_4 - f_4 \cdot g_2}{f_1 \cdot g_2 - f_2 \cdot g_1}.$$

References

Abreu, Dilip, and David Pearce. 2002. "Bargaining, Reputation and Equilibrium Selection in Repeated Games," *mimeo, Princeton University.*

Abreu, Dilip, and David Pearce. 2007. "Bargaining, Reputation, and Equilibrium Selection in Repeated Games with Contracts," *Econometrica* 75, 653-710.

Alger, Ingela, and Jörgen W. Weibull. 2010. "Kinship, Incentives and Evolution," *American Economic Review,* 100:1727-1760.

Axelrod, R. and William D. Hamilton. 1981. "The Evolution of Cooperation," *Science*, 211:1390-1396.

Bartle, Robert G. 1976. *The Elements of Real Analysis, Second Edition.* New York: John Wiley and Sons.

Bergstrom, Theodore C. 1995. "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review,* 85:58-81.

Bergstrom, Theodore C. 2003. "The Algebra of Assortative Encounters and the Evolution of Cooperation," *International Game Theory Review,* 5(3):211-228.

Bernheim, Douglas and Oded Stark. 1988. "Altruism within the Family Reconsidered: Do Nice Guys Finish Last?", *American Economic Review* 78, 1034-1045.

Bester, Helmut and Werner Güth. 1998. "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization,* 34:193–209.

Bisin, Alberto, Giorgio Topa, and Thierry Verdier. 2004. "Cooperation as a Transmitted Cultural Trait," *Rationality and Society*, 16:477-507.

Boyd, R., and P.J. Richerson. 1985. *Culture and the Evolutionary Process.* Chicago: Chicago University Press.

Cavalli-Sforza, Luigi L., and Marcus W. Feldman. 1981. *Cultural Transmission and Evolution: A Quantitative Approach.* Princeton, NJ: Princeton University Press.

Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences With Simple Tests," *Quarterly Journal of Economics*, 117:817-869.

Choi, Jung-Kyoo. 2008. "Play Locally, Learn Globally: Group Selection and Structural Basis of Cooperation," *Journal of Bioeconomics,* 10:239-257.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* London: John Murray.

Darwin, Charles. 1871. *The Descent of Man, and Selection in Relation to Sex.* London: John Murray.

Dekel, Eddie, Jeffrey C. Ely, and Okan Yilankaya. 2007. "Evolution of Preferences," *Review of Economic Studies*, 74:685-704.

Durrett, Richard and Simon A. Levin. 2005. "Can Stable Social Groups be Maintained by Homophilous Imitation Alone?" *Journal of Economic Behavior & Organization,* 57:267–286.

Ely, Jeffrey C. and Okan Yilankaya. 2001. "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory,* 97:255-272.

Eshel, Ilan, and L.L. Cavalli-Sforza. 1982. "Assortment of Encounters and Evolution of Cooperativeness," *Proceedings of the National Academy of Sciences*, 79:1331-1335.

Fehr, Ernst and Klaus Schmidt. 1999. "A theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics,* 114:817-868.

Fershtman, Chaim and Yoram Weiss. 1998. "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics,* 70:53–73.

Fletcher, Jeffrey A. and Michael Doebeli. 2009. "A Simple and General Explanation for the Evolution of Altruism," *Proceedings of the Royal Society Biology,* 276:13–19.

Grafen, Alan. 1979. "The Hawk-Dove Game Played between Relatives," *Animal Behavior,* 27:905–907.

Grafen, Alan. 2006. "Optimization of Inclusive Fitness," *Journal of Theoretical Biology,* 238:541–563.

Güth, Werner. 1995. "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24:323-44.

Güth, Werner, and Bezalel Peleg. 2001. "When Will Payoff Maximization Survive? An Indirect Evolutionary Analysis," *Journal of Evolutionary Economics*, 11:479-499.

Haldane, John B.S. 1955. "Population Genetics," *New Biology*, 18:34-51.

Hamilton, William D. 1964a. "The Genetical Evolution of Social Behaviour. I." *Journal of Theoretical Biology*, 7:1-16.

Hamilton, William D. 1964b. "The Genetical Evolution of Social Behaviour. II." *Journal of Theoretical Biology*, 7:17-52.

Heifetz, Aviad, Chris Shannon, and Yossi Spiegel. 2006. "The Dynamic Evolution of Preferences," *Economic Theory*, 32:251-286.

Heifetz, Aviad, Chris Shannon, and Yossi Spiegel. 2007. "What to maximize if you must," *Journal of Economic Theory*, 133:31-57.

Hines, W. Gord S., and John Maynard Smith. 1979. "Games between Relatives," *Journal of Theoretical Biology*, 79:19-30.

Holt, C.A. 2007. Markets, Games, & Strategic Behavior. Boston: Pearson.

Koçkesen, Levent, Efe A. Ok, and Rajiv Sethi. 2000. "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory* 92:274-299.

Kyle, Albert S. and F. Albert Wang. 1997. "Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test?" *Journal of Finance*, 52:2073-90.

Levine D. 1998. "Modelling altruism and spite in experiments", *Review of Economic Dynamics,* 1:593-622.

Nakamaru, Mayuko and Simon A. Levin. 2004. "Spread of two linked social norms on complex interaction networks," *Journal of Theoretical Biology,* 230:57–64.

Nash, John. 1953. "Two-Person Cooperative Games," *Econometrica*, 21:128-140.

Nowak, Martin A. 2006. "Five Rules for the Evolution of Cooperation," *Science,* 314: 1560-1563.

Ok, Efe A. and Fernando Vega-Redondo. "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory,* 97:231-254.

Ostrom, E., J. Walker, and R. Gardner. 1994. *Rules, Games, and Common-Pool Resources*. Ann Arbor: University of Michigan Press.

Robson, Arthur J. 1990. "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144:379-396.

Robson, Arthur J., and Larry Samuelson. 2009. "The Evolution of Time Preference with Aggregate Uncertainty," *American Economic Review,* 99:1925-1953.

Rotemberg, Julio J. 1994. "Human Relations in the Workplace," *Journal of Political Economy*, 102:684-717.

Rowthorn, Robert. 2006. "The Evolution of Altruism between Siblings: Hamilton's Rule Revisited," *Journal of Theoretical Biology,* 241:774–790.

Samuelson, Larry and Jeroen Swinkels. 2006. "Information, Evolution and Utility," *Theoretical Economics,* 1:119-142.

Sethi, Rajiv and E. Somanathan. 2001. "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97:273-297.

Tabellini, Guido. 2008. "The Scope of Cooperation: Values and Incentives," *Quarterly Journal of Economics,* 123:905-950.

Tarnita, Corina E., Tibor Antal, Hisashi Ohtsuki, and Martin A. Nowak. 2009. "Evolutionary dynamics in set structured populations," *Proceedings of the National Academy of Sciences*, 106:8601-8604.

Tullock, G. 1980. "Efficient Rent-Seeking," in *Towards a Theory of the Rent-Seeking Society,* ed. by James M. Buchanan, Robert D. Tollison, and Gordon Tullock. College Station: Texas A&M University Press, 97-112.

Trivers, Robert L. 1971. "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, 46:35-57.

Weibull, Jörgen W. and Marcus Salomonsson. 2006. "Natural Selection and Social Preferences," *Journal of Theoretical Biology*, 239:79-92.

Williams, George C. and Doris C. Williams. 1957. "Natural Selection of Individually Harmful Social Adaptations Among Sibs With Special Reference to Social Insects," *Evolution*, 11:32-39.

Wilson, David S. 1977. "Structured Demes and the Evolution of Group-Advantageous Traits," *American Naturalist*, 111:157-185.

Wilson, David S., and Lee A. Dugatkin. 1997. "Group Selection and Assortative Interactions," *American Naturalist*, 149:336-351.

Wright, Sewall G. 1921. "Systems of Mating," *Genetics*, 6:111-178.

Wright, Sewall G. 1922. "Coefficients of Inbreeding and Relationship," *American Naturalist,* 56:330-338.