JOURNAL OF PUBLIC ECONOMIC THEORY
Department of Economics 415 Calhoun Hall Vanderbilt University Nashville, TN 37235 USA (615) 343-8495 (fax)

Submission Number:JPET-09-00145

Public goods games, altruism, and evolution

Ingela Alger
Carleton University

Abstract

I analyze the evolution of altruistic preferences in a population where individuals are matched pairwise to play a one-shot public goods game. I determine the evolutionarily stable degree of altruism, allowing for assortative matching. The stable degree of altruism is strictly smaller than the degree of assortativity. In particular, if matching is completely random, spite is stable, and a positive degree of assortativity is necessary for pure selfishness to be stable. Furthermore, the stable degree of altruism is increasing in the degree of assortativity, and it depends on the specifics of the public goods game.

I wish to thank Sam Bowles, Jörgen Weibull, and two anonymous referees for helpful comments. I am grateful to Carleton University and to the Knut and Alice Wallenberg Research Foundation for financial support, and to the Stockholm School of Economics for its hospitality during my visits there.
Citation: Ingela Alger, (2012) "Public goods games, altruism, and evolution", *Journal of Public Economic Theory*, Vol. 12 No. 4 pp. 789-813.
Contact: Ingela Alger - ingela_alger@carleton.ca.
Submitted: July 27, 2009. **Published:** March 08, 2012.

1 Introduction

Life involves cooperation at most levels: animals hunt together, friends and family help each other, individuals work together to create and produce goods, countries sign trade agreements. Cooperation is ubiquitous, and it is next to impossible to imagine a world without it. Not surprisingly, economists, biologists, political scientists, psychologists, anthropologists, sociologists, and philosophers alike are seeking to understand the forces that allow for cooperation to be sustained. This paper proposes some new theoretical results on the evolution of cooperation among humans.

Humans are peculiar beings, with extraordinary cognitive abilities, as well as an ability to remember the past and imagine the future. As such, they are ideal candidates for one of the leading theories of cooperation. According to this theory, cooperation may be sustained by way of reciprocity in repeated interactions between perfectly selfish individuals (Trivers, 1971, Axelrod and Hamilton, 1981, Fudenberg and Maskin, 1986). Consistent with this theory are experimental findings that humans tend to punish those who fail to cooperate (Fehr and Gächter, 2000, Gächter and Herrmann, 2009). However, there is also substantial evidence that a large fraction of individuals cooperate, against their own material interest, in one-shot interactions (Marwell and Ames, 1979, Gächter et al., 2004, Walker and Halloran, 2004, Cubitt et al., 2008, Gächter and Herrmann, 2009). Such evidence suggests that humans may have a *preference* for cooperative over selfish behavior. Recent research in neuroscience provides support for this hypothesis (Rilling et al., 2002, Moll et al., 2006, Harbaugh et al., 2007, Fehr and Camerer, 2007).

These observations prompt several questions. First, if material welfare affects individual success, can altruistic preferences survive? Second, can spite (negative altruism) also arise? Third, can we predict the level of altruism based on some exogenously given specifics?

In this paper I provide a theory for the endogenous formation of altruistic preferences in a population where individuals are matched pairwise to play a public goods game.[1] This game captures key qualitative features of many common human interactions in the past, such as teamwork in food production, cooperative childrearing, and warfare: whereas the collective of individuals benefits from cooperation in these interactions, each individual would be materially better off by free-riding on the others. The analysis proceeds in two steps. First,

[1] Analysis of data on private transfers conducted by Cox, Hansen, and Jimenez (2004) supports the hypothesis of altruism.

equilibrium behavior and material payoffs are determined for given altruistic preferences. Second, assuming that an individual's material payoff affects his or her reproductive success, I determine evolutionarily stable degrees of altruism.

In the baseline model, two individuals select contributions towards the production of a public good.[2] The individuals are altruistic, in the sense that they care about the other's material welfare. Altruism may be due to fraternal love, or, more generally, to ethics, e.g., in the case of teamwork in the workplace. I also allow for spite (negative altruism). Each individual knows the other's degree of altruism.

Altruism increases an individual's *perceived benefit* from contributing, and hence, an individual's equilibrium contribution is increasing in his or her own altruism (empathy effect of own altruism). By the same token, an individual's equilibrium contribution is decreasing in the other's altruism (free-rider effect of the other's altruism): a higher contribution made by the other lowers an individual's marginal benefit from contributing.

The evolutionary analysis closely follows the methodology developed by Alger and Weibull (2009) to determine *evolutionarily stable degrees of altruism*. This method is reminiscent of standard evolutionary game theory (see, e.g., Weibull, 1995): it checks whether a population consisting of identical individuals (incumbents) would withstand the invasion by a small number of mutants. It deviates from standard evolutionary game theory, however, by endowing each individual with *preferences* rather than with a strategy. Hence, an individual chooses a strategy that maximizes his or her utility, given own and other's degree of altruism. Players are matched pairwise to play the basic game introduced above. An incumbent degree of altruism α is evolutionarily stable against a mutant degree of altruism α', if an individual with altruism α gets a higher expected material payoff than an individual with altruism α'. Matching may be assortative: even though the number of mutants is small, the likelihood that a mutant is matched with another mutant may be significant. An *index of assortativity* measures the probability with which a mutant is matched with another mutant.

It has long been known that assortative matching may be an important factor behind the evolution of *altruistic behavior*. In closely related work, Hamilton (1964a,b) and Bergstrom

[2] A large part of the literature on the evolution of cooperation uses the prisoner's dilemma game, with two strategies. By contrast, I use a general public goods game with a continuous contribution variables. The public goods game may be viewed as a generalized version of the prisoner's dilemma game (Camerer and Fehr, 2004).

(1995)[3] analyze models where related individuals interact, so that the index of assortativity corresponds to *Wright's coefficient of relationship, r* (Wright, 1922), and where individuals are programmed to play a strategy. In such a setting, *Hamilton's rule* applies*:* an altruistic action will be taken if and only if $rb > c$, where c is the reduction of the actor's fitness, and b is the increase in the recipient's fitness. Hence, this rule predicts that individuals will behave *as though* they attached a weight of r to their sibling's material welfare.

By comparison, in the model at hand the stable degree of altruism is *strictly smaller* than the index of assortativity (except in the extreme case of perfect assortativity). In particular, if matching is random, spite is stable. This may seem counterintuitive, since a selfish individual behaves so as to maximize own material welfare. To see why selfishness is not stable, think of a population with selfish incumbents and random matching. Then, a slightly spiteful mutant would almost certainly be matched with an incumbent, who would adjust his or her behavior to the mutant's lower degree of altruism by making a larger contribution than when playing against another incumbent. Hence, if matching is completely random, mutating towards a slightly negative degree of altruism involves a benefit, but no cost.

By the same token, for selfishness to be stable, there must be some cost involved in mutating towards a lower degree of altruism, which requires the mutant to meet another, relatively selfish, mutant with some positive probability. Furthermore, an increase in this probability means that a mutant becomes less likely to meet an incumbent: *ceteris paribus*, this implies that it becomes less beneficial to mutate towards a lower degree of altruism. Accordingly, the stable degree of altruism is strictly increasing in the degree of assortativity.

The model further predicts that, given a level of assortativity, the stable degree of altruism depends on the specifics of the basic game. This is because those specifics determine how strongly a mutant's opponent responds to the mutation at hand. The stronger this response is, the higher is the benefit from mutating towards a lower degree of altruism, and the lower is the stable degree of altruism.

Qualitatively, the results in this paper confirm those derived by Alger and Weibull (2009), who find that the stable degree of altruism is strictly smaller than the degree of assortativity, in a population where individuals are matched pairwise to play a game of risk sharing. By contrast, Bester and Güth (1998), Possajennikov (2000), Bolle (2000) and Heifetz, Shannon, and Spiegel (2006) find that the stable degree of altruism is strictly positive, in a model

[3]Other pieces on the importance of assortative matching include Haldane (1955), Williams and Williams (1957), Wilson (1977), Robson (1990), Grafen (2006), and Nowak (2006).

where players are matched pairwise in a completely random fashion.[4] This result is driven by their assumption that the marginal benefit of a player's action is increasing in the other's action. By contrast, in the model at hand the marginal benefit of a player's contribution is decreasing in the other's contribution.

Sethi and Somanathan (2001) use a game similar to the one at hand, but where a player's altruism may depend on the opponent's altruism. They find that "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. They do not determine evolutionarily stable preferences, however. Bisin, Topa and Verdier (2004), and Tabellini (2008), use a framework of cultural value-transmission to analyze parents' incentives to raise children with a taste for cooperation in a prisoner's dilemma game, with a finite number of strategies. Choi (2008) also studies cultural transmission of cooperation; his focus is on how the frequency of cooperators depends on whether individuals learn and interact at the local or the global level. Weibull and Salomonsson (2006) propose a model of population dynamics that allows for group selection forces (within and between effects). They show how such a model may explain how social preferences, with altruism as an important special case, can emerge from natural selection even in the absence of kinship.

The remainder is organized as follows. In the next section the basic game between two altruistic individuals is analyzed. Section 3 is devoted to determining evolutionarily stable degrees of altruism. Section 4 studies how the evolutionarily stable degrees of altruism depend on the environment. Section 5 provides a discussion, and Section 6 concludes. All the mathematical proofs are in the appendix.

2 A game between mutually altruistic individuals

Two individuals, A and B, play a public goods game. Player i's strategy is his or her contribution $z_i \in [0, \bar{z}]$, $i = A, B$. The amount of output depends on the sum of the contributions, $Z = z_A + z_B$, through a twice differentiable production function $F : [0, 2\bar{z}] \to \mathbb{R}$, where F is strictly increasing and strictly concave, $F' > 0$ and $F'' < 0$, and marginal

[4]Relatedly, Eaton, Eswaran and Oxoby (2009) propose a model to analyze the evolution of altruism in competing groups. Their simulation results indicate that the stable degree of intra-group altruism is higher than the stable degree of inter-group altruism.

product $F'(\bar{z})$ is finite. Individual i incurs a material cost by contributing z_i, described by the twice differentiable function $c : [0, \bar{z}] \to \mathbb{R}$. Marginal cost $c'(z_i) \geq 0$ is strictly increasing in the individual's contribution, $c''(z_i) > 0$ for all $z_i \in [0, \bar{z}]$, and $c'(0) = 0$ and $c'(\bar{z}) = +\infty$.

If i contributes z_i and j contributes z_j, individual i obtains *material welfare*

$$Y(z_i, z_j) = F(z_i + z_j) - c(z_i), \tag{1}$$

and *utility*

$$Y(z_i, z_j) + \alpha_i Y(z_j, z_i), \tag{2}$$

where $\alpha_i \in (-1, L]$ is i's degree of altruism towards j, for some $L \geq 1$.[5] Individuals observe each other's altruism level.

An altruistic individual may truly care about the welfare of the other individual, as would be the case if the game were played by relatives or friends. However, α_i may also measure the extent to which an individual internalizes the external effects of his or her actions, e.g., on a business partner. As such it may reflect a business ethic, which specifies the "right thing to do," and which does not require people to care about each other.

An individual's utility is also his or her payoff in the game. Assume that the players make their contributions simultaneously. Denote by Γ the game thus defined.

2.1 Equilibrium contributions

Given a contribution z_j, the necessary first-order condition for an interior best response $z_i \in (0, \bar{z})$ for individual i is:

$$(1 + \alpha_i) F'(z_i + z_j) = c'(z_i). \tag{3}$$

A Nash equilibrium $(\tilde{z}_A, \tilde{z}_B) \in (0, \bar{z})^2$ must satisfy the following set of first-order conditions:

$$\begin{cases} (1 + \alpha_A) F'(\tilde{z}_A + \tilde{z}_B) = c'(\tilde{z}_A) \\ (1 + \alpha_B) F'(\tilde{z}_A + \tilde{z}_B) = c'(\tilde{z}_B). \end{cases} \tag{4}$$

The next proposition establishes existence and uniqueness of an interior equilibrium.

[5] I will generally refer to α_i as altruism although $\alpha_i < 0$ indicates spite. Degrees of altruism $\alpha_i > 1$ may exist, for instance in parents who are willing to sacrifice their lives for the sake of their children. I rule out the uninteresting case where $\alpha_i \leq -1$ by assumption, since such high levels of spite would lead to the absence of contributions.

Proposition 1 *For each* $(\alpha_A, \alpha_B) \in (-1, L]^2$*, there exists a unique Nash equilibrium* $(\tilde{z}_A, \tilde{z}_B) \in (0, \bar{z})^2$*, where* $\tilde{z}_A \geq \tilde{z}_B \Leftrightarrow \alpha_A \geq \alpha_B$*.*

An individual's degree of altruism determines his or her perceived benefit from contributing. Since the individuals face the same cost, the most altruistic individual makes the largest contribution.

Let $\tilde{z} : (-1, L]^2 \to \mathbb{R}$ be the function that describes the equilibrium contribution of an individual with altruism α playing against an individual with altruism β. This function is implicitly defined by:

$$\begin{cases} (1+\alpha) F' \left(\tilde{z} (\alpha, \beta) + \tilde{z} (\beta, \alpha) \right) = c' \left(\tilde{z} (\alpha, \beta) \right) \\ (1+\beta) F' \left(\tilde{z} (\alpha, \beta) + \tilde{z} (\beta, \alpha) \right) = c' \left(\tilde{z} (\beta, \alpha) \right). \end{cases} \tag{5}$$

Likewise, let $V : (-1, L]^2 \to \mathbb{R}$ be the function that specifies the equilibrium material utility of an individual with altruism α playing against an individual with altruism β:

$$V (\alpha, \beta) = F \left(\tilde{z} (\alpha, \beta) + \tilde{z} (\beta, \alpha) \right) - c \left(\tilde{z} (\alpha, \beta) \right). \tag{6}$$

Below, an index $n = 1, 2$ indicates a partial derivative with respect to the n-th argument for either of these functions.

2.2 Comparative statics

How would the equilibrium contributions described in Proposition 1 change if individuals became more altruistic?

Proposition 2 *Ceteris paribus, an increase in an individual's degree of altruism raises the individual's equilibrium contribution,* $\tilde{z}_1 (\alpha, \beta) > 0$*, and lowers the other's equilibrium contribution,* $\tilde{z}_2 (\beta, \alpha) < 0$*. The sum of the contributions strictly increases:* $\tilde{z}_1 (\alpha, \beta) + \tilde{z}_2 (\beta, \alpha) > 0$*.*

An increase in an individual's own altruism increases his or her incentive to make a contribution for all levels of the other's contribution: there is an *empathy effect* of own altruism. This in turn implies that the other's incentive to contribute declines: there is a *free-rider effect* due to the other's altruism. I illustrate these effects with a numerical example.

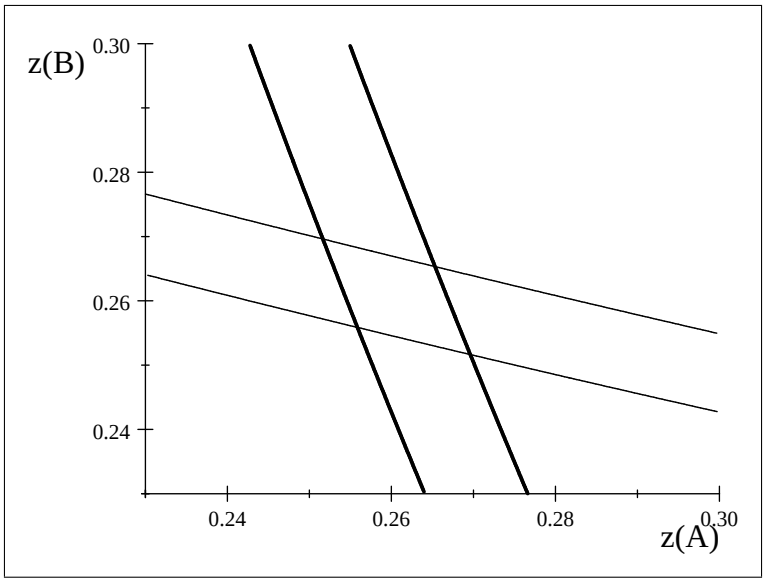

Figure 1: Reaction functions for A (thick curves), and B (thin curves), for $\alpha = 0.4$ (curves close to the origin) and for $\alpha = 0.5$, for $\tau = 0.1$ in Example 1.

Example 1 *Suppose that* $F(Z) = Z^{\tau}$ *for some* $\tau \in (0,1)$*, and that* $c(z) = \frac{1}{2}z^2$*. Then individual* i*'s best response to* j*'s contribution* z_j*, is defined by*

$$\frac{(1+\alpha_i)\,\tau}{(z_i+z_j)^{1-\tau}} = z_i,$$

and equilibrium contributions are defined by

$$\tilde{z}(\alpha,\beta) = \frac{(1+\alpha)\cdot(\tau)^{\frac{1}{2-\tau}}}{(2+\alpha+\beta)^{(1-\tau)/(2-\tau)}}.$$

Figure 1 shows a close-up of the reaction functions for $\tau = 0.1$. The thin curves are the reaction functions of the individual whose contribution is on the vertical axis (say, B), when his altruism level is $\alpha_B = 0.4$, and $\alpha_B = 0.5$, respectively. The thick curves are A's reaction functions, for $\alpha_A = 0.4$ (left curve), and $\alpha_A = 0.5$. If $\alpha_B = 0.4$, an increase in A's altruism from 0.4 to 0.5 causes an outward shift in A's reaction function. This results in A contributing more, and B less. Proposition 2 also says that when one individual becomes more altruistic, the sum of the contributions increases. To see this, note in Figure 1 that the shift in the contributions occurs along B's reaction function; the absolute value of this slope being less than one (see the proof of Proposition 1), the increase in A's contribution outweighs the decline in B's contribution. Intuitively, while the increase in A's altruism induces a first-order change in A's marginal benefit from contributing, the increase in A's contribution causes only a second-order change in B's marginal benefit from contributing.

Proposition 2 implies:[6]

Corollary 1 *Ceteris paribus, if both individuals became more altruistic, they would both increase their contributions.*

2.3 Material welfare

The behavioral changes described in Proposition 2 affects material welfare as follows.

Proposition 3 *An individual always benefits materially from an increase in the other's altruism:* $V_2(\alpha, \beta) > 0$ *for any* $(\alpha, \beta) \in (-1, L]^2$. *An altruistic individual suffers materially from a further increase in own altruism:* $V_1(\alpha, \beta) < 0$, *for any* $(\alpha, \beta) \in [0, L] \times (-1, L]$.

An individual's material welfare would be maximized if he or she were selfish ($\alpha = 0$). Compared to this, an altruistic individual ($\alpha > 0$) makes a material sacrifice, for the benefit of the other individual. An increase in an individual's altruism further strenghtens the individual's sacrifice, as well as the other's benefit. Proposition 3 implies that if both individuals attach a positive weight to the other's material welfare, then the more altruistic one gets a lower material utility: there is a *within-pairs* detrimental effect of altruism on individual material welfare.

How would welfare be affected if both individuals became more altruistic?

Proposition 4 *Suppose the individuals are equally altruistic,* $\alpha_A = \alpha_B = \alpha$. *Equilibrium material welfare is maximized for* $\alpha = 1$, *it is strictly increasing in* α *for any* $\alpha \in (-1, 1)$, *and strictly decreasing in* α *for any* $\alpha \in (1, L]$.

If, in a pair of individuals, the common degree of altruism changes, equilibrium material welfare changes: this may be interpreted as a *between-pairs* effect of altruism on individual material welfare. This effect is positive or negative, depending on whether or not the common degree of altruism is brought closer to $\alpha = 1$, the degree of altruism that maximizes

[6] In a model where two mutually altruistic individuals share risk, and choose risk-reducing efforts, Alger and Weibull (2009) find that an increase in the common level of altruism may cause equilibrium effort to decline. Hwang and Bowles (2009) show that equilibrium contributions in a public goods game with repeated interactions may decline if pure altruism increases, as this reduces the incentive to punish stingy contributors.

equilibrium material welfare. Below it will be shown how the within-pairs and between-pairs effects together affect the evolution of altruism.[7]

2.4 Expected material welfare

Assume now that the players are drawn from a large population, and are matched according to some rule. What is the expected material welfare of an individual with a certain degree of altruism? Here I describe the matching process that will be used in the evolutionary analysis.

Let there be only two degrees of altruism in the population, α and α', and denote by $x \in [0,1]$ the proportion of α-altruists. Although random matching may sometimes be a reasonable assumption, assortative matching may happen naturally for many reasons. Let $P(\alpha|\alpha)$ denote the conditional probability that an α-individual is matched with another α-individual. Then, the expected material payoff of an α-altruist is

$$P(\alpha|\alpha) V(\alpha,\alpha) + [1 - P(\alpha|\alpha)] V(\alpha,\alpha'), \tag{7}$$

where $V(\alpha,\beta)$, $\beta \in \{\alpha,\alpha'\}$, is the material welfare of an individual with altruism α playing against an individual with altruism β (see (6)). Bergstrom (2003) shows that the conditional probability $P(\alpha|\alpha)$ can be derived from the *index of assortativity,* which measures the difference between the conditional probability that an α-individual is matched with another α-individual, and the conditional probability that an α'-individual is matched with an α-individual. Let $\sigma(x) \in [0,1]$ denote this index. Then:

$$\sigma(x) = P(\alpha|\alpha) - P(\alpha|\alpha').$$

The two alternative ways to calculate the number of matches between α-altruists and α'-altruists leads to the following identity:

$$x[1 - P(\alpha|\alpha)] = (1-x) P(\alpha|\alpha').$$

These two equations imply

$$P(\alpha|\alpha) = \sigma(x) + x[1 - \sigma(x)],$$

[7] The within-group and between-group effects of altruism are well-known concepts in evolutionary biology. For recent surveys, see, Sober and Wilson (1998), Bergstrom (2002), and Wilson and Wilson (2007).

and

$$P(\alpha|\alpha') = x[1 - \sigma(x)].$$

Hence, an individual's expected material payoff, as a function of his or her degree of altruism, the index of assortativity $\sigma(x)$, and the population share x of α-individuals, is

$$[\sigma(x) + x(1 - \sigma(x))] V(\alpha, \alpha) + (1 - x)(1 - \sigma(x)) V(\alpha, \alpha') \tag{8}$$

for an α-individual, and

$$[1 - x(1 - \sigma(x))] V(\alpha', \alpha') + x(1 - \sigma(x)) V(\alpha', \alpha) \tag{9}$$

for an α'-individual. These expressions are key in the evolutionary analysis below.

A special case arises when the index of assortativity does not depend on x. This case has been used in analyses by Wright (1921), Cavalli-Sforza and Feldman (1981), Bergstrom (2003), and Alger and Weibull (2009). One can interpret this case as follows: suppose there is a fraction $\sigma \in [0, 1]$ of individuals who are matched with an individual with the same degree of altruism, while the remaining fraction gets a random match. Then, for a randomly drawn α-altruist,

$$P(\alpha|\alpha) = \sigma + x(1 - \sigma).$$

This special case applies, in particular, to games played by relatives. In the following example, it is shown that, for siblings, $\sigma = 1/2$, independent of x. The second example illustrates another instance of assortative matching.

Example 2 ***Teamwork between relatives.*** *Consider a society where grown-up siblings engage in teamwork. Assume that mating is random, and that each child inherits the mother's degree of altruism or that of the father with equal probability (inheritance may be due to education, imitation, or genetics). Suppose that a share x of the adult population carries degree of altruism α', while the remaining population share carries degree of altruism α. A child who has altruism α must have at least one parent with altruism α. If the other parent also has altruism α, which happens with probability $1 - x$, the child's sibling must also have altruism α. If the other parent has altruism α', which happens with probability x, the child's sibling is an α-altruist with probability $1/2$, and an α'-altruist with probability $1/2$. Hence, the probability that an α-altruist's sibling also is an α-altruist is $1 - x/2$. Similarly, a child who is an α'-altruist must have at least one parent with altruism an α'. The probability that the other parent also has altruism an α' is x. Hence, the probability that an α'-altruist's*

sibling has altruism α is $(1/2) - x/2$. In this example, then, the index of assortativity is $\sigma(x) = 1/2$, the coefficient of relationship between siblings (Wright, 1922).[8]

Example 3 ***Teamwork in the workplace.*** *In a workplace, α may be interpreted as the extent to which an employee internalizes the external effect of his behavior on another member of his or her team. A new batch of employees is hired each year. Every new employee is trained by one senior employee; suppose that each junior employee assimilates his or her mentor's attitude towards teamwork (the parameter α in the model). Moreover, assume that, following training, in each group trained by one senior trainer, 80% of the junior employees are kept in that group, while the remaining 20% are randomly assigned to the other groups. Then, in each group thus formed, junior employees are matched pairwise to work together on some task. If each group has n members, the index of assortativity is $\sigma(x) = \frac{0.8n-1}{n-1} - \frac{0.2n-1}{n-1}$ (assuming that $n > 5$).*

3 Evolutionarily stable altruism

Here I closely follow the methodology proposed by Alger and Weibull (2009) to determine evolutionarily stable degrees of altruism.

Consider a sequence of non-overlapping generations. In each generation, individuals are matched pairwise. Each matched pair plays game Γ introduced above once, and each individual obtains the associated equilibrium expected material utility (see (6)). I assume that a higher expected material utility leads to a higher reproductive success, as measured by the expected number of offspring.

Initially the population consists entirely of individuals with some incumbent degree of altruism $\alpha \in (-1, L]$. In the next generation a mutant degree of altruism $\alpha' \neq \alpha$ appears in a small share $\varepsilon > 0$ of the population.[9] The incumbent degree of altruism α is *evolutionarily stable against degree of altruism* α' if an incumbent α-individual gets a higher expected material payoff than a mutant α'-individual, for all sufficiently small ε. The incumbent

[8] For details on how to calculate the index of assortativity with alternative mating and transmission models, I refer to Bergstrom (1995, 2003).

[9] Situations where mutant preferences appear in a large fraction of the population are not impossible, but this model does not allow for them. Oechssler and Riedel (2002) introduce a concept similar to evolutionary stability when mutants may appear in large shares of the population.

degree of altruism α is *evolutionarily stable* if this is true for every $\alpha' \neq \alpha$. As in standard evolutionary game theory, it is assumed that mutations are rare, in the sense that at most one mutant degree of altruism may occur in any given generation.

Focusing on situations with a constant index of assortativity σ, a sufficient condition for the incumbent altruism α to be *evolutionarily stable against* $\alpha' \neq \alpha$ is that

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

The left-hand side of this inequality is (approximately) the expected material payoff of an incumbent, while the right-hand side is the expected material payoff of a mutant, when the share of mutants is close to zero (let the proportion x of the incumbent altruism α go to 1 in expressions (8) and (9)).

A degree of altruism $\alpha \in (-1, L]$ is *locally evolutionarily stable* if inequality (10) holds for all $\alpha' \neq \alpha$ near α, and *evolutionarily stable* if it holds for all $\alpha' \neq \alpha$.

Determining stable degrees of altruism is facilitated by noting that, given the incumbent degree of altruism α, the right-hand side of inequality (10) may be viewed as a function of α'. Hence, inequality (10) says that a degree of altruism α is *evolutionarily stable* if and only if this function reaches its unique global maximum at $\alpha' = \alpha$. Let $\mathcal{A}$ denote the set of degrees of altruism $\alpha \in (-1, L]$ such that $V : (-1, L]^2 \to \mathbb{R}$, defined in (6), is differentiable at the point (α, α). Taking the derivative of the right-hand side of (10) with respect to α' yields the *evolutionary drift function* $D : \mathcal{A} \to \mathbb{R}$, introduced by Alger and Weibull (2009):

$$D(\alpha) = \sigma \cdot \frac{dV(\alpha, \alpha)}{d\alpha} + (1 - \sigma) V_1(\alpha, \alpha). \tag{11}$$

If the incumbent degree of altruism is α, then $D(\alpha)\, d\alpha$ is the effect of a slight increase in a mutant's degree of altruism on the mutant's material payoff. If $D(\alpha) > 0$, a mutant with $\alpha' > \alpha$ gets a higher expected material payoff than an incumbent. Conversely, if $D(\alpha) < 0$, a mutant with $\alpha' < \alpha$ outperforms the incumbents. Stability requires that there be no drift:

Proposition 5 *(Alger and Weibull, 2009) A necessary condition for a degree of altruism $\alpha \in \mathcal{A}$ to be locally evolutionarily stable is $D(\alpha) = 0$. A necessary and sufficient condition for a degree of altruism $\alpha \in int(\mathcal{A})$ to be locally evolutionarily stable is (i)-(iii), where:*

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

(ii) $D(\alpha') > 0$ *for all nearby* $\alpha' < \alpha$

(iii) $D(\alpha') < 0$ *for all nearby* $\alpha' > \alpha$.

The first term in (11) shows how an individual's material utility would be affected by a mutation in altruism, should the mutation be present in both players, an event that happens with probability σ. This *between-pairs* effect of altruism favors drift towards higher degrees of altruism if $\alpha < 1$, and towards lower degrees of altruism if $\alpha > 1$ (see Proposition 4). The second term shows how an individual's material utility would be affected by a mutation in altruism, should the mutation not be present in the other player, an event that happens with probability $1 - \sigma$. This *within-pairs* effect tends to favor drift towards lower degrees of altruism (see Proposition 3).

How do these effects play out with the strategic interaction at hand? Straightforward calculations lead to the following expression for the drift function, when applied to game Γ analyzed in the preceding section.

Lemma 1 *For any $\alpha \in \mathcal{A}$, if players are matched pairwise according to the index of assortativity $\sigma \in [0,1]$, and if each matched pair plays game Γ once, then the evolutionary drift function is*

$$D(\alpha) = F'(2\tilde{z}(\alpha,\alpha))\left[(\sigma - \alpha)\cdot \tilde{z}_1(\alpha,\alpha) + (1 - \sigma\cdot\alpha)\cdot \tilde{z}_2(\alpha,\alpha)\right]. \tag{12}$$

The expression in (12) shows how the change in a mutant's material welfare depends on how the mutant (called M in the following discussion) as well as the mutant's opponent (called O in the following discussion) adjust their behaviors to the mutation. The terms with a factor α represent M's own adjustment. With certainty, M adapts to the change in own altruism ($\tilde{z}_1(\alpha,\alpha)$); with probability σ, O is also a mutant, and M then adapts to the change in O's altruism ($\tilde{z}_2(\alpha,\alpha)$). These adjustments are multiplied by $-\alpha$, for whenever $\alpha \neq 0$, the mutant's behavior diverges from material welfare maximization by a factor α. The other terms represent O's behavioral adjustments: whether O is a mutant or not, he or she adapts to M's degree of altruism ($\tilde{z}_2(\alpha,\alpha)$); with probability σ, O is a mutant, who then adapts to the change in his or her own altruism ($\tilde{z}_1(\alpha,\alpha)$).

Hamilton's rule predicts that natural selection should lead to a degree of altruism equal to the index of assortativity σ. This rule does not apply here, since

$$D(\sigma) = F'(2\tilde{z}(\sigma,\sigma))\left(1-\sigma^2\right)\cdot \tilde{z}_2(\alpha,\alpha) < 0 \tag{13}$$

for any $\sigma < 1$. More precisely:

Proposition 6 *If $\sigma = 1$, then $\alpha^* = 1$ is the unique evolutionarily stable degree of altruism. For any $\sigma \in [0, 1)$, any locally evolutionarily stable degree of altruism $\alpha^* \in \mathcal{A}$ is such that $\alpha^* < \sigma$.*

Equation (13) shows that divergence from Hamilton's rule arises because individuals adapt behavior to their opponent's degree of altruism. To see why Hamilton's rule would obtain absent this adjustment, set $\tilde{z}_2(\alpha, \alpha) = 0$ in (12): then, at the stable degree of altruism $\alpha = \sigma$, a mutant's cost of becoming slightly more altruistic (the term $-\alpha \cdot \tilde{z}_1(\alpha, \alpha)$) would be exactly offset by the benefit of meeting another mutant (the term $\sigma \cdot \tilde{z}_1(\alpha, \alpha)$).

Proposition 6 leads to two notable, and perhaps surprising, observations:

Corollary 2 *If matching is random ($\sigma = 0$), any stable degree of altruism is negative ($\alpha^* < 0$). Furthermore, for purely self-regarding preferences to be evolutionarily stable ($\alpha^* = 0$), there must be positive assortative matching ($\sigma > 0$).*

To see why this happens, consider a society with random matching, and suppose that initially individuals are selfish. Then:

$$D(0) = F'(2\tilde{z}(0, 0)) \cdot \tilde{z}_2(0, 0) < 0. \tag{14}$$

In such a population a slightly spiteful mutant is almost certain to meet a purely selfish incumbent, who would adjust to the mutant's degree of altruism by making a higher effort than against another incumbent. There is a benefit and no cost involved in mutating towards a lower degree of altruism. By the same token, for selfishness to be stable, there must be some cost involved in mutating towards a lower degree of altruism: this requires the between-pairs effect to be at work, which happens only if there is some positive assortative matching.

Does the stable degree of altruism vary in some systematic way with the degree of assortativity? A higher σ means that the between-pairs effect becomes more important relative to the within-pairs effect (see (11)). Mutating towards more selfishness becomes less beneficial, since the likelihood of being matched with another, relatively selfish, mutant increases. Hence:

Proposition 7 *Assume that for some index of assortativity $\sigma < 1$ there exists an evolutionarily stable degree of altruism $\alpha^* < 1$. An increase in σ would lead to an increase in the evolutionarily stable degree of altruism.*

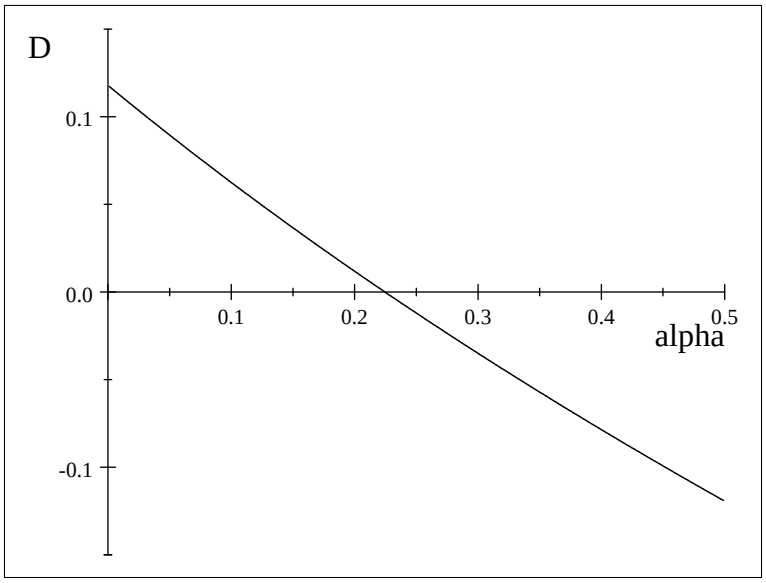

Figure 2: The evolutionary drift $D(\alpha)$.

To illustrate, in the parametric Example 1 the drift function is

$$D(\alpha) = \left(\frac{\tau}{[2(1+\alpha)]^{1-\tau}}\right)^{\frac{1}{2-\tau}} [(\sigma-\alpha)(3-\tau) - (1-\sigma\cdot\alpha)(1-\tau)]. \tag{15}$$

The assumptions $\tau \in (0,1)$ and $\sigma \in [0,1]$ imply that D is strictly decreasing in α, and that it tends towards a strictly negative number as α tends to σ. Figure 2 shows this function for $(\sigma,\tau) = (0.5, 0.1)$. Where Hamilton's rule would predict that altruism be equal to one-half, here the unique stable degree of altruism is smaller than 0.25.

From (15), for any $\tau \in (0,1)$ and $\sigma \in [0,1]$ the unique evolutionarily stable degree of altruism α^* is given by:

$$\alpha^* = \frac{3\sigma - 1 + \tau(1-\sigma)}{3 - \sigma - \tau(1-\sigma)}. \tag{16}$$

Figure 3 shows the stable degree of altruism as a function of the index of assortativity σ, for $\tau = 0.5$. For sufficiently small σ the stable degree of altruism is negative.

4 Altruism and the specifics of the strategic interaction

The preceding analysis shows how the between-pairs and the within-pairs effects of altruism on material welfare together determine the stable degree of altruism. These effects are driven

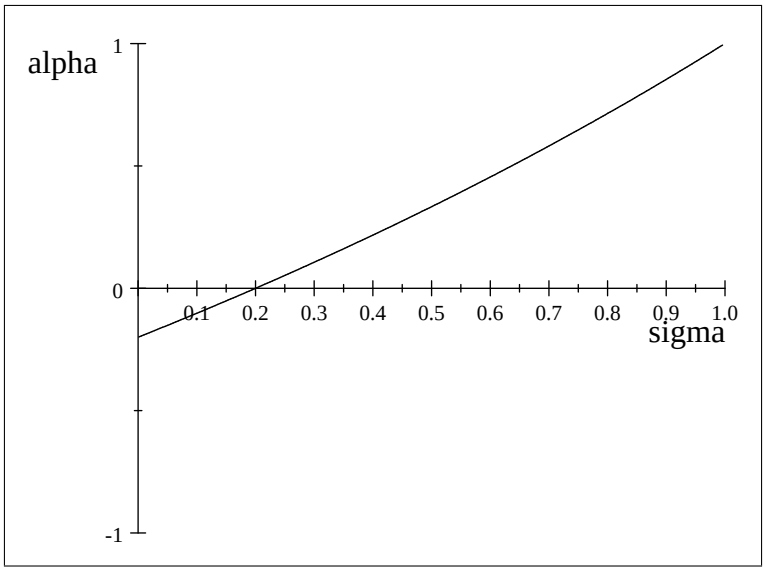

Figure 3: The evolutionarily stable degree of altruism as a function of σ, in Example 1 for $\tau = 0.5$.

by modifications in equilibrium behavior following changes in altruism. Such modifications in turn depend on the specifics of the strategic interaction at hand.

Consider first Example 1, where the stable degree of altruism α^* is given by (16). For a given degree of assortativity σ, the stable degree of altruism is increasing in the production function parameter τ. Figure 4 shows the evolutionarily stable degree of altruism as a function of σ and τ; the leftmost curve is the set of values of (σ, τ) for which $\alpha^* = -0.3$, while the rightmost curve is the set corresponding to $\alpha^* = 0.9$.

In this example the parameter τ measures how quickly the benefits to contributing diminish. It thus also measures the extent to which the marginal product of an individual is affected by the other's contribution. A high value of τ means that the effect is small: the marginal benefit for an individual i of increasing input z_i does almost not depend on the other's input. Hence, τ affects the strength of the behavioral responses to changes in α, and therefore also the stable degree of altruism.

As an illustration, assume that the incumbent degree of altruism is $\alpha = 0.4$, and that a mutant degree of altruism $\alpha' = 0.5$ appears. The reaction functions corresponding to the case $\tau = 0.9$ are illustrated in Figure 5. A pair of incumbents would make contributions at the intersection closest to the origin. Consider now a mutant; suppose it is individual A. If B is an incumbent, A suffers a material loss compared to incumbents: she then contributes more and her opponent contributes less than incumbents do. This is the within-pairs detrimental

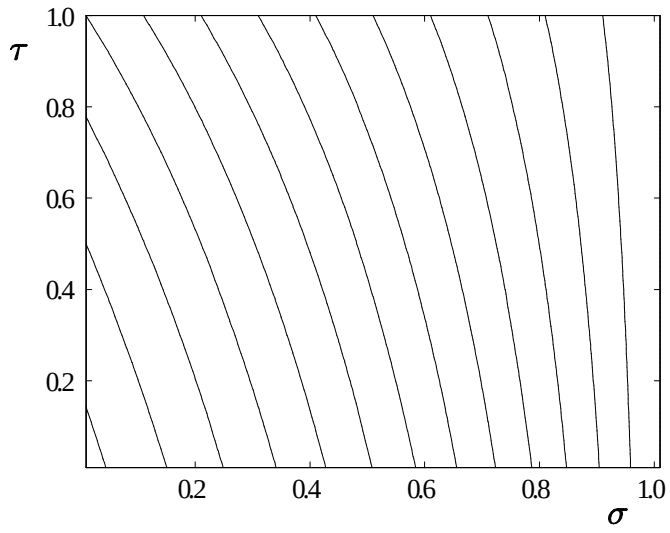

Figure 4: The evolutionarily stable degree of altruism, as a function of σ and τ in Example 1.

effect of own altruism. By contrast, if B is also a mutant, they both contribute more than incumbents do, and they are better off materially. This is the between-pairs beneficial effect of altruism.

In Figure 5 the reaction functions are very flat: whether playing against a mutant or an incumbent, an incumbent makes almost the same contribution. Compare this with Figure 1, which shows the much steeper reaction functions associated with $\tau = 0.1$. When $\tau = 0.1$ an incumbent contributes much less when meeting a mutant, than when meeting an incumbent. Hence, the within-pairs detrimental effect of mutating towards a higher degree of altruism is stronger when $\tau = 0.1$ than when $\tau = 0.9$, and the stable degree of altruism is lower.

The example suggests that the stable degree of altruism depends on the specifics of the strategic interaction, because they affect how strongly individuals respond to changes in one's own and in the other's altruism. The following proposition confirms this, by showing how the evolutionary drift generally depends on the shapes of the production and cost functions.

Proposition 8 *If individuals are matched pairwise to play game Γ, according to the index of assortativity σ, the evolutionary drift may be written in terms of the production and the cost functions as follows:*

$$D(\alpha) = \frac{\left[F'\left(2\tilde{z}\left(\alpha,\alpha\right)\right)\right]^2\left[\left(1-\sigma\right)\left(1+\alpha\right)^2 F''\left(2\tilde{z}\left(\alpha,\alpha\right)\right)+\left(\sigma-\alpha\right)c''\left(\tilde{z}\left(\alpha,\alpha\right)\right)\right]}{c''\left(\tilde{z}\left(\alpha,\alpha\right)\right)\cdot\left[c''\left(\tilde{z}\left(\alpha,\alpha\right)\right)-2\left(1+\alpha\right)F''\left(\tilde{z}\left(\alpha,\beta\right)+\tilde{z}\left(\beta,\alpha\right)\right)\right]}. \qquad (17)$$

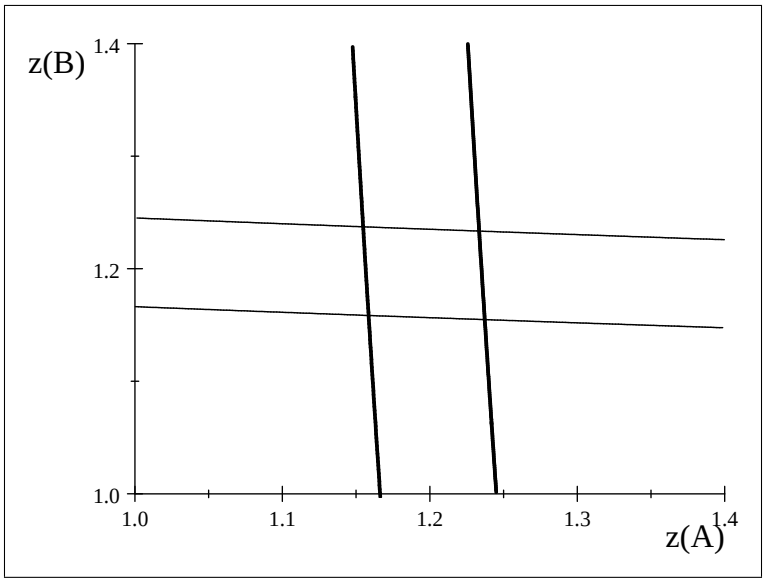

Figure 5: Reaction functions for A (thick curves), and B (thin curves), for $\alpha = 0.4$ (curves close to the origin) and for $\alpha = 0.5$, for $\tau = 0.9$ in Example 1.

Hence, $D(\alpha) = 0$ *if and only if*

$$(1-\sigma)(1+\alpha)^2 F''(2\tilde{z}(\alpha,\alpha)) + (\sigma-\alpha)c''(\tilde{z}(\alpha,\alpha)) = 0.$$

Proposition 8 shows that, together with the degree of assortativity σ, the second derivatives of the production and cost functions, F'' and c'', are crucial in determining the stable degree of altruism.[10] This is because these derivatives affect the strength of the individuals' behavioral responses to changes in one's own and the other's altruism.

5 Conclusion

This paper builds on Alger and Weibull (2009) to determine evolutionarily stable degrees of altruism in a population where individuals are matched pairwise to play a one-shot public goods game. More altruism means that individuals internalize the external effects of their actions to a larger extent, and thus leads to more cooperation. While the evolutionary analysis confirms existing theories about the positive impact of assortative matching on altruism, it also yields two qualitatively different predictions, which confirm results derived

[10] While the magnitude of the marginal product, F' (see (12)), affects the steepness of the drift function, it is inconsequential for the sign of the drift.

in Alger and Weibull (2009). First, the stable degree of altruism is lower than what is suggested by theories relying on selection of strategies rather than preferences (Hamilton, 1964,a,b, Bergstrom, 1995). Second, the stable degree of altruism depends on the shapes of the benefit and cost functions in the public goods game. The reason is that these functions determine an individual's incentive to free-ride on the other. If this incentive is large, then the within-pair effect detrimental effect of own altruism on material welfare is large, and the resulting stable degree of altruism is small.

A growing body of evidence shows significant cross-cultural variability in altruistic behavior (Henrich et al., 2005, Gächter and Herrmann, 2009). Besides the oft-cited multiple social equilibria (see, e.g., Henrich et al., 2005), the results derived above suggest that such cross-cultural variability may stem from exogenously given differences in the conditions in which groups evolved in the past, such as differences in the degree of assortativity, or in the specifics of commonly played strategic interactions. Since cultural values and preferences tend to persist over time,[11] data about a people's ecological past may perhaps help us understand their current cultural values.[12] Because cultural values and preferences in turn affect behavior, such research may ultimately provide a better understanding of economic development.

In the model at hand two individuals with pure altruism interact in a one-shot game. It would be desirable to extend the methodology developed in Alger and Weibull (2009), and used in this paper, in a number of directions, e.g., to games with more than two individuals, to games with repeated interactions, to other classes of games, and to an index of assortativity that depends on the fraction of mutants. Some of these extensions are explored in Alger and Weibull (2010). Furthermore, the method could be extended to study the formation of other preferences than pure altruism, such as reciprocal preferences (Levine, 1998, Sethi and Somanathan, 2001, Weibull, 2004, Hwang and Bowles, 2009), work ethic (Lindbeck and Nyberg, 2006), or a desire for social esteem (Ellingsen and Johannesson, 2008).

---

[11] Using three waves of data representative of 75% of the world's population, Inglehart and Baker (2000) find evidence of persistence of core cultural values. Furthermore, variation in values is higher between than within groups.

[12] Such research is already being conducted in the field of human behavioral ecology. For a survey, see Winterhalder and Smith (2000).

6 Appendix

6.1 Proposition 1

For any $\alpha_i \in (-1, L]$, equation (3) implicitly defines i's best response z_i^R as a function of z_j, $i, j = A, B$, $i \neq j$. The functions F and c being twice differentiable, z_i^R is continuous. Since $F'(z_i + z_j)$ is strictly positive and finite, and since $c'(0) = 0$ and $c'(\bar{z}) = +\infty$, $z_i^R(z_j) \in (0, \bar{z})$ for all $z_j \in [0, \bar{z}]$. Moreover,

$$\frac{dz_i^R}{dz_j} = \frac{(1+\alpha_i) F''(z_i + z_j)}{c''(z_i) - (1+\alpha_i) F''(z_i + z_j)} < 0. \tag{18}$$

These arguments together imply that there exists a solution $(\tilde{z}_A, \tilde{z}_B)$ to the system of equations (4). Since $\left|\frac{dz_i^R}{dz_j}\right| < 1$, the solution is unique.

If $\alpha_A = \alpha_B$ the system of equations (4) implies $c'(\tilde{z}_A) = c'(\tilde{z}_B)$, which, by strict convexity of c implies $\tilde{z}_A = \tilde{z}_B$. Likewise, if $\alpha_i > \alpha_j$, $i = A, B$, $i \neq j$, the system of equations (4) implies $c'(\tilde{z}_i) > c'(\tilde{z}_j)$, which, by strict convexity of c implies $\tilde{z}_i > \tilde{z}_j$.

6.2 Proposition 2

For any z_j, equation (3) defines z_i implicitly as a function of α_i. Applying the implicit function theorem:

$$\frac{dz_i}{d\alpha_i} = \frac{F'(z_i + z_j)}{c''(z_i) - (1+\alpha_i) F''(z_i + z_j)} > 0.$$

Furthermore, $\frac{dz_j}{d\alpha_i} = 0$. Hence, $\tilde{z}_1(\alpha, \beta) > 0$. Since $\frac{dz_i^R}{dz_j} \in (-1, 0)$ (see the proof of Proposition 1), $\tilde{z}_2(\beta, \alpha) < 0$, and $\tilde{z}_1(\alpha, \beta) + \tilde{z}_2(\beta, \alpha) > 0$.

6.3 Proposition 3

Since,

$$V(\alpha, \beta) = F(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha)) - c(\tilde{z}(\alpha, \beta)),$$

the partial derivatives V_1 and V_2 write

$$V_1(\alpha, \beta) = F'(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha)) [\tilde{z}_1(\alpha, \beta) + \tilde{z}_2(\beta, \alpha)] - c'(\tilde{z}(\alpha, \beta)) \tilde{z}_1(\alpha, \beta),$$

and

$$V_2(\alpha, \beta) = F'(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha)) [\tilde{z}_2(\alpha, \beta) + \tilde{z}_1(\beta, \alpha)] - c'(\tilde{z}(\alpha, \beta)) \tilde{z}_2(\alpha, \beta).$$

Using the first-order condition for an α-altruist playing against a β-altruist,

$$(1+\alpha) F'\left(\tilde{z}\left(\alpha,\beta\right)+\tilde{z}\left(\beta,\alpha\right)\right)=c'\left(\tilde{z}\left(\alpha,\beta\right)\right), \tag{19}$$

these expressions may be written

$$V_1\left(\alpha,\beta\right)=F'\left(\tilde{z}\left(\alpha,\beta\right)+\tilde{z}\left(\beta,\alpha\right)\right)\left[\tilde{z}_2\left(\beta,\alpha\right)-\alpha\tilde{z}_1\left(\alpha,\beta\right)\right] \tag{20}$$

and

$$V_2\left(\alpha,\beta\right)=F'\left(\tilde{z}\left(\alpha,\beta\right)+\tilde{z}\left(\beta,\alpha\right)\right)\left[\tilde{z}_1\left(\beta,\alpha\right)-\alpha\tilde{z}_2\left(\alpha,\beta\right)\right]. \tag{21}$$

Recall that $\tilde{z}_1\left(\alpha,\beta\right)>0$, $\tilde{z}_2\left(\beta,\alpha\right)<0$, and $\left|\tilde{z}_1\left(\alpha,\beta\right)\right|>\left|\tilde{z}_2\left(\beta,\alpha\right)\right|$, and $F'>0$. Therefore, $V_1\left(\alpha,\beta\right)<0$ for all $\alpha\in[0,L]$. Furthermore, $V_2\left(\alpha,\beta\right)>0$ for all $\alpha\in(-1,1]$.

Finally, I show that $V_2\left(\alpha,\beta\right)>0$ for $\alpha\in(1,L]$. Let $Y\left(z,v\right)=F\left(z+v\right)-c\left(z\right)$ denote the material utility of an individual who contributes z and whose opponent contributes v. The assumptions on F and c imply that, given v, there exists a unique $z^m\left(v\right)$ that maximizes $Y\left(z,v\right)$; it satisfies

$$F'\left(z^m\left(v\right)+v\right)=c'\left(z^m\left(v\right)\right), \tag{22}$$

$\frac{\partial Y(z,v)}{\partial z}>0$ for all $z<z^m\left(v\right)$, and $\frac{\partial Y(z,v)}{\partial z}<0$ for all $z>z^m\left(v\right)$.

In game Γ between an α-altruist and a β-altruist, the α-altruist's equilibrium contribution satisfies

$$(1+\alpha) F'\left(\tilde{z}\left(\alpha,\beta\right)+\tilde{z}\left(\beta,\alpha\right)\right)=c'\left(\tilde{z}\left(\alpha,\beta\right)\right).$$

Clearly, for any $\alpha>1$, $\tilde{z}\left(\alpha,\beta\right)>z^m\left(\tilde{z}\left(\beta,\alpha\right)\right)$, and $\frac{\partial Y(z,v)}{\partial z}_{|z=\tilde{z}(\alpha,\beta)}<0$. Finally, note that Y is increasing in v. Since $\tilde{z}_1\left(\beta,\alpha\right)>0$ and $\tilde{z}_2\left(\alpha,\beta\right)<0$, the equilibrium material utility of the α-altruist increases as a result of an increase in β ($V_2\left(\alpha,\beta\right)>0$).

6.4 Proposition 4

If $\alpha_A=\alpha_B=\alpha$, the unique symmetric Nash equilibrium $\tilde{z}\left(\alpha,\alpha\right)$ satisfies

$$(1+\alpha) F'\left(2\tilde{z}\left(\alpha,\alpha\right)\right)=c'\left(\tilde{z}\left(\alpha,\alpha\right)\right). \tag{23}$$

Comparison with the equation defining the contribution that maximizes the sum of the material utilities,

$$2F'\left(2z^*\right)=c'\left(z^*\right), \tag{24}$$

yields: $\tilde{z}\left(\alpha,\alpha\right)=z^*\Leftrightarrow\alpha_A=\alpha_B=1$, $\alpha<1\Rightarrow\tilde{z}\left(\alpha,\alpha\right)<z^*$, and $\alpha>1\Rightarrow\tilde{z}\left(\alpha,\alpha\right)>z^*$.

Let $W(z)$ denote individual material utility if both contribute the same amount z:

$$W(z) = F(2z) - c(z).$$

By the assumptions on F and c, W is strictly concave in z. Since W is maximized for $z = z^*$, $W'(z) > 0$ for all $z < z^*$, and $W'(z) < 0$ for all $z > z^*$. Since $\tilde{z}(\alpha, \alpha)$ is strictly increasing in α, the result in the proposition obtains.

6.5 Lemma 1

Since

$$\begin{aligned} D(\alpha) &= \sigma \cdot \frac{dV(\alpha, \alpha)}{d\alpha} + (1 - \sigma) V_1(\alpha, \alpha) \\ &= V_1(\alpha, \alpha) + \sigma \cdot V_2(\alpha, \alpha), \end{aligned}$$

equations (20) and (21) may be used to write

$$\begin{aligned} D(\alpha) &= V_1(\alpha, \beta)_{|\beta=\alpha} + \sigma V_2(\alpha, \beta)_{|\beta=\alpha} \\ &= F'(2\tilde{z}(\alpha, \alpha)) [(\sigma - \alpha) \tilde{z}_1(\alpha, \alpha) + (1 - \sigma\alpha) \tilde{z}_2(\alpha, \alpha)]. \end{aligned}$$

6.6 Proposition 6

I first show that a stable degree of altruism cannot exceed 1. Since

$$D(\alpha) = \sigma \cdot \frac{dV(\alpha, \alpha)}{d\alpha} + (1 - \sigma) V_1(\alpha, \alpha),$$

Propositions 3 and 4 imply $D(\alpha) < 0$ for all $\alpha \in (1, L]$.

Second, suppose that $\sigma = 1$. Then,

$$D(\alpha) = F'(2\tilde{z}(\alpha, \alpha))(1 - \alpha) [\tilde{z}_1(\alpha, \alpha) + \tilde{z}_2(\alpha, \alpha)].$$

Since $F' > 0$ and since $\tilde{z}_1(\alpha, \alpha) + \tilde{z}_2(\alpha, \alpha) > 0$ (see Proposition 2), this is strictly positive for all $\alpha < 1$, and equal to 0 if $\alpha = 1$.

Finally, suppose that $\sigma \in [0, 1)$. Then,

$$D(\alpha) = F'(2\tilde{z}(\alpha, \alpha)) [(\sigma - \alpha) \tilde{z}_1(\alpha, \alpha) + (1 - \sigma\alpha) \tilde{z}_2(\alpha, \alpha)]$$

is strictly negative for any $\alpha \in [\sigma, 1]$, since $(\sigma - \alpha) \tilde{z}_1(\alpha, \alpha) \leq 0$ and $(1 - \sigma\alpha) \tilde{z}_2(\alpha, \alpha) < 0$.

6.7 Proposition 7

Assume that for some functions F, and c, and some $\sigma < 1$, there exists a unique evolutionarily stable degree of altruism; let $\hat{\alpha}$ denote this stable degree of altruism. From Proposition 5, $D(\alpha) > 0$ for all nearby $\alpha < \hat{\alpha}$, and $D(\alpha) < 0$ for all nearby $\alpha > \hat{\alpha}$.

Since $D(\alpha) = V_1(\alpha, \alpha) + \sigma \cdot V_2(\alpha, \alpha)$, and $V_2(\alpha, \alpha) > 0$, an increase in σ leads to an increase in $D(\alpha)$, for every α. Hence, if D is continuous, there exists $\varepsilon > 0$ such that $D(\hat{\alpha} + \varepsilon) = 0$, $D(\alpha) > 0$ for $\alpha < \hat{\alpha} + \varepsilon$, and $D(\alpha) < 0$ for $\alpha > \hat{\alpha} + \varepsilon$.

6.8 Proposition 8

Using the system of equations in (5),and applying the implicit function theorem:

$$\frac{\partial \tilde{z}(\alpha, \beta)}{\partial \alpha} = \frac{\left[c''(\tilde{z}(\beta, \alpha)) - (1+\beta) F''(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha))\right] \cdot F'(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha))}{K}$$

where

$$\begin{aligned} K &= c''(\tilde{z}(\alpha, \beta)) \cdot c''(\tilde{z}(\beta, \alpha)) \\ &\quad -F''(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha)) \cdot [(1+\alpha) c''(\tilde{z}(\beta, \alpha)) + (1+\beta) c''(\tilde{z}(\alpha, \beta))] \end{aligned}$$

and

$$\frac{\partial \tilde{z}(\alpha, \beta)}{\partial \beta} = \frac{(1+\alpha) \cdot F'(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha)) \cdot F''(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha))}{K}.$$

Hence:

$$\begin{aligned} D(\alpha) &= F'(2\tilde{z}(\alpha, \alpha)) \left[(\sigma - \alpha) \tilde{z}_1(\alpha, \alpha) + (1 - \sigma\alpha) \tilde{z}_2(\alpha, \alpha)\right] \\ &= \frac{\left[F'(2\tilde{z}(\alpha, \alpha))\right]^2 \left[(1-\sigma)(1+\alpha)^2 F''(2\tilde{z}(\alpha, \alpha)) + (\sigma - \alpha) c''(\tilde{z}(\alpha, \alpha))\right]}{\left[c''(\tilde{z}(\alpha, \alpha))\right]^2 - 2(1+\alpha) F''(\tilde{z}(\alpha, \beta) + \tilde{z}(\beta, \alpha)) c''(\tilde{z}(\alpha, \alpha))}. \end{aligned}$$

Since $F' > 0$, $F'' < 0$, and $c'' > 0$, the sign of $D(\alpha)$ is determined by the sign of

$$(1-\sigma)(1+\alpha)^2 F''(2\tilde{z}(\alpha, \alpha)) + (\sigma - \alpha) c''(\tilde{z}(\alpha, \alpha)).$$

References

Alger I. and J.W. Weibull (2009) "Kinship, Incentives and Evolution" forthcoming, *American Economic Review*.

Alger I. and J.W. Weibull (2010) "Evolutionarily Stable Altruism" manuscript, Carleton University and Stockholm School of Economics.

Axelrod, R. and W.D. Hamilton (1981) "The Evolution of Cooperation" *Science* **211**, 1390-1396.

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

Bergstrom, T.C. (2002) "Evolution of Social Behavior: Individual and Group Selection" *Journal of Economic Perspectives* **16**, 67-88.

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

Bester, H. and W. Güth (1998) "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization* **34**, 193–209.

Bisin, A., G. Topa, and T. Verdier (2004) "Cooperation as a Transmitted Cultural Trait" *Rationality and Society* **16**, 477-507.

Bolle, F. (2000) "Is Altruism Evolutionarily Stable? And Envy and Malevolence? Remarks on Bester and Güth" *Journal of Economic Behavior and Organization* **42**, 131-133.

Camerer, C.F. and E. Fehr (2004) "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists" in *Foundations of Human Sociality: Experimental and Ethnographic Evidence from 15 Small-scale Societies*, by J. Henrich, R. Boyd, S. Bowles, C.F. Camerer, and E. Fehr, Eds., Oxford University Press: Oxford.

Cavalli-Sforza, L. L. and M. W. Feldman (1981) *Cultural Transmission and Evolution: A Quantitative Approach*, Princeton University Press: Princeton, N.J.

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

Cox, D., B.E. Hansen, and E. Jimenez (2004) "How Responsive are Private Transfers to Income? Evidence from a Laissez-Faire Economy" *Journal of Public Economics* **88**, 2193-2219.

Cubitt, R., M. Drouvelis and S. Gächter (2008) "Framing and Free Riding: Emotional Responses and Punishment in Social Dilemma Games" CeDEx Discussion Paper No. 2008-02.

Eaton, B.C., M. Eswaran, and R.J. Oxoby (2009) " 'Us' and 'Them': The Origin of Identity, and its Economic Implications" University of Calgary Working Paper 2009-03.

Ellingsen, T. and M. Johannesson (2008) "Pride and Prejudice: The Human Side of Incentive Theory" *American Economic Review* **98**, 990-1008.

Fehr, E. and C.F. Camerer (2007) "Social Neuroeconomics: The Neural Circuitry of Social Preferences" *TRENDS in Cognitive Sciences* **11**, 419-426.

Fehr, E. and S. Gächter (2000) "Cooperation and Punishment in Public Goods Experiments" *American Economic Review* **90**, 980-994.

Fudenberg, D. and E. Maskin (1986) "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information" *Econometrica* **54**, 533–554.

Gächter, S. and B. Herrmann (2009) "Reciprocity, Culture and Human Cooperation: Previous Insights and a New Cross-Cultural Experiment" *Philosophical Transactions of the Royal Society B* **364**, 791-806;

Gächter, S., B. Herrmann, and C. Thöni (2004) "Trust, Voluntary Cooperation, and Socio-Economic Background: Survey and Experimental Evidence" *Journal of Economic Behavior and Organization* **55**, 505-531.

Grafen, A. (2006) "Optimization of Inclusive Fitness" *Journal of Theoretical Biology* **238**, 541–563.

Haldane, J.B.S. (1955) "Population Genetics" *New Biology* **18**, 34-51.

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

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

Harbaugh, W.T., U. Mayr, and D.R. Burghart (2007) "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations" *Science* **316**, 1622-1625.

Heifetz, A., C. Shannon, and Y. Spiegel (2006) "The Dynamic Evolution of Preferences" *Economic Theory* **32**, 251-286.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. Smith Henrich, K. Hill, F. Gil-White, M. Gurven, F.W. Marlowe, J.Q. Patton, D. Tracer (2005) " 'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies" *Behavioral and Brain Sciences* **28**, 795–855.

Hwang, S.-H. and S. Bowles (2009) "Is Altruism Bad for Cooperation?" mimeo, U Mass Amherst, and Santa Fe Institute.

Inglehart, R. and W.E. Baker (2000) "Modernization, Cultural Change, and the Persistence of Traditional Values" *American Sociological Review* **65**, 19-51.

Levine, D. (1998) "Modeling Altruism and Spitefulness in Experiments" *Review of Economic Dynamics* **1**, 593-622.

Lindbeck, A. and S. Nyberg (2006) "Raising Children to Work Hard: Altruism, Work Norms and Social Insurance" *Quarterly Journal of Economics* **121**, 1473-1503.

Marwell, G. and R.E. Ames (1979) "Experiments on the Provision of Public Goods" *American Journal of Sociology* **84**, 1335-1360.

Moll, J., F. Krueger, R. Zahn, M. Pardini, R. de Oliveira-Souza, and J. Grafman (2006) "Human Fronto-mesolimbic Networks Guide Decisions about Charitable Donation" *Proceedings of the National Academy of Sciences* **103**, 15623-15628.

Nowak, M.A. (2006) "Five Rules for the Evolution of Cooperation" *Science* **314**, 1560-1563.

Oechssler, J. and F. Riedel (2002) "On the Dynamic Foundation of Evolutionary Stability in Continuous Models" *Journal of Economic Theory* **107**, 223-252.

Possajennikov, A. (2000) "On the Evolutionary Stability of Altruistic and Spiteful Preferences" *Journal of Economic Behavior and Organization* **42**, 125-129.

Rilling, J.K., D.A. Gutman, T.R. Zeh, G. Pagnoni, G.S. Berns, and C.D. Kilts (2002) "A Neural Basis for Social Cooperation" *Neuron* **35**, 395–405.

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

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

Sober, E., and D.S. Wilson (1998) *Unto Others*, Harvard University Press: Cambridge.

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

Trivers, R.L. (1971) "The Evolution of Reciprocal Altruism" *Quarterly Review of Biology* **46**, 35-57.

Walker, J.M. and M.A. Halloran (2004) "Rewards and Sanctions and the Provision of Public Goods in One-Shot Settings" *Experimental Economics* **7**, 235-247.

Weibull, J.W. (1995) *Evolutionary Game Theory,* MIT Press: Cambridge.

Weibull, J. (2004) "Testing Game Theory," in *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality. Essays in Honor of Werner Güth* by S. Huck, Ed., Palgrave Macmillan: New York.

Weibull, J.W. and M. Salomonsson (2006) "Natural Selection and Social Preferences" *Journal of Theoretical Biology* **239**, 79-92.

Williams, G.C. and D.C. Williams (1957) "Natural Selection of Individually Harmful Social Adaptations Among Sibs With Special Reference to Social Insects" *Evolution* **11**, 32-39.

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

Wilson, D.S. and E.O. Wilson (2007) "Rethinking the Theoretical Foundation of Sociobiology" *Quarterly Review of Biology* **82**, 327-348.

Winterhalder, B. and E.A. Smith (2000) "Analysing Adaptive Strategies: Human Behavioral Ecology at Twenty-Five" *Evolutionary Anthropology* **9**, 51-72.

Wright, S. (1921) "Systems of Mating" *Genetics* **6**:111-178.

Wright, S. (1922) "Coefficients of Inbreeding and Relationship" *American Naturalist* **56**, 330-338.