Journal of Theoretical Biology 299 (2012) 42–54

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A generalization of Hamilton's rule—Love others how much?

Ingela Alger [a,*], Jörgen W. Weibull [b,c,d]

[a] Carleton University, Ottawa, Canada
[b] Stockholm School of Economics, Stockholm, Sweden
[c] The Royal Institute of Technology, Stockholm, Sweden
[d] École Polytechnique, Paris, France

ARTICLE INFO

Available online 27 May 2011

Keywords:
Evolutionary stability
Altruism
Spite
Cooperation
Hamilton's rule

ABSTRACT

According to Hamilton's (1964a, b) rule, a costly action will be undertaken if its fitness cost to the actor falls short of the discounted benefit to the recipient, where the discount factor is Wright's index of relatedness between the two. We propose a generalization of this rule, and show that if evolution operates at the level of behavior rules, rather than directly at the level of actions, evolution will select behavior rules that induce a degree of cooperation that may differ from that predicted by Hamilton's rule as applied to actions. In social dilemmas there will be less (more) cooperation than under Hamilton's rule if the actions are strategic substitutes (complements). Our approach is based on natural selection, defined in terms of personal (direct) fitness, and applies to a wide range of pairwise interactions.

1. Introduction

A central concern in theoretical biology is to explain the evolution of cooperative behavior. This is particularly challenging when cooperation requires altruism in the sense that an individual incurs a cost for the benefit of another. Hamilton (1964a,b) provided a key prediction regarding how much altruism one should expect between kin.[1] For the class of one-sided interactions that Hamilton considered, he predicted that a costly act that benefits a relative will be carried out if and only if the fitness cost incurred by the actor is outweighed by the discounted fitness benefit bestowed on the relative, where the discount factor is Wright's coefficient of relatedness. We will refer to this as Hamilton's rule *at the behavioral level*. In this study, we consider a more general class of pairwise interactions and show that when evolution operates at the level of behavior rules (to be made precise below), rather than directly on acts, as is usually assumed, the level of cooperation generally violates Hamilton's rule at the behavioral level.

More specifically, we develop a general paradigm for the evolution of heritable traits in pairwise interactions, where a trait can be any feature of relevance for the fitness consequences of the interaction. We introduce a methodology that allows the researcher to easily identify evolutionarily stable traits. This permits a formulation of Hamilton's rule *at the level of traits*, the level at which evolution occurs. When a trait is an action always to be taken, this generalized rule coincides with Hamilton's rule at the behavioral level. By contrast, for more general traits, it may violate Hamilton's rule at the behavioral level.

Our theoretical model shares many aspects with models that exist in the literature. However, to the best of our knowledge, the way these aspects are here brought together is novel. We begin by defining evolutionary stability of heritable traits in asexually or sexually reproducing populations of finite or infinite size, under pairwise random matching that may, but need not, be uniform. In other words, the population may, but need not, be well mixed. A single parameter, the *index of assortativity*, represents the degree of assortment, or trait correlation, in the random-matching process in question. Roughly speaking, this index reflects the likelihood that an individual with the mutant trait will be matched with another mutant, in population states where mutants are rare. A canonical example of non-uniform random matching is interaction between relatives in the same generation of a sexually reproducing population, in which case the index of assortativity equals Wright's coefficient of relatedness (Wright, 1921, 1922). We formalize a global and a local version of evolutionary stability. Both versions are close in spirit to existing definitions in the literature.[2] While the global version requires robustness to small invasions of *arbitrary* traits—maybe very different from the resident trait—the local version only requires robustness to small invasions of *similar* traits.[3]

* Corresponding author.
E-mail address: ingela_alger@carleton.ca (I. Alger).

[1] In fact, Hamilton's (1964a,b) articles contain two distinct results. Van Veelen (2007) clarifies these and their connection.

[2] See Taylor and Jonker (1978), Taylor and Frank (1996), Day and Taylor (1998), Geritz et al. (1998), Rousset (2004), Gardner et al. (2007), and Van Veelen (2011b).

[3] We represent traits as points in Euclidean (attribute) spaces, so one trait is similar to another if the Euclidean distance between the two attribute-vectors is small.

0022-5193/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2011.05.008

A trait is (locally) evolutionarily stable if a monomorphic population of individuals who all have that trait cannot be invaded by a small population share of mutants who have a (slightly) different trait, where invasion refers to success in terms of personal, or direct, fitness.[4]

We apply this machinery to two concretizations of the abstract notion of a heritable trait. First, we consider the case when evolution operates directly at the level of *behaviors*, by which we mean strategies in symmetric two-player games (with finite or continuum strategy spaces). This is the classical setting in evolutionary game theory, but here allowing for non-uniform random matching. We call this *strategy evolution*. Under quite general conditions, evolutionary stability induces the behavior that would result if each individual strived to maximize a weighted sum of own personal fitness and that of the other individual in the match, where the weight attached to the latter is the index of assortativity. This result confirms earlier results due to Grafen (1979), Bergstrom (1995), and Day and Taylor (1998). For interactions between relatives, this is nothing but Hamilton's rule at the behavioral level: the actor's marginal fitness cost equals the relatedness-discounted marginal fitness benefit to the relative.

We contrast strategy evolution with evolution that operates at a higher level, that of behavior rules, rather than directly on strategies in the game in question. A behavior rule specifies what strategy to use, depending on the game at hand and on the strategy used by the other individual. We assume that each individual's behavior rule is consistent with the maximization of some goal function. It is thus as if nature provided each individual with a goal-seeking behavior rule and delegated the choice of action in the situation at hand to the individual. In order to enable a clear and precise analysis of Hamilton's rule, as formulated at different levels, we focus on the special case when the goal function is a weighted sum of own personal fitness and that of the other individual, where the heritable trait is the weight attached to the other individual's personal fitness. Such a trait can thus be interpreted as the individual's *degree of altruism*, where a positive weight represents *altruism*, a negative weight *spite*, and zero weight *selfishness*.

Such *preference evolution* turns out to induce the same behavior as strategy evolution only in certain interactions. In particular, this behavioral equivalence holds for interactions in which the fitness costs and benefits of an individual's action is independent of others' actions ("no synergies"). However, in many real-life situations, the costs and benefits of an individual's action do depend on others' actions. Under preference evolution, each individual in a match then adapts his or her own action to the other's action. In particular, in a population with a resident and a mutant preference, individuals with the resident behavioral rule may end up taking different actions against mutants than against each other.

For the sake of illustration, consider a social dilemma in which the net value of a further contribution to the public good decreases with the total contribution already made—an arguably realistic feature of many social dilemmas. Examples of situations of this sort are foraging and gathering, as well as defence of a common resource pool. In such interactions, preference evolution leads to less cooperation than prescribed by Hamilton's rule at the behavioral level. The reason for the lower degree of cooperation is that, unlike under strategy evolution, an individual with the resident degree of altruism will increase her contribution to the public good when matched with a less altruistic mutant, in order to partly make up for the mutant's under-provision. Less altruistic mutants can thus more effectively exploit the residents than under strategy evolution, in which case the residents do not change their behavior when faced with a less altruistic mutant. As a second illustration, consider a social dilemma in which the net value of a further contribution instead *increases* as others' contributions increase—such as group hunting of big game (or any collective activity with some division of labor). Then preference evolution again violates Hamilton's rule at the level of behaviors, but now in the opposite direction; less altruistic mutants have a harder time exploiting residents than under strategy evolution. A resident faced with such a mutant will now reduce his or her contribution. The evolutionarily stable level of cooperation is consequently higher than that implied by Hamilton's rule at the behavioral level.[5]

Although Hamilton's rule does not hold at the behavioral level under preference evolution, an increase in the degree of assortativity has the same qualitative effect as under strategy evolution; it leads to more cooperation.[6] At least this is what we find in our parametrically specified social dilemmas. In particular, when applied to interactions between kin, the evolutionarily stable degree of altruism is increasing in relatedness.

Our work, which builds on and extends Alger and Weibull (2010), see also Alger (2010), combines two approaches that have hitherto been explored in two separate literatures. On the one hand, a great deal of attention has been paid to various assortment mechanisms and their consequences for the evolution of altruistic (or spiteful) behavior under strategy evolution, see Hamilton (1964a,b), Wilson (1977), Nowak and May (1992), Nowak and Sigmund (2000), Gardner and West (2004), Rousset (2004), Nowak (2006), Fletcher and Doebeli (2009), Tarnita et al. (2009a,b), and Van Veelen (2009, 2011a,b). Inspired by the seminal work of Hamilton, special attention has been paid to altruistic behavior towards kin, see, e.g., Grafen (1979, 2006), Hines and Maynard Smith (1979), Bergstrom (1995), Taylor and Frank (1996), Day and Taylor (1998), Rousset and Billiard (2000), Rowthorn (2006), and Lehmann and Rousset (2010). On the other hand, preference evolution in infinite populations with uniform random matching has been analyzed by Güth and Yaari (1992), Güth (1995), Bester and Güth (1998), Ok and Vega-Redondo (2001), Dekel et al. (2007), Heifetz et al. (2006, 2007), McNamara et al. (1999), Taylor and Day (2004), and Akçay et al. (2009).[7] The paper that is closest to ours is Van Veelen (2006), who analyzes preference evolution under assortative matching. However, his analysis is restricted to one-sided donor–recipient interactions. He shows that only goal functions that are linear in own and other's personal fitness (as we here assume in our application to preference evolution) can be explained by kinship selection.

The rest of the paper is organized as follows. Section 2 formalizes the matching process, the notions of global and local evolutionary stability of heritable traits, and the nature of the pairwise interaction. In Section 3 we examine the case of strategy evolution and in Section 4 we analyze preference evolution. In Section 5 we illustrate the results by way of examples. The relation of our "perturbation stability" condition and convergence stability (Eshel and Motro, 1981) is discussed in Section 6. Section 7 discusses kin recognition as well as multiple parallel interactions, and Section 8 concludes. Mathematical proofs have been relegated to an Appendix.

[4] Other researchers who also base their methodology on personal fitness rather than inclusive fitness include Day and Taylor (1998) and Nowak et al. (2010); for a discussion, see Taylor et al. (2006).

[5] For other discussions of the validity of Hamilton's rule, see Wenseleers (2006), Van Veelen (2009), and Nowak et al. (2010).

[6] See, e.g., Haldane (1955), Williams and Williams (1957), Wilson (1977), Robson (1990), Nakamaru and Levin (2004), Durrett and Levin (2005), Nowak (2006), and Tarnita et al. (2009a,b).

[7] Behavior rules for continuum action sets have also been used to analyze behavior in repeated interactions by Roberts and Sherratt (1998), Wahl and Nowak (1999), and André and Day (2007).

2. Evolutionary stability of traits

Consider a population of individuals who are matched into pairs. Each pair is engaged in some interaction. Each individual carries some heritable trait $\theta \in T$, where T is a set of potential traits. We here use the word "trait" in a general sense, allowing it to mean anything that may influence the outcome of the interaction, such as an ability, strength, action, signal, rule of conduct, etc. Let $f(\theta,\theta')$ be the increase of the *personal* (*direct*) *fitness* of individuals with trait θ from interaction with an individual with trait θ'. The average effect upon the personal fitness of an individual with a given trait depends on the population distribution of traits and on the matching process. We treat the matching process as exogenous and random, but allow for non-uniform random matching of individuals with respect to their types. For example, it may be more likely that one's match has the same trait. We use the *algebra of assortative encounters* developed by Bergstrom (2003) when analyzing matching processes (see also Eshel and Cavalli-Sforza, 1982, and Van Veelen, 2009, 2011b).

2.1. The matching process

Suppose that only two traits are present in the population, θ and θ', both elements of T. Let $1-\varepsilon$ be the population share of individuals with trait θ and ε the share of individuals with trait θ', where $\varepsilon \in (0,1)$. This defines a *population state* $(\theta,\theta',\varepsilon)$. We will first keep both θ and θ' fixed and vary ε. Let $\sigma(\varepsilon)$ be the difference between the conditional probabilities for an individual to be matched with an individual with trait θ, given that the individual him- or herself either also has trait θ, or, alternatively, trait θ':

$$\sigma(\varepsilon) = \Pr[\theta|\theta,\varepsilon] - \Pr[\theta|\theta',\varepsilon]. \tag{1}$$

This defines an *assortment function* $\sigma : (0,1) \to [-1,1]$ which to each population share ε assigns an *index of assortativity*, $\sigma(\varepsilon)$.

The following equation is a necessary balancing condition for the number of pairwise matches between individuals with traits θ and θ':

$$(1-\varepsilon)\cdot[1-\Pr[\theta|\theta,\varepsilon]] = \varepsilon \cdot \Pr[\theta|\theta',\varepsilon]. \tag{2}$$

Eqs. (1) and (2) together give

$$\begin{cases} \Pr[\theta|\theta,\varepsilon] = \sigma(\varepsilon) + (1-\varepsilon)[1-\sigma(\varepsilon)], \\ \Pr[\theta|\theta',\varepsilon] = (1-\varepsilon)[1-\sigma(\varepsilon)]. \end{cases} \tag{3}$$

Using these expressions, one may write the expected increase in personal fitness, from a match in the population, for an individual with the trait θ, in population state $(\theta,\theta',\varepsilon)$, as

$$F(\theta,\theta',\varepsilon) = [\sigma(\varepsilon) + (1-\varepsilon)[1-\sigma(\varepsilon)]]\cdot f(\theta,\theta) + \varepsilon[1-\sigma(\varepsilon)]\cdot f(\theta,\theta') \tag{4}$$

and likewise for an individual with trait θ':

$$G(\theta,\theta',\varepsilon) = [\sigma(\varepsilon) + \varepsilon[1-\sigma(\varepsilon)]]\cdot f(\theta',\theta') + (1-\varepsilon)[1-\sigma(\varepsilon)]\cdot f(\theta',\theta). \tag{5}$$

We are particularly interested in population states $(\theta,\theta',\varepsilon)$ in which $\varepsilon > 0$ is small, then viewing θ as the *resident* trait, predominant in the population, and θ' as a rare *mutant* trait. The matching process is taken to be such that the index of assortativity, $\sigma(\varepsilon)$, is continuous in ε, with limit $\sigma_0 = \lim_{\varepsilon\to 0}\sigma(\varepsilon)$ as the mutant becomes vanishingly rare. It follows from (3) that $\sigma_0 \in [0,1]$. We clearly have

$$\begin{cases} \lim_{\varepsilon\to 0} F(\theta,\theta',\varepsilon) = f(\theta,\theta), \\ \lim_{\varepsilon\to 0} G(\theta,\theta',\varepsilon) = f(\theta',\theta) + \sigma_0\cdot[f(\theta',\theta') - f(\theta',\theta)]. \end{cases} \tag{6}$$

As a special case, consider an infinite population in which all matches are equally likely. Then $\Pr[\theta|\theta,\varepsilon] = \Pr[\theta|\theta',\varepsilon] = 1-\varepsilon$, and hence $\sigma(\varepsilon) = 0$ for all $\varepsilon \in (0,1)$. We will accordingly refer to matching processes with $\sigma_0 = 0$ as *uniform* random matching, and to processes with $\sigma_0 \neq 0$ as *assortative,* or *non-uniform*, random matching. A canonical example of the latter is sibling interaction in a sexually reproducing population with haploid genetics. For if traits are inherited with equal probability from the two parents, mating probabilities in the parent population are independent of traits, and the parent population is infinite, then $\sigma(\varepsilon) = 1/2$ for all $\varepsilon \in (0,1)$, and thus $\sigma_0 = r = 1/2$, where $r = 1/2$ is Wright's index of relatedness between siblings. (See Appendix for a calculation.) In the same vein, Grafen (1979) argues that for pairwise interactions between relatives with arbitrary index of relatedness $r \in [0,1]$ one obtains $\sigma_0 = r$ (see also Van Veelen, 2011b).[8]

Remark 1. For a finite population of arbitrary size $N > 1$, in which all matches are equally likely, one has $\varepsilon = n/N$ for some positive integer $n \leq N$, $\Pr[\theta|\theta,\varepsilon] = [(1-\varepsilon)N-1]/(N-1)$ and $\Pr[\theta|\theta',\varepsilon] = (1-\varepsilon)N/(N-1)$. Hence, $\sigma(\varepsilon) = -1/(N-1) < 0$. As the population size N tends to infinity, the index of assortativity—not surprisingly—tends to zero. Here the order of limits matters, a subtlety noted in the evolutionary game theory literature by Schaffer (1988). See also Hamilton (1971), Grafen (1985), Wild and Taylor (2004), and Van Veelen (2009).

2.2. Evolutionary stability

For a given matching process, represented by the assortment function σ:[9]

Definition 1. A trait $\theta \in T$ is **evolutionarily stable** if for every trait $\theta' \neq \theta$ there exists some $\bar{\varepsilon} \in (0,1)$ such that for all $\varepsilon \in (0,\bar{\varepsilon})$:

$$F(\theta,\theta',\varepsilon) > G(\theta,\theta',\varepsilon). \tag{7}$$

To see that this indeed is a generalization of the classical ESS concept (Maynard Smith and Price, 1973; Maynard Smith, 1974), consider the case of uniform random matching in an infinite population, which, as was seen above, gives $\sigma(\varepsilon) = 0$ for all $\varepsilon > 0$. In that special case, inequality (7) boils down to

$$(1-\varepsilon)\cdot f(\theta,\theta) + \varepsilon\cdot f(\theta,\theta') > \varepsilon\cdot f(\theta',\theta') + (1-\varepsilon)\cdot f(\theta',\theta). \tag{8}$$

Let T be the mixed-strategy simplex in a finite and symmetric game, and let $f(\theta,\theta')$ be the payoff to any mixed strategy θ when playing against any mixed strategy θ'. Since payoff functions in two-player games are bilinear in the mixed strategies, inequality (8) is equivalent with

$$f(\theta,(1-\varepsilon)\theta + \varepsilon\theta') > f(\theta',(1-\varepsilon)\theta + \varepsilon\theta').$$

In other words, strategy θ should earn a higher payoff (personal fitness) than the mutant strategy, in the post-entry population. To require this inequality to hold for all $\varepsilon > 0$ sufficiently small is precisely the usual definition of an ESS (see, e.g., Weibull, 1995).

Returning to the general case of arbitrary traits, and using the limit equations (6), one obtains that a *necessary* condition for a trait $\theta \in T$ to be evolutionarily stable is that the weak inequality

$$f(\theta,\theta) \geq (1-\sigma_0)\cdot f(\theta',\theta) + \sigma_0\cdot f(\theta',\theta') \tag{9}$$

holds for all traits θ'. In other words, it is necessary that the residents, on average, earn at least as much personal fitness in the interactions as the mutants do, when the latter are virtually absent from the population. A *sufficient* condition for evolutionary

[8] The same result holds for diploid genetics and dominant genes, see Bergstrom (1995, 2003).

[9] This definition is formally identical to Definition 4 in Taylor and Jonker (1978).

stability is that this inequality holds strictly:

$$f(\theta,\theta) > (1-\sigma_0)\cdot f(\theta',\theta)+\sigma_0\cdot f(\theta',\theta') \tag{10}$$

for all traits $\theta' \neq \theta$. We suggest the following definitions:

Definition 2. A trait $\theta \in T$ is *strictly* evolutionarily stable (SES) if inequality (10) holds for all $\theta' \neq \theta$, and θ is *locally strictly* evolutionarily stable (LSES) if (10) holds for all $\theta' \neq \theta$ in some neighborhood of θ.

Note that the right-hand side of inequality (10), viewed as a function of the mutant trait θ', has a strict (local) maximum at $\theta' = \theta$ if and only if θ is (locally) strictly evolutionarily stable. This simple observation provides analytical power when traits are represented as points in a normed vector space. In the following, we will assume that $T \subset \mathbb{R}^k$ for some positive integer k, that the fitness function $f : T^2 \to \mathbb{R}$ is continuous, and that it is continuously differentiable on $\overset{\circ}{(T)}^2$, where $\overset{\circ}{T}$ is the interior of the set T.[10] For interior traits, one can then take the derivative of the right-hand side of (10) with respect to θ', and obtain

$$D(\theta,\theta') = (1-\sigma_0)\cdot \nabla f_1(\theta',\theta)+\sigma_0\cdot[\nabla f_1(\theta',\theta')+\nabla f_2(\theta',\theta')], \tag{11}$$

where ∇f_1 is the gradient of f with respect to the individual's own trait, and ∇f_2 the gradient of f with respect to the other individual's trait.[11] The vector $D(\theta,\theta') \in \mathbb{R}^k$ is the gradient of the (average personal) fitness to the mutant trait θ' in a bimorphic population state with vanishingly few mutants. This vector points in the direction, in trait space, of steepest ascent of the personal fitness of the mutant trait θ'. We will call it the *evolution gradient*, and note that it takes a particularly simple form when the mutant trait is the same as the resident trait:

$$D(\theta,\theta) = \nabla f_1(\theta,\theta)+\sigma_0\cdot\nabla f_2(\theta,\theta). \tag{12}$$

Writing "$\cdot$" for the inner product in $\mathbb{R}^k$ and (boldface) $\mathbf{0}$ for the origin in $\mathbb{R}^k$, the following result follows from standard calculus[12]:

Proposition 1. *Suppose that $T \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$ and let $\theta \in \overset{\circ}{T}$. Condition* (i) *below is necessary for θ to be evolutionarily stable, and for θ to be LSES. Conditions* (i) *and* (ii) *are together sufficient for $\theta \in \overset{\circ}{T}$ to be LSES.*

(i) $D(\theta,\theta) = \mathbf{0}$.

(ii) $(\theta-\theta')\cdot D(\theta,\theta') > 0$ *for all $\theta' \neq \theta$ in some neighborhood of θ.*

The first condition states that the evolution gradient of the mutant should vanish at the resident trait θ; there should be no direction of fitness improvement for a rare mutant at the resident trait. The second condition states that any nearby mutation's evolution gradient should make an acute angle with the direction back towards the resident trait. Hence, if some nearby rare mutant $\theta' \neq \theta$ would arise in a vanishingly small population share, then the mutant's fitness would be increasing in the direction towards the resident trait, θ. We will refer to (i) and (ii) as the *evolutionary stationarity* and *perturbation-robustness* conditions, respectively. (For a discussion of the second condition, in particular its relationship to convergence stability, see Section 6).

In sum: a necessary condition for a trait θ to be locally strictly evolutionarily stable is that it satisfies the evolutionary stationarity condition and a sufficient condition is that it also satisfies the perturbation-robustness condition. We will henceforth focus mainly on LSES, and only when this property has been established, consider its global counterpart, SES. Writing the evolutionary stationarity condition as

$$-\nabla f_1(\theta,\theta) = \sigma_0\cdot\nabla f_2(\theta,\theta) \tag{13}$$

we see that it may be interpreted as a generalized version of *Hamilton's rule at the trait level*. The left-hand side is the marginal fitness cost of trait θ to the carrier and the right-hand side is the marginal fitness benefit of this trait to the other individual, multiplied by the index of assortativity.[13]

2.3. The pairwise interaction

We now apply the above machinery to pairwise interactions represented as symmetric two-player games $\Gamma = (X,\pi)$ in which the set X of strategies (available to each player) lies in some Euclidean space, and $\pi : X^2 \to \mathbb{R}$ is the payoff function. The payoff is taken to represent the incremental effect on personal fitness: $\pi(x,y)$ for the x-strategist and $\pi(y,x)$ for the y-strategist. We assume the payoff function to be continuous, and to be twice continuously differentiable on the interior of its domain. One special case of this is the classical setting of evolutionary game theory, where X is the unit simplex of mixed strategies in a finite and symmetric two-player game. Another special case is when X is a continuum of pure strategies, or actions, in some Euclidean space. We will call the game Γ the *fitness game*.

Hamilton's rule at the behavioral level can be written in the following form:

$$-\nabla\pi_1(x,x) = \sigma_0\cdot\nabla\pi_2(x,x). \tag{14}$$

The left-hand side is the marginal fitness cost of strategy x to the actor and the right-hand side is the marginal fitness benefit to the other individual, multiplied by the index of assortativity.[14]

There are various ways in which a trait may guide behavior. We will consider two possibilities and apply our methodology to each of them in turn. In the first, and classical approach, the hereditary traits are strategies in the fitness game Γ in question. It is thus as if individuals were (genetically or culturally) programmed to strategies that they take against any opponent, irrespective of the opponent's strategy. In the second approach, the heritable traits are behavior rules. Each such rule prescribes a strategy for each strategy that the other individual may use in the game at hand. We assume that paired individuals will use strategies that are mutually compatible in the sense that each individual's strategy agrees with his or her behavior rule, given the other individual's strategy. We call the first approach *strategy evolution* and the second *preference evolution*.

3. Strategy evolution

Suppose that evolution operates directly on strategies. The set of potential traits is $T=X$, the strategy set in some fitness game Γ. It is immaterial for the analysis if the strategies are pure or mixed. The increase in personal fitness for an x-strategist, when matched with a y-strategist, for arbitrary strategies $x,y \in T$, is then identical with the game payoff; $f = \pi$. If x is the resident strategy and $x' \neq x$

[10] In the classical case of evolutionary stability of mixed strategies in symmetric and finite two-player games with m pure strategies, the set T is the unit simplex in $\mathbb{R}^m$, and hence $k=m$. The subsequent calculus also holds in infinite-dimensional normed vector spaces.

[11] More exactly, for any $\tau,\theta \in T$: $\nabla f_1(\tau,\theta) = (\partial f(\tau,\theta)/\partial\tau_1,\ldots,\partial f(\tau,\theta)/\partial\tau_k)$ and $\nabla f_2(\tau,\theta) = (\partial f(\tau,\theta)/\partial\theta_1,\ldots,\partial f(\tau,\theta)/\partial\theta_k)$.

[12] See, e.g., Theorem 2 in Section 7.4 of Luenberger (1969).

[13] For other formulations of a marginal Hamilton's rule, see Taylor and Frank (1996), and Frank (1998).

[14] Queller (1984, 1985) claimed that if the cost of acting altruistically depends on whether or not the recipient also acts altruistically, Hamilton's rule can no longer be stated under the simple form $rb > c$ (see Gardner et al., 2007, p. 219; Van Veelen, 2011a, for reformulations of Hamilton's rule in Queller's model). Our model encompasses such situations of strategic "synergies" in a general manner.

a mutant strategy, condition (10) thus becomes

$$\pi(x,x) > (1-\sigma_0)\cdot\pi(x',x)+\sigma_0\cdot\pi(x',x') \tag{15}$$

the evolution gradient becomes

$$D(x',x) = (1-\sigma_0)\cdot\nabla\pi_1(x',x)+\sigma_0\cdot[\nabla\pi_1(x',x')+\nabla\pi_2(x',x')] \tag{16}$$

and the evolutionary stationarity condition boils down to

$$\nabla\pi_1(x,x)+\sigma_0\cdot\nabla\pi_2(x,x) = \mathbf{0}. \tag{17}$$

Comparing with (14), we note that this is nothing but Hamilton's rule at the behavioral level.

In the special case of uniform random matching ($\sigma_0 = 0$), strict evolutionary stability is equivalent with the requirement that (x,x) be a strict equilibrium of the underlying game, Γ. For finite games with X being the simplex of mixed strategies, this is a well-known sufficient condition for a strategy x to be an ESS. For a matching process with an arbitrary index of assortativity, $\sigma_0 \in [0,1]$, the necessary evolutionary stationarity condition (17) is identical with the necessary first-order condition for Nash equilibrium between two individuals, each of whom strives to maximize the weighted sum of their own personal fitness, $\pi(x,y)$, and that of the other individual, $\pi(y,x)$, the latter given weight σ_0. In other words, it is as if each individual had the goal function

$$u_{\sigma_0}(x,y) = \pi(x,y)+\sigma_0\cdot\pi(y,x), \tag{18}$$

with x for the own strategy and y for that of the other individual, and played a symmetric Nash equilibrium when matched. This is consistent with findings by Grafen (1979), Bergstrom (1995), and Day and Taylor (1998).

4. Preference evolution

We now turn to the second application of our methodology, namely, when the hereditary trait is a behavior rule for some pairwise interaction that takes the form of a symmetric fitness game $\Gamma = (X,\pi)$, as defined in Section 2.3. A *behavior rule* is a correspondence $\varphi : X \rightrightarrows X$ that to each strategy y used by the other individual assigns a non-empty subset $\varphi(y) \subset X$ of strategies to use. We assume that when two individuals with behavior rules φ and ψ are matched, they use strategies $x^* \in X$ and $y^* \in X$ that are mutually compatible in the sense that

$$(x^*,y^*) \in \varphi(y^*) \times \psi(x^*). \tag{19}$$

We do not explicitly model how pairs of individuals find such mutually compatible strategies. However, a possibility is indicated in Fig. 1. The steep curve is (the graph of) a behavior rule φ that assigns a (unique) strategy x, on the horizontal axis, to be used against any strategy y that the opponent might use. The flat curve is, likewise, a behavior rule ψ that assigns a strategy y (on the vertical axis) to use against any strategy x of the opponent. The intersection of the two curves defines a pair of mutually compatible strategies. The thin horizontal and vertical line segments show how this strategy pair can be reached by means of mutual strategy adaptation. This particular "cob web" starts from strategy $y_0 = 1$, lets the player with behavior rule φ choose her "response" strategy $x_1 = \varphi(y_0) \approx 0.595$, whereupon the player with behavior rule ψ adapts her strategy to $y_1 = \psi(x_1)$, etc. Such iterative adaptation quickly leads both players to choose strategies close to the unique intersection (x^*,y^*) between the two curves. Our assumption that any pair of individuals use mutually compatible strategies can thus be thought of as such mutual strategy adaptation occurring on a faster time-scale than the evolutionary adaptation of the behavior rules themselves. See Mohlin (2010) for an analysis of the role of the relative speed of two such adaptation processes. We note that this kind of behavioral adaptation requires no knowledge of the other

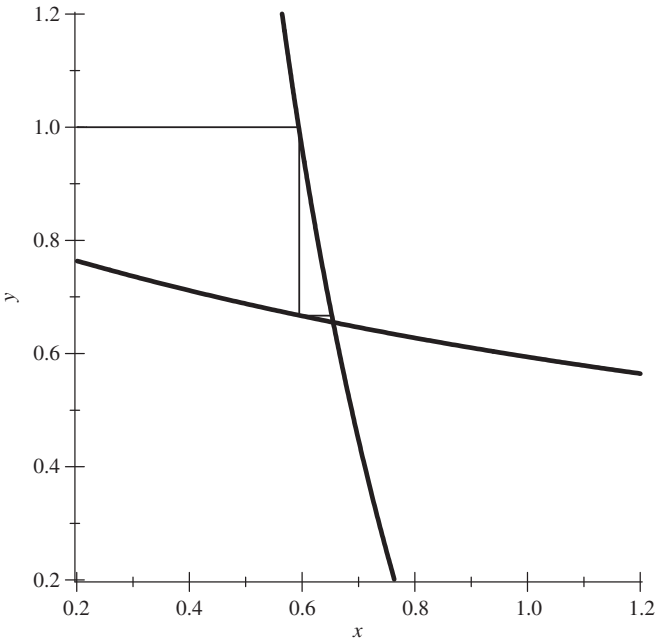

Fig. 1. Convergence to mutually compatible strategies.

individual's behavior rule. We also note that strategy evolution corresponds to the special case when the behavior rule φ is a vertical straight line and the behavior rule ψ is a horizontal straight line.

We will focus on behavior rules such that the product set on the right-hand side in (19) always is a singleton. The so determined strategy pair will be denoted $(x^*(\varphi,\psi),y^*(\psi,\varphi))$. Since both individuals' behavior is uniquely determined, so is the effect on their personal fitness[15]:

$$f(\varphi,\psi) = \pi[x^*(\varphi,\psi),y^*(\psi,\varphi)].$$

The present notion of "behavior rule" being quite general, we impose more structure by restricting attention to *goal-oriented* rules, that is, rules φ such that

$$\varphi(y) = \underset{x\in X}{\operatorname{argmax}} \quad u(x,y) \quad \forall y \in X \tag{20}$$

for some function $u : X^2 \to \mathbb{R}$. Hence, it is *as if* each individual with behavior rule φ would *choose* his or her strategy x so as to maximize the value $u(x,y)$ of some (subjective) goal function, where y is the strategy used by the other individual. We call this approach *preference evolution*, because an individual with behavior rule φ who is matched with an individual who uses strategy y in a fitness game $\Gamma = (X,\pi)$ behaves like an individual who prefers a strategy x over a strategy x' if and only if $u(x,y) > u(x',y)$. In this approach, it is as if nature selects individuals' goal functions (or preferences) and delegates the choice of strategy, in each particular interaction, to the individual, who acts in accordance with the goal function that nature gave him or her. We apply this approach in the next subsection to goal functions expressing potential altruism or spite.

Remark 2. For behavior rules that satisfy (20), the paired individuals in effect play a Nash equilibrium of the (in general asymmetric) two-player game in which each player's strategy set is X and the payoff functions are the goal functions associated with their behavior rules.

[15] More generally, if the product set on the right-hand side of (19) contains more than one strategy pair, $f(\varphi,\psi)$ can be defined as an average over all such pairs.

4.1. Altruism and spite

We noted in Section 3 that strategy evolution leads to the same behavior as when each individual strives to maximize the goal function (18) that attaches weight one to own personal fitness and weight σ_0 to that of the other individual. If this goal function were driving the behavior of all individuals in a population, would it be evolutionarily stable? Suppose that all individuals have goal functions of the form

$$u_\alpha(x,y) = \pi(x,y) + \alpha \cdot \pi(y,x), \tag{21}$$

for some $\alpha \in A = (-1,1)$, and where, as above, $x \in X$ is the individual's strategy, and $y \in X$ is the strategy of the other individual. Such an α- altruist attaches weight one to own personal fitness and weight α to that of the other. A positive α thus expresses *altruism*, a negative α *spite*, and $\alpha = 0$ *selfishness*. For each $\alpha \in A$, the induced behavior rule $\varphi_\alpha : X \rightrightarrows X$ is defined by

$$\varphi_\alpha(y) = \underset{x \in X}{\operatorname{argmax}} \quad [\pi(x,y) + \alpha \cdot \pi(y,x)]. \tag{22}$$

We treat the coefficient $\alpha \in A$ as the hereditary trait, so now $T = A$.

Remark 3. Under strategy evolution we noted that an evolutionarily stable strategy maximizes the goal function in (18). Comparing this with (21), we see that the behavior induced by preference evolution is the same as induced by strategy evolution if the evolutionarily stable degree of altruism equals the index of assortativity of the matching process.

We henceforth assume that $\varphi_\alpha(y) \subset \overset{\circ}{T}$ (the interior of X) for all $\alpha \in A$ and $y \in X$. A necessary condition for a strategy x to belong to the subset $\varphi_\alpha(y)$ then is that the derivative of the goal function in (22), with respect to x, is zero. Formally:

$$x \in \varphi_\alpha(y) \Rightarrow \nabla\pi_1(x,y) + \alpha \cdot \nabla\pi_2(y,x) = \mathbf{0}.$$

Consider a matched pair of individuals, with traits $\alpha \in A$ and $\beta \in A$, respectively. A necessary condition for a strategy pair, $(x,y) \in (\overset{\circ}{X})^2$, to satisfy the mutual compatibility condition (19) is thus that the strategy pair meets the following pair of first-order optimality conditions, one for x, given α and y, and one for y, given β and x:

$$\begin{cases} \nabla\pi_1(x,y) + \alpha \cdot \nabla\pi_2(y,x) = \mathbf{0}, \\ \nabla\pi_1(y,x) + \beta \cdot \nabla\pi_2(x,y) = \mathbf{0}. \end{cases} \tag{23}$$

Let Π^o denote the class of payoff functions π such that, for each $(\alpha,\beta) \in A^2$, there exists a unique pair of strategies that solves (23), where, moreover, this solution is interior and regular.[16] Let $x^*(\alpha,\beta)$ denote the so determined unique strategy of an individual with altruism trait α when matched with an individual with trait β.

The resulting increase in personal fitness is

$$v(\alpha,\beta) = \pi[x^*(\alpha,\beta), x^*(\beta,\alpha)] \tag{24}$$

for the α-altruist, and $v(\beta,\alpha) = \pi[x^*(\beta,\alpha), x^*(\alpha,\beta)]$ for the β-altruist. Hence, in this second application of our general machinery, the function f that maps pairs of traits to personal fitness is identical with this value function v; we now have $f = v$.

Thus, if $\alpha \in A$ is a resident degree of altruism and $\alpha' \neq \alpha$ a mutant degree, condition (10) becomes

$$v(\alpha,\alpha) > (1-\sigma_0) \cdot v(\alpha',\alpha) + \sigma_0 \cdot v(\alpha',\alpha'). \tag{25}$$

[16] A solution is *regular* if the Jacobian of the left-hand side in (23) has non-zero determinant at the solution point. By the implicit function theorem, there then exists a neighborhood of (α,β) and a continuously differentiable function x^* that to each parameter pair (α',β') in this neighborhood assigns the associated (unique) solution to (23).

With subscripts for partial derivatives (the trait space is here one-dimensional), the evolution gradient is

$$D(\alpha',\alpha) = (1-\sigma_0) \cdot v_1(\alpha',\alpha) + \sigma_0 \cdot [v_1(\alpha',\alpha') + v_2(\alpha',\alpha')] \tag{26}$$

and the evolutionary stationarity condition, for a degree of altruism $\alpha^* \in A$, becomes

$$v_1(\alpha^*,\alpha^*) + \sigma_0 \cdot v_2(\alpha^*,\alpha^*) = 0. \tag{27}$$

While stable degrees of altruism may be determined by using Eqs. (24) and (27) directly, the following lemma makes the definition of locally stable degrees of altruism operational in the class of games described above.[17]

Lemma 1. *If $\pi \in \Pi^o$ and $\sigma_0 \in [0,1]$, then*

$$D(\alpha,\alpha) = [(\sigma_0-\alpha) \cdot x_1^*(\alpha,\alpha) + (1-\sigma_0\alpha) \cdot x_2^*(\alpha,\alpha)] \cdot \pi_2[x^*(\alpha,\alpha), x^*(\alpha,\alpha)]. \tag{28}$$

In particular, for fitness games Γ with $\pi \in \Pi^o$ and $\pi_2 \neq 0$, the evolutionary stationarity condition, for a degree of altruism $\alpha^* \in A$, boils down to

$$(\sigma_0-\alpha^*) \cdot x_1^*(\alpha^*,\alpha^*) + (1-\sigma_0\alpha^*) \cdot x_2^*(\alpha^*,\alpha^*) = 0. \tag{29}$$

Below we use this condition to derive a general result on how the stable degree of altruism relates to σ_0.

4.2. Violation of Hamilton's rule at the behavioral level

As noted above, preference evolution leads to the same behavior as strategy evolution if the evolutionarily stable degree of altruism/spite equals the index of assortativity of the matching process. We here proceed to establish general conditions for the evolutionarily stable degree of altruism/spite, under preference evolution, to fall short of, equal and exceed, respectively, the index of assortativity, σ_0. The analysis concerns fitness games $\Gamma = (X,\pi)$ in which the strategy set X is an interval on the real line.

Definition 3. The strategies are *strategically neutral* if $\pi_{12}(x,y) = 0$ for all $x,y \in \overset{\circ}{X}$. The strategies are *strategic substitutes* if $\pi_{12}(x,y) < 0$ for all $x,y \in \overset{\circ}{X}$ and *strategic complements* if $\pi_{12}(x,y) > 0$ for all $x,y \in \overset{\circ}{X}$.

In order to establish the result, it is useful to first consider interactions between individuals with the same degree of altruism/spite. Granted that the other individual's strategy has *some* effect on one's own personal fitness, the effect of a change in the common degree of altruism on the mutually compatible strategy, $x^*(\alpha,\alpha)$, is opposite for strategic substitutes and complements:

Lemma 2. *Suppose that $\pi \in \Pi^o$ and $\pi_2 \neq 0$. Then*

(i) $x_1^*(\alpha,\alpha) \cdot x_2^*(\alpha,\alpha) < 0$ *if the strategies in $\Gamma = (X,\pi)$ are strategic substitutes.*

(ii) $x_1^*(\alpha,\alpha) \cdot x_2^*(\alpha,\alpha) > 0$ *if the strategies in $\Gamma = (X,\pi)$ are strategic complements.*

(iii) $x_1^*(\alpha,\alpha) \neq 0$ *and* $x_2^*(\alpha,\alpha) = 0$ *if the strategies in $\Gamma = (X,\pi)$ are strategically neutral.*

Combining Lemmas 1 and 2 we obtain

Theorem 1. *Suppose that $\pi \in \Pi^o$ and $\pi_2 \neq 0$. If the matching process has index of assortativity σ_0 and a degree α^* of altruism/spite is LSES, then*

1. $\alpha^* < \sigma_0$ *if the strategies in $\Gamma = (X,\pi)$ are strategic substitutes.*
2. $\alpha^* > \sigma_0$ *if the strategies in $\Gamma = (X,\pi)$ are strategic complements.*
3. $\alpha^* = \sigma_0$ *if the strategies in $\Gamma = (X,\pi)$ are strategically neutral.*

[17] For ease of exposition, we here focus on fitness games $G = (X,\pi)$ where $X \subset \mathbb{R}$.

It follows that, in all interactions where strategies are either strategic substitutes or strategic complements, preference evolution leads to different behaviors than strategy evolution, and thus violation of Hamilton's rule at the level of behaviors. Indeed, consider a population consisting of individuals with the "Hamiltonian" degree of altruism $\alpha=\sigma_0$. The evolution gradient would then be

$$D(\sigma_0,\sigma_0)=(1-\sigma_0^2)\cdot x_2^*(\sigma_0,\sigma_0)\cdot\pi_2[x^*(\sigma_0,\sigma_0),x^*(\sigma_0,\sigma_0)].$$

Since $x_2^*\neq 0$, and $\pi_2\neq 0$ by hypothesis, $D(\sigma_0,\sigma_0)\neq 0$, so the necessary stationarity condition for evolutionary stability would be violated if the degree of altruism in the population were σ_0. Evolution operating at the level of behavior rules would thus push the degree of altruism away from its "Hamiltonian" value σ_0.

We now turn to applications in order to illustrate the above results.

5. Applications

We first analyze a class of one-sided interactions of the kind discussed by Hamilton (1964a,b) and thereafter two broad classes of social dilemmas.

5.1. Donor–recipient interactions

Let $\phi(w)$ denote the personal fitness of an individual possessing the amount $w\geq 0$ of a certain resource, where $\phi(w)$ is increasing in w. Suppose an individual holds an amount $w^H>0$ of this resource and has the possibility to transfer a fixed quantity $x\in[0,w^H]$ of this resource to another individual, to whom the coefficient of relatedness is $r\in[0,1]$. With $w^L<w^H$ denoting the other individual's initial possession of the resource, the personal-fitness benefit, b, to the recipient of such a transfer x would be $\phi(w^L+x)-\phi(w^L)$, and the personal-fitness cost, c, to the donor would be $\phi(w^H)-\phi(w^H-x)$. Hamilton (1964a,b) suggested that the transfer x will be made if and only if

$$\phi(w^H)-\phi(w^H-x)<r\cdot[\phi(w^L+x)-\phi(w^L)],$$

that is, if the fitness cost falls short of the "relatedness-discounted" fitness benefit to the recipient.

Suppose now that the donor is free to choose the size x of the transfer, as a continuous variable, as long as $0\leq x\leq w^H$, and that the marginal fitness effect of the resource is decreasing in the amount already held. More exactly, assume ϕ to be differentiable with $\phi'>0$ and $\phi''<0$. According to Hamilton's rule, the donor will give nothing if $\phi'(w^H)\geq r\cdot\phi'(w^L)$, since then any transfer would incur a fitness cost exceeding the "relatedness-discounted" fitness benefit to the recipient. In the more interesting case when $\phi'(w^H)<r\phi'(w^L)$, Hamilton's rule stipulates that the donor will give a positive amount x^* such that his or her marginal fitness cost of increasing the transfer further equals the "relatedness-discounted" marginal fitness benefit to the recipient:

$$\phi'(w^H-x^*)=r\cdot\phi'(w^L+x^*). \tag{30}$$

Given a fitness function, ϕ, and initial resource holdings, w^H and $w^L<w^H$, what would our game-theoretic approach predict? To answer this question, we need to cast the interaction in the form of a symmetric personal-fitness game $\Gamma=(X,\pi)$. For this purpose, imagine that with probability 1/2, player 1's initial wealth is w^H and 2's is w^L, and with probability 1/2 it is the other way round; then player 2's initial wealth is w^H and that of player 1 is w^L. Assume, moreover, that the richer of the two players has the opportunity to transfer any amount of his or her resource holding to the other.[18] For each matched pair of individuals, this interaction takes place only once, so a potential donation is not motivated by any hope of reciprocity. Let x be player 1's transfer if rich and let y be player 2's transfer if rich. These are the players' strategies. We thus have $X=[0,w^H]$ and the payoff function π is defined by

$$\pi(x,y)=\tfrac{1}{2}[\phi(w^H-x)+\phi(w^L+y)].$$

Applying the necessary evolutionary stationarity condition (17) for a strategy x to be evolutionarily stable, we again obtain (30): evolutionary stability under strategy evolution agrees with Hamilton's rule. What about preference evolution? Using the above framework for altruistic/spiteful preferences of the form (21), a rich α-individual chooses the transfer x that maximizes

$$u_\alpha(x,y)=\frac{1}{2}[\phi(w^H-x)+\phi(w^L+y)]+\frac{\alpha}{2}[\phi(w^H-y)+\phi(w^L+x)].$$

Since the optimal transfer x is independent of the other individual's strategy y, there is strategic neutrality; $\pi_{12}(x,y)=0$. Hence, by Theorem 1, $\alpha^*=\sigma_0=r$ is the unique evolutionarily stable degree of altruism. In situations when $\phi'(w^H)<r\cdot\phi'(w^L)$, a rich individual will thus, in equilibrium, choose the transfer level x that satisfies Eq. (30). Hence, in this game also evolutionary stability under preference evolution agrees with Hamilton's rule. This finding is consistent with Van Veelen (2006), who analyzes the evolution of preferences in a population where individuals are faced with similar donor decision problems.

5.2. Social dilemmas when contributions are perfect substitutes

Increasing the maintenance or defence of a common resource pool has less effect when the initial maintenance/defence level is already high. Such diminishing returns occur when the social benefit of the public good in question is a power function of the sum of individual contributions, $B(x+y)\equiv(x+y)^\tau$, for some power $0<\tau\leq 1$. Moreover, individual contributions are then perfect substitutes in the sense that any change in one contribution can be compensated by the opposite and equally large change in the other. We here analyze this case in combination with a quadratic cost function, $C(x)\equiv\kappa\cdot x^2$. In sum

$$\pi(x,y)=(x+y)^\tau-\kappa\cdot x^2 \tag{31}$$

and $X\in\mathbb{R}_+$. It follows that

$$\pi_{12}(x,y)=\tau(\tau-1)(x+y)^{\tau-2}.$$

Hence, the strategies are strategic substitutes if $\tau<1$ and strategically neutral if $\tau=1$. The goal functions in (21) become

$$\begin{cases} u_\alpha(x,y,\pi)=(1+\alpha)(x+y)^\tau-\kappa\cdot(x^2+\alpha y^2),\\ u_\beta(y,x,\pi)=(1+\beta)(x+y)^\tau-\kappa\cdot(y^2+\beta x^2). \end{cases} \tag{32}$$

In this special case, the solution (x,y) lies in the interior of X^2 for all $\alpha\in A$. The condition (23) for mutual compatibility is now both necessary and sufficient, and boils down to

$$\begin{cases} (1+\alpha)\tau(x+y)^{\tau-1}=2\kappa x,\\ (1+\beta)\tau(x+y)^{\tau-1}=2\kappa y. \end{cases} \tag{33}$$

The first equation implicitly defines the behavior rule of the α-altruist (what contribution x to make, given α and y), and the second equation the behavior rule of the β-altruist.

Under strategic neutrality ($\tau=1$), the two optimality conditions in (33) are decoupled: the first equation becomes $1+\alpha=2\kappa x$, thus determining x irrespective of the value of y, and likewise for the second equation. A resident therefore behaves the same way, whether matched with another resident or with a mutant, and irrespective of what others do, just as under strategy evolution. Hence, preference evolution in strategically neutral social dilemmas leads to the same contributions

[18] In the parlance of experimentalists, such an interaction would be called a "random dictator game".

as strategy evolution; Hamilton's rule holds at the behavioral level.

Consider now the arguably more typical case when contributions are strategic substitutes: $\tau<1$. The two optimality conditions in (33) are then coupled: optimality of x (the α-altruist's contribution) depends on the contribution y, and likewise for optimality of y (the β-altruist's contribution). Hence, each individual's behavior is adapted to the other's. Indeed, this is precisely what Fig. 1 shows, for the special case when $\alpha=\beta=\tau=\kappa=0.5$. Fig. 2 shows how these curves change as α and β change. The thin curves represent the case $\alpha=\beta=0.4$, while the thick curves correspond to the case where $\alpha=\beta=0.5$. Each of the four intersections is a mutually consistent pair of actions.

The key difference, as compared with the strategically neutral case ($\tau=1$) is that here a resident behaves differently towards another resident than towards a mutant. To see this, suppose that the resident degree of altruism is $\alpha=0.5$, as would be the case under Hamilton's rule applied directly to the contributions by two siblings, and let the mutant degree of altruism be $\alpha'=0.4$. A pair of residents would make (high) contributions at the intersection of the thick curves, and a pair of mutants would make (low) contributions at the intersection of the thin curves. A pair with one mutant and one resident would make the contributions at the intersection of a thin and a thick curve, and the resident would make a higher contribution when matched with a mutant than when matched with a resident. Compared to the linear case, $\tau=1$, then, there is now an additional benefit of mutating towards a lower degree of altruism. Solving (33) for the contribution made by the α-altruist (when matched with and a β-altruist), we obtain

$$x^*(\alpha,\beta)=\frac{1}{2}\left(\frac{\tau}{\kappa}\right)^{1/(2-\tau)}(1+\alpha)\left(1+\frac{\alpha+\beta}{2}\right)^{-(1-\tau)/(2-\tau)}. \tag{34}$$

Using this expression, it is easily verified that an individual's contribution is increasing in her own degree of altruism and decreasing in the other individual's degree of altruism (as was also seen in Fig. 2 above):

$$x_1^*(\alpha,\beta)=\frac{1}{2}\left(\frac{\tau}{\kappa}\right)^{1/(2-\tau)}\left(1+\frac{\alpha+\beta}{2}\right)^{(\tau-1)/(2-\tau)}\left[1+\left(\frac{\tau-1}{2-\tau}\right)\left(\frac{1+\alpha}{2+\alpha+\beta}\right)\right]>0 \tag{35}$$

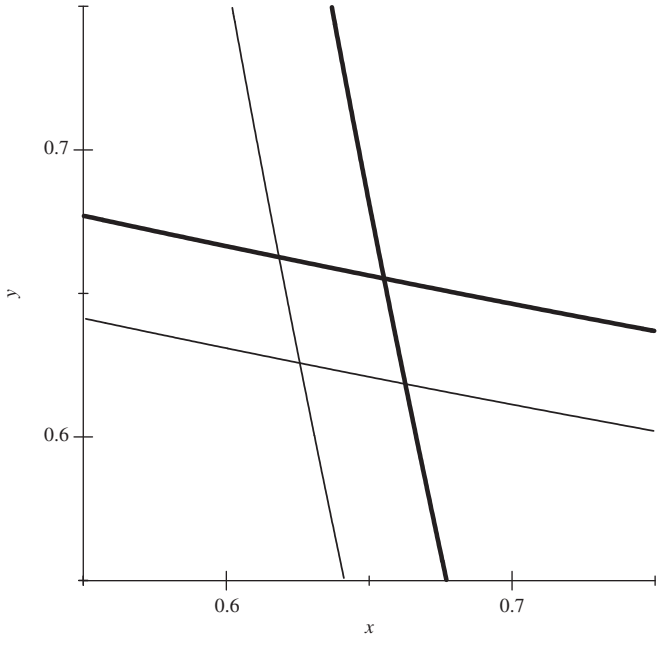

Fig. 2. Best-response curves in a social dilemma with additive contributions ($\tau=\kappa=1/2$).

and (for $\tau<1$)

$$x_2^*(\alpha,\beta)=\frac{1}{2}\left(\frac{\tau}{\kappa}\right)^{1/(2-\tau)}\left(1+\frac{\alpha+\beta}{2}\right)^{(\tau-1)/(2-\tau)}\left(\frac{\tau-1}{2-\tau}\right)\left(\frac{1+\alpha}{2+\alpha+\beta}\right)<0.$$

Evaluating these expressions at $\beta=\alpha$ and dividing through by the common positive factor $(1+\alpha)^{(\tau-1)/(2-\tau)}$, the evolutionary stationarity condition (29) becomes

$$(\sigma_0-\alpha^*)\left(1+\frac{\tau-1}{4-2\tau}\right)+(1-\sigma_0\alpha^*)\frac{\tau-1}{4-2\tau}=0.$$

Solving for α^* gives the unique degree of altruism that can be evolutionarily stable[19]:

$$\alpha^*=\frac{\tau-1+(3-\tau)\sigma_0}{(\tau-1)\sigma_0+3-\tau}. \tag{36}$$

In particular, $\alpha^*\le\sigma_0$, with strict inequality if and only if $\tau<1$. We note that preference evolution and strategy evolution predict the same behavior when strategies are strategically neutral, $\tau=1\Rightarrow\alpha^*=\sigma_0$, and that spite may be evolutionarily stable: α^* is negative if and only if $\sigma_0<(1-\tau)/(3-\tau)$. It also follows from (36) that the evolutionarily stable degree of altruism is increasing in the index of assortativity σ_0.

Fig. 3 shows how the level of cooperation induced by the evolutionarily stable degree of altruism under preference evolution (thick curve) compares to the level of cooperation according to Hamilton's rule applied at the behavioral level, both plotted against the index of assortativity, σ_0 (for $\kappa=1$ and $\tau=0.5$).[20]

5.3. Social dilemmas when contributions are complements

Assume now that the contributions to the public good are complementary in the sense that the more the other contributes, the higher is the return to one's own contribution. Hunting big game could be an example. Formally, let

$$\pi(x,y)=(xy)^\tau-\kappa x^2, \tag{37}$$

where $\tau\in(0,1)$, $\kappa>0$, and $X=\mathbb{R}_+$. We now have $\pi_{12}(x,y)=\tau^2(xy)^{\tau-1}$. Hence, the contributions are strategic complements. An α-individual chooses x so as to maximize

$$\pi(x,y)+\alpha\cdot\pi(y,x)=(1+\alpha)(xy)^\tau-\kappa x^2. \tag{38}$$

The necessary condition (23) now becomes

$$\begin{cases}(1+\alpha)\tau x^{\tau-1}y^\tau=2\kappa x,\\(1+\beta)\tau x^\tau y^{\tau-1}=2\kappa y.\end{cases} \tag{39}$$

The behavior rules for this game, with $\tau=1/4$, $\kappa=1$, are illustrated in Fig. 4 (again, the thick curves are for residents, with altruism $\alpha=0.5$, and the thin curves for mutants, with altruism, $\alpha'=0.4$). As in the previous example, an individual's contribution is increasing in own altruism, $x_1^*(\alpha,\alpha)>0$. However, here an individual increases his or her contribution in response to the opponent's larger contribution; $x_2^*(\alpha,\alpha)>0$. This is because the marginal benefit of contribution now is increasing in the opponent's contribution. Eq. (29) then implies that the stable degree of altruism exceeds σ_0.

Solving (39) for x and y gives

$$x^*(\alpha,\beta)=\left(\frac{\tau}{2\kappa}\right)^{1/2(1-\tau)}(1+\alpha)^{(2-\tau)/4(1-\tau)}(1+\beta)^{\tau/4(1-\tau)}, \tag{40}$$

[19] Numerical simulations suggest that also the perturbation stability condition in Proposition 1 holds, and thus the degree of altruism is in fact evolutionarily stable.

[20] For $\kappa=1$, $\tau=0.5$ and $\sigma_0=0.5$, the stable degree of altruism is $\alpha^*=1/3$. The associated level of cooperation under preference evolution is $x^*(1/3,1/3)\simeq0.38$, while under strategy evolution it would have been $x^*(1/2,1/2)\simeq0.41$.

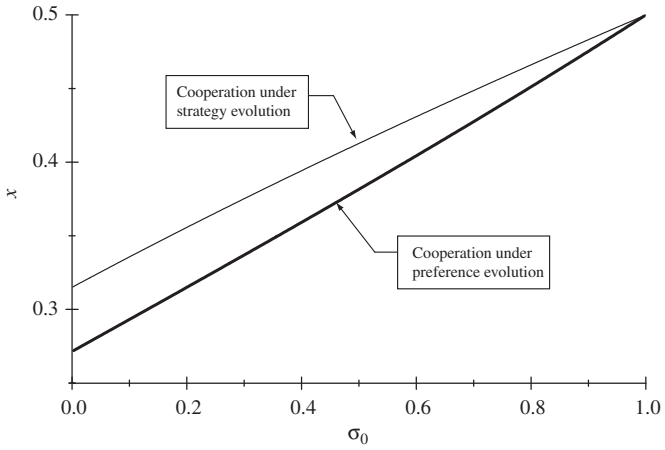

Fig. 3. Cooperation in a social dilemma with strategic substitutes.

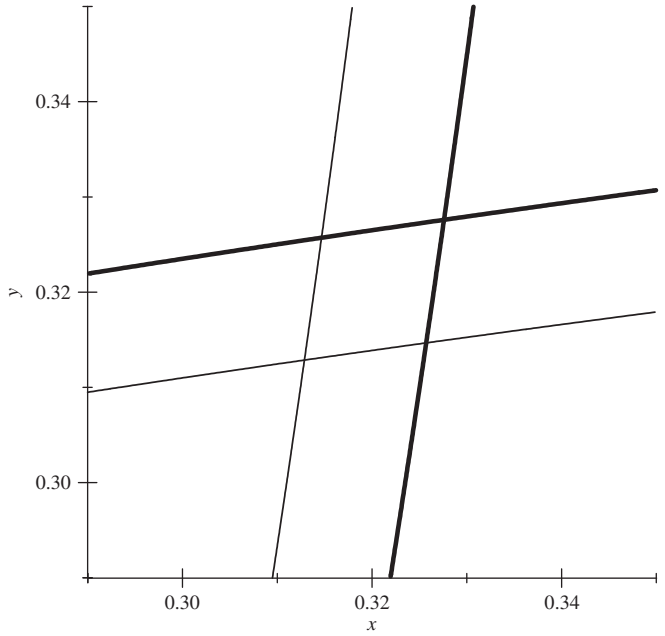

Fig. 4. Best-response curves in a social dilemma with multiplicative contributions ($\tau = 1/4, \kappa = 1$).

so that

$$x_1^*(\alpha,\beta) = \frac{2-\tau}{4(1-\tau)}\left(\frac{\tau}{2\kappa}\right)^{1/2(1-\tau)}(1+\alpha)^{(3\tau-2)/4(1-\tau)}(1+\beta)^{\tau/4(1-\tau)} > 0,$$

and

$$x_2^*(\alpha,\beta) = \frac{\tau}{4(1-\tau)}\left(\frac{\tau}{2\kappa}\right)^{1/2(1-\tau)}(1+\alpha)^{(2-\tau)/4(1-\tau)}(1+\beta)^{(5\tau-4)/4(1-\tau)} > 0.$$

The evolutionary stationarity condition (29) boils down to

$$(\sigma_0-\alpha^*)(2-\tau)+(1-\sigma_0\alpha^*)\tau = 0,$$

with the unique solution[21]

$$\alpha^* = \frac{\tau+(2-\tau)\sigma_0}{\tau\sigma_0+(2-\tau)}. \tag{41}$$

In particular, $\alpha^* \geq \sigma_0$ for all $\tau \in (0,1)$, with strict inequality for all $\sigma_0 < 1$. Here spite is not evolutionarily stable, and a positive degree of altruism is stable under uniform random matching.

[21] Here also numerical simulations confirm that α^* is evolutionarily stable.

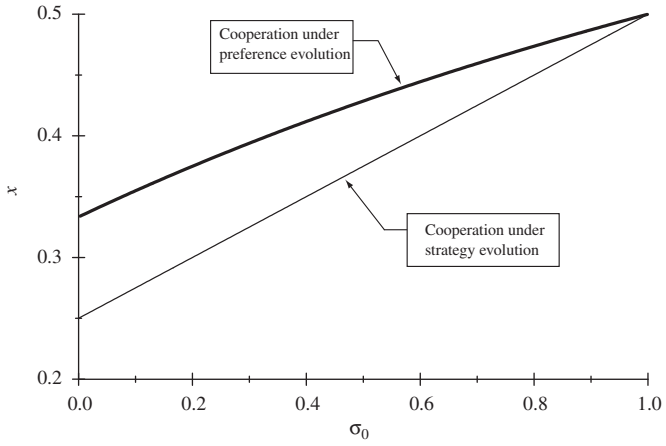

Fig. 5. Cooperation in a social dilemma with strategic complements.

As in the previous example, the evolutionarily stable degree of altruism is increasing in the index of assortativity σ_0.

Fig. 5 shows how the degree of cooperation induced by the evolutionarily stable degree of altruism under preference evolution (thick curve) compares to the degree of cooperation according to Hamilton's rule applied at the behavioral level, both plotted against the index of assortativity, σ_0 (for $\kappa = 1$ and $\tau = 0.5$).

6. Perturbation stability and convergence stability

It has been argued that to be robust, an evolutionarily stable trait should also be convergence stable.[22] A trait $\theta \in \mathbb{R}$ is *convergence stable* if, for some alternative resident trait $\hat{\theta}$ close to θ, mutants between $\hat{\theta}$ and θ fare better than mutants that are further away from θ than $\hat{\theta}$. Here we formalize this notion within our model and compare it with perturbation stability, condition *(ii)* in Proposition 1. Consider, thus, a trait $\hat{\theta}$ in a neighborhood of a given trait $\theta \neq \hat{\theta}$. We will say that there is upward (downward) drift at $\hat{\theta}$, as a resident trait, if a small population share of mutants with a trait slightly above (below) $\hat{\theta}$ fare better than the residents, with trait $\hat{\theta}$, and mutants with a trait slightly below (above) $\hat{\theta}$ fare worse than residents in this population. More precisely, there is *upward drift* at $\hat{\theta}$ if

$$(1-\sigma_0)\cdot f(\hat{\theta}+\delta,\hat{\theta})+\sigma_0\cdot f(\hat{\theta}+\delta,\hat{\theta}+\delta) > f(\hat{\theta},\hat{\theta})$$

for all $0 < \delta < |\theta-\hat{\theta}|$. Likewise, there is *downward drift* at $\hat{\theta}$ if

$$(1-\sigma_0)\cdot f(\hat{\theta}-\delta,\hat{\theta})+\sigma_0\cdot f(\hat{\theta}-\delta,\hat{\theta}-\delta) > f(\hat{\theta},\hat{\theta})$$

for all $0 < \delta < |\theta-\hat{\theta}|$. The following definition is formally equivalent with the conditions for *m*-stability in Taylor (1989) and conditions (4a) and (4b) in Day and Taylor (1998).

Definition 4. A trait $\theta \in T$, for $T \subset \mathbb{R}$ open, is **convergence stable** if there is upward drift at traits $\hat{\theta}$ slightly below θ , *and downward drift at* traits $\hat{\theta}$ slightly above θ.

This property can be characterized in terms of the evolution gradient $D : T^2 \to \mathbb{R}$ defined in Eq. (11):

Proposition 2. *Condition* (i) *below is necessary for a trait* θ *to be convergence stable. Conditions* (i) *and* (ii') *are together sufficient for this property.*

(i) $D(\theta,\theta) = 0$.

(ii') $(\theta-\theta')\cdot D(\theta',\theta') > 0$ *for all* θ' *in a neighborhood of* θ.

[22] Eshel and Motro (1981), Eshel (1983), Taylor (1989), Christiansen (1991), and Day and Taylor (1998).

We note that condition (i) is the evolutionary stationarity condition that was seen to be necessary for the trait θ to be evolutionarily stable. Let $K(t) \equiv D(t,t)$ and assume that condition (i) holds ($K(\theta) = 0$). Condition (ii') then holds if and only if the function K is strictly decreasing at $t = \theta$, a sufficient condition for which is $K'(\theta) < 0$, or, equivalently,

$$\nabla D_1(\theta,\theta) + \nabla D_2(\theta,\theta) < 0.$$

Comparing this with the perturbation stability condition (ii) in Proposition 1, we immediately obtain:

Proposition 3. *If θ is perturbation stable and $\nabla D_2(\theta,\theta) < 0$, then θ is also convergence stable. If $\nabla D_2(\theta,\theta) > 0$ and θ is convergence stable, then θ is also perturbation stable.*

Applying these observations to the above analysis of strategy and preference evolution (and assuming that the fitness function f is twice continuously differentiable), one readily obtains that perturbation and convergence stability, which in general are distinct properties, are in fact identical for interactions in which strategies are strategically neutral.

7. Kin and game recognition

Here we discuss how the methodology developed above applies to quite general settings, allowing for multiple games and degrees of kinship.

Suppose, first, that the interaction always is the same fitness game $\Gamma = (X,\pi)$ but that it is played in a population consisting of relatives with varying coefficients of relatedness between each other. For the sake of concreteness, assume that sometimes a matched pair has a given coefficient of relatedness r, say siblings ($r = 1/2$), and sometimes the two individuals in a pair are completely unrelated ($r = 0$). For any given individual, let $\lambda \in [0,1]$ be the probability that the other individual in a match is a sibling. If siblings always recognize each other, then evolutionary forces may be analyzed as above, separately for pairs of siblings, with $\sigma_0 = 1/2$, and for pairs of completely unrelated individuals, $\sigma_0 = 0$. Each individual may thus behave differently, in the same fitness game, towards siblings than towards unrelated individuals. It is as if two evolutionary selection processes operate in parallel. By contrast, if individuals do not recognize their degree of relatedness, then our machinery still applies, but now with $\sigma_0 = \lambda \cdot 1/2 + (1-\lambda) \cdot 0 = \lambda/2$. Evidently, this reasoning can be extended to any fixed population mixture of relatives of various degrees.

Secondly, suppose that multiple fitness games are being played in the population. If individuals recognize the fitness game at hand in each pairwise match, then our machinery applies to each such fitness game separately. If individuals do not recognize the fitness game, but all games played have the same strategy space X, and are played with given frequencies, then this defines a composite fitness game. For instance, if two fitness games, $\Gamma = (X,\pi)$ and $\Gamma^* = (X,\pi^*)$, are played with frequencies ρ and $1-\rho$, respectively, then the composite fitness game is $\tilde{\Gamma} = (X,\tilde{\pi})$, where $\tilde{\pi} : X^2 \to \mathbb{R}$ is defined by $\tilde{\pi}(x,y) = \rho \cdot \pi(x,y) + (1-\rho) \cdot \pi^*(x,y)$, to which our machinery can be applied.

Thirdly, suppose that multiple fitness games are being played in a population with mixed kinship relations. Again, if individuals always both recognize kin and the fitness game, then the machinery applies to each combination of game and kinship relation separately. It may then be that siblings, for example, treat each other with different degrees of altruism in different interactions. In particular, in strategic-substitutes games our prediction is that the same siblings will behave less altruistically towards each other than in strategic-complements games. This is of course not the case if individuals do not recognize the fitness game. Nevertheless, our methodology can be applied also if individuals do not recognize the game or their degree of relatedness, as long as the frequencies of kin-game combinations are stable.

8. Conclusion

In this article, we develop a general paradigm for the evolution of heritable traits in pairwise interactions, where a trait can be any feature of relevance for the fitness consequences of the interaction. For example the trait may be ability, strength, talent, information processing, a signal, an action to take, etc. We define evolutionary stability and introduce an evolution gradient that allows the researcher to easily identify evolutionarily stable traits. The requirement that this gradient vanishes is shown to be a generalization of Hamilton's rule. We apply this paradigm to *actions* and *behavior rules*, where a behavior rule is a prescription how to adapt one's action to the other party's action. In the latter case, we assume that a matched pair of individuals quickly reach a pair of mutually compatible actions, and we focus on behavior rules that express altruism or spite towards the other individual. We call the first application *strategy evolution* and the second *preference evolution*. Following Hamilton (1964a,b), the standard approach has been to assume strategy evolution.

We show how and why strategy evolution and preference evolution may lead to different behaviors. In a class of interactions, which we call strategically neutral (and which includes one-sided interactions and linear public-goods games), evolutionary stability under preference evolution leads to the same behavioral predictions as in Hamilton's model. By contrast, in interactions where actions are strategic substitutes (complements), preference evolution leads to other predictions. More specifically, preference evolution leads to less cooperation than strategy evolution when actions are strategic substitutes, while it leads to more cooperation when actions are strategic complements.

Moreover, our approach can be applied to interactions that sometimes occur between close relatives and sometimes between distant relatives or even complete strangers. If individuals recognize their degree of relatedness, then the effect of evolution can be analyzed by applying our approach separately to each coefficient of relatedness. If relatives do not recognize their relatedness, then the effect of evolution can be analyzed with our approach by scaling the index of assortativity of the matching process accordingly.

Our analysis may be extended in many directions. While preference evolution, as modelled here, presumes that interacting individuals fully adapt their strategies to each other, in accordance with their behavior rules, strategy evolution assumes the complete lack of such adaptation. Full strategy adaptation is formally identical with Nash equilibrium play under complete information. Analyses of varying degrees of incomplete information, in the pairwise interactions, would seem relevant for understanding many strategic interactions. For modelling approaches in such directions and discussions, see Ok and Vega-Redondo (2001), Heifetz et al. (2006), Dekel et al. (2007), Gärdenfors (2008) and West et al. (forthcoming). Moreover, while we here focus on games in normal form with finite-dimensional strategy spaces, many interactions are dynamic and sometimes repeated, which may require infinite-dimensional strategy spaces.[23] Generalizations to such interactions seem important for many applications. Another aspect is that the present analysis

[23] Mixed-strategy spaces are finite-dimensional in any finite extensive-form game, so the present approach applies to all such symmetric games, including multi-stage games with a finite number of actions in each period, etc.

is restricted to pairwise interactions, while in practice interactions often involve more individuals. Our approach readily generalizes to symmetric strategic interactions with an arbitrary number of participants. We are currently working on such a generalization, and also on generalizations to other kinds of social preferences than altruism/spitefulness, such as inequity aversion (Fehr and Schmidt, 1999) and conditional altruism, that is, altruism that partly depends on the other party's altruism (Levine, 1998; Sethi and Somanathan, 2001). Preliminary calculations point to results that could provide evolutionary foundations for some of the social preferences suggested in behavioral economics. Finally, an exciting and challenging direction for future research would be to apply the present methodology to more complex interaction and matching schemes (such as in Rousset, 2004).

Acknowledgments

We thank Peter Gärdenfors, Andy Gardner, Alan Grafen, Laurent Lehmann, François Rousset, Matthijs van Veelen, and Stuart West for comments to earlier drafts, and the Knut and Alice Wallenberg Research Foundation for financial support. Ingela Alger thanks Carleton University and SSHRC for financial support and the Stockholm School of Economics for its hospitality.

Appendix A

A.1. The index of assortativity for siblings

We here substantiate the claim that $\sigma(\varepsilon)=r=1/2$ for all $\varepsilon\in(0,1)$ for siblings in a sexually reproducing population with haploid genetics, where traits are inherited with equal probability from the two parents, mating probabilities in the parent population are independent of traits, and the parent population is infinite.[24]

Suppose that the population share $1-\varepsilon$ in the parent population has trait θ, while the remaining population share, ε, has trait θ'. Assume also that the type distribution is the same among males and females, and that all couples are equally likely to form.[25] There are four types of couples in this population: in a proportion $(1-\varepsilon)^2$ of couples both are θ, in a proportion ε^2 both parents are θ', and in a proportion $2\varepsilon(1-\varepsilon)$ one parent is θ while the other is θ'. In the latter case, each offspring has probability 1/2 to inherit any given trait. At least one of an offspring's parents must have the same trait as the offspring. Consider first a θ-offspring. The probability that both its parents have trait θ is

$$p=\frac{(1-\varepsilon)^2}{(1-\varepsilon)^2+2(1-\varepsilon)\varepsilon/2}=1-\varepsilon.$$

In this case also the sibling has trait θ. Moreover, since both parents cannot have the other trait θ', the probability that exactly one of them has trait θ is $1-p=\varepsilon$. In this case, the probability that the sibling has trait θ is one half. Thus

$$\Pr[\theta|\theta,\varepsilon]=1-\varepsilon+\varepsilon/2=1-\varepsilon/2.$$

Secondly, consider a θ'-offspring. The probability that both parents have trait θ' is

$$q=\frac{\varepsilon^2}{\varepsilon^2+2(1-\varepsilon)\varepsilon/2}=\varepsilon.$$

[24] We thank François Rousset for spotting an error in our initial calculations.

[25] The following calculations can be adapted to situations in which the type distributions differ among males and females and where mating probabilities are type dependent.

The probability that exactly one of them has trait θ' is thus $1-q=1-\varepsilon$. Therefore, the probability that the sibling has trait θ' is

$$\Pr[\theta'|\theta',\varepsilon]=\varepsilon+(1-\varepsilon)/2=(1+\varepsilon)/2.$$

Consequently,

$$\Pr[\theta|\theta',\varepsilon]=1-\Pr[\theta'|\theta',\varepsilon]=(1-\varepsilon)/2.$$

Putting this together, we obtain $\sigma(\varepsilon)=1-\varepsilon/2-(1-\varepsilon)/2=1/2$.

A.2. Lemma 1

For any fitness game $\Gamma=(X,\pi)$ with $\pi\in\Pi^o$, and traits $\alpha,\beta\in(-1,1)$:

$$v_1(\beta,\alpha)=\pi_1[x^*(\beta,\alpha),x^*(\alpha,\beta)]\cdot x_1^*(\beta,\alpha)+\pi_2[x^*(\beta,\alpha),x^*(\alpha,\beta)]\cdot x_2^*(\alpha,\beta), \tag{44}$$

and

$$v_2(\beta,\alpha)=\pi_1[x^*(\beta,\alpha),x^*(\alpha,\beta)]\cdot x_2^*(\beta,\alpha)+\pi_2[x^*(\beta,\alpha),x^*(\alpha,\beta)]\cdot x_1^*(\alpha,\beta). \tag{45}$$

Using the β-altruist' s first-order condition (see (23)),

$$\pi_1[x^*(\beta,\alpha),x^*(\alpha,\beta)]+\beta\cdot\pi_2[x^*(\alpha,\beta),x^*(\beta,\alpha)]=0. \tag{46}$$

(44) and (45) may be re-written

$$v_1(\beta,\alpha)=-\beta\cdot\pi_2[x^*(\alpha,\beta),x^*(\beta,\alpha)]\cdot x_1^*(\beta,\alpha)+\pi_2[x^*(\beta,\alpha),x^*(\alpha,\beta)]\cdot x_2^*(\alpha,\beta)$$

and

$$v_2(\beta,\alpha)=-\beta\cdot\pi_2[x^*(\alpha,\beta),x^*(\beta,\alpha)]\cdot x_2^*(\beta,\alpha)+\pi_2[x^*(\beta,\alpha),x^*(\alpha,\beta)]\cdot x_1^*(\alpha,\beta)$$

so that

$$v_1(\alpha,\alpha)=[x_2^*(\alpha,\alpha)-\alpha\cdot x_1^*(\alpha,\alpha)]\cdot\pi_2[x^*(\alpha,\alpha),x^*(\alpha,\alpha)] \tag{47}$$

and

$$v_2(\alpha,\alpha)=[x_1^*(\alpha,\alpha)-\alpha\cdot x_2^*(\alpha,\alpha)]\cdot\pi_2[x^*(\alpha,\alpha),x^*(\alpha,\alpha)], \tag{48}$$

which, together with (27), imply (28).

A.3. Lemma 2

Consider a fitness game $\Gamma=(X,\pi)$ with $\pi\in\Pi^o$. For any $\alpha,\beta\in(-1,1)$, let $(x^*(\alpha,\beta),x^*(\beta,\alpha))$ be the unique solution to the system of Eq. (23):

$$\begin{cases}\pi_1[x^*(\alpha,\beta),x^*(\beta,\alpha)]+\alpha\cdot\pi_2[x^*(\beta,\alpha),x^*(\alpha,\beta)]=0,\\ \pi_1[x^*(\beta,\alpha),x^*(\alpha,\beta)]+\beta\cdot\pi_2[x^*(\alpha,\beta),x^*(\beta,\alpha)]=0.\end{cases}$$

By the implicit function theorem,

$$\frac{\partial x^*(\alpha,\beta)}{\partial\alpha}=\frac{-[\pi_{11}[x^*(\beta,\alpha),x^*(\alpha,\beta)]+\beta\pi_{22}[x^*(\alpha,\beta),x^*(\beta,\alpha)]]\pi_2[x^*(\beta,\alpha),x^*(\alpha,\beta)]}{K}$$

and

$$\frac{\partial x^*(\alpha,\beta)}{\partial\beta}=\frac{[\pi_{12}[x^*(\alpha,\beta),x^*(\beta,\alpha)]+\alpha\pi_{21}[x^*(\beta,\alpha),x^*(\alpha,\beta)]]\pi_2[x^*(\alpha,\beta),x^*(\beta,\alpha)]}{K},$$

where

$$\begin{aligned}K=&[\pi_{11}[x^*(\alpha,\beta),x^*(\beta,\alpha)]+\alpha\cdot\pi_{22}[x^*(\beta,\alpha),x^*(\alpha,\beta)]]\\&\cdot[\pi_{11}[x^*(\beta,\alpha),x^*(\alpha,\beta)]+\beta\cdot\pi_{22}[x^*(\alpha,\beta),x^*(\beta,\alpha)]]\\&-[\pi_{12}[x^*(\alpha,\beta),x^*(\beta,\alpha)]+\alpha\cdot\pi_{21}[x^*(\beta,\alpha),x^*(\alpha,\beta)]]\\&\cdot[\pi_{12}[x^*(\beta,\alpha),x^*(\alpha,\beta)]+\beta\cdot\pi_{21}[x^*(\alpha,\beta),x^*(\beta,\alpha)]]\end{aligned}$$

and $K\neq0$ from the regularity assumption. Evaluating these expressions at $\beta=\alpha$, we obtain (writing x^* for $x^*(\alpha,\alpha)$):

$$\begin{aligned}x_1^*(\alpha,\alpha)&=\frac{\partial x^*(\alpha,\beta)}{\partial\alpha}_{|\beta=\alpha}\\&=\frac{-[\pi_{11}(x^*,x^*)+\alpha\cdot\pi_{22}(x^*,x^*)]\cdot\pi_2(x^*,x^*)}{[\pi_{11}(x^*,x^*)+\alpha\cdot\pi_{22}(x^*,x^*)]^2-(1+\alpha)^2[\pi_{12}(x^*,x^*)]^2}\end{aligned} \tag{49}$$

and

$$x_2^*(\alpha,\alpha) = \frac{\partial x^*(\alpha,\beta)}{\partial \beta}_{|\beta=\alpha} = \frac{(1+\alpha)\cdot\pi_{12}(x^*,x^*)\cdot\pi_2(x^*,x^*)}{[\pi_{11}(x^*,x^*)+\alpha\cdot\pi_{22}(x^*,x^*)]^2-(1+\alpha)^2[\pi_{12}(x^*,x^*)]^2}. \quad (50)$$

(We have used the fact that $\beta=\alpha \Rightarrow \pi_{12}[x^*(\alpha,\beta),x^*(\beta,\alpha)] = \pi_{21}[x^*(\beta,\alpha),x^*(\alpha,\beta)]$.)

A necessary second-order condition for x to be a local maximum of $\pi(x,y)+\alpha\cdot\pi(y,x)$, given y, is $\pi_{11}(x,y)+\alpha\cdot\pi_{22}(y,x)\le 0$. It follows from our regularity assumption that $\pi_{11}(x,y)+\alpha\cdot\pi_{22}(y,x)\neq 0$, so $\pi_{11}(x,y)+\alpha\cdot\pi_{22}(y,x)<0$. Since (by hypothesis) $\pi_2\neq 0$, (49) implies $x_1^*(\alpha,\alpha)\neq 0$. Moreover, by (50), $\pi_{12}=0 \Rightarrow x_2^*(\alpha,\alpha)=0$. Finally, since $\pi_{11}(x,y)+\alpha\cdot\pi_{22}(y,x)<0$, it is immediate from these expressions that if $\pi_{12}<0$, $x_1^*(\alpha,\alpha)$ and $x_2^*(\alpha,\alpha)$ have opposite signs, while if $\pi_{12}>0$, $x_1^*(\alpha,\alpha)$ and $x_2^*(\alpha,\alpha)$ have the same sign.

A.4. Theorem 1

Eq. (28) and the assumption $\pi_2\neq 0$ imply

$$D(\alpha,\alpha)=0 \Leftrightarrow (\alpha-\sigma_0)\cdot x_1^*(\alpha,\alpha) = (1-\sigma_0\alpha)\cdot x_2^*(\alpha,\alpha).$$

Since $1-\sigma_0\alpha>0$, it follows that:

1. when actions are strategic substitutes (so that $x_1^*(\alpha,\alpha)\cdot x_2^*(\alpha,\alpha)<0$): $D(\alpha,\alpha)=0 \Leftrightarrow \alpha-\sigma_0<0$;
2. when actions are strategic complements (so that $x_1^*(\alpha,\alpha)\cdot x_2^*(\alpha,\alpha)>0$): $D(\alpha,\alpha)=0 \Leftrightarrow \alpha-\sigma_0>0$;
3. when actions are strategically independent (so that $x_2^*(\alpha,\alpha)=0$ and $x_1^*(\alpha,\alpha)\neq 0$): $D(\alpha,\alpha)=0 \Leftrightarrow \alpha-\sigma_0=0$.

References

Akçay, E., Van Cleve, J., Feldman, M.W., Roughgarden, J., 2009. A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proceedings of the National Academy of Sciences 106, 19061–19066.

Alger, I., 2010. Public goods games, altruism, and evolution. Journal of Public Economic Theory 12, 789–813.

Alger, I., Weibull, J.W., 2010. Kinship, incentives, and evolution. American Economic Review 100, 1725–1758.

André, J.-B., Day, T., 2007. Perfect reciprocity is the only evolutionarily stable strategy in the continuous iterated prisoner's dilemma. Journal of Theoretical Biology 247, 11–22.

Bergstrom, T.C., 1995. On the evolution of altruistic ethical rules for siblings. American Economic Review 85, 58–81.

Bergstrom, T.C., 2003. The algebra of assortative encounters and the evolution of cooperation. International Game Theory Review 5 (3), 211–228.

Bester, H., Güth, W., 1998. Is altruism evolutionarily stable? Journal of Economic Behavior and Organization 34, 193–209.

Christiansen, F.B., 1991. On conditions for evolutionary stability for a continuously varying character. The American Naturalist 138, 37–50.

Day, T., Taylor, P.D., 1998. Unifying genetic and game theoretic models of kin selection for continuous traits. Journal of Theoretical Biology 194, 391–407.

Dekel, E., Ely, J.C., Yilankaya, O., 2007. Evolution of preferences. Review of Economic Studies 74, 685–704.

Durrett, R., Levin, S.A., 2005. Can stable social groups be maintained by homophilous imitation alone? Journal of Economic Behavior & Organization 57, 267–286.

Eshel, I., 1983. Evolutionary and continuous stability. Journal of Theoretical Biology 103, 99–111.

Eshel, I., Cavalli-Sforza, L.L., 1982. Assortment of encounters and evolution of cooperativeness. Proceedings of the National Academy of Sciences 79, 1331–1335.

Eshel, I., Motro, U., 1981. Kin selection and strong evolutionary stability of mutual help. Theoretical Population Biology 19, 420–433.

Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817–868.

Fletcher, J.A., Doebeli, M., 2009. A simple and general explanation for the evolution of altruism. Proceedings of the Royal Society Biology 276, 13–19.

Frank, S.A., 1998. Foundations of Social Evolution. Princeton University Press, Princeton.

Gärdenfors, P., 2008. The role of intersubjectivity in animal and human cooperation. Biological Theory 3, 51–62.

Gardner, A., West, S.A., 2004. Spite and the scale of competition. Journal of Evolutionary Biology 17, 1195–1203.

Gardner, A., West, S.A., Barton, N.H., 2007. The relation between multilocus population genetics and social evolution theory. American Naturalist 169, 207–226.

Geritz, S.A.H., Kisdi, É., Meszéna, G., Metz, J.A.J., 1998. Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree. Evolutionary Ecology 12, 35–57.

Grafen, A., 1979. The hawk–dove game played between relatives. Animal Behavior 27, 905–907.

Grafen, A., 1985. A geometric view of relatedness. Oxford Surveys in Evolutionary Biology 2, 28–89.

Grafen, A., 2006. Optimization of inclusive fitness. Journal of Theoretical Biology 238, 541–563.

Güth, W., 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. International Journal of Game Theory 24, 323–344.

Güth, W., Yaari, M., 1992. An evolutionary approach to explain reciprocal behavior in a simple strategic game. In: Witt, U. (Ed.), Explaining Process and Change—Approaches to Evolutionary Economics. University of Michigan Press, Ann Arbor.

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.

Hamilton, W.D., 1971. Selection of selfish and altruistic behavior in some extreme models. In: Eisenberg, J.F., Dillon, W.S. (Eds.), Man and Beast: Comparative Social Behavior. Smithsonian Press, Washington, DC.

Heifetz, A., Shannon, C., Spiegel, Y., 2006. The dynamic evolution of preferences. Economic Theory 32, 251–286.

Heifetz, A., Shannon, C., Spiegel, Y., 2007. What to maximize if you must. Journal of Economic Theory 133, 31–57.

Hines, W.G.S., Maynard Smith, J., 1979. Games between relatives. Journal of Theoretical Biology 79, 19–30.

Lehmann, L., Rousset, F., 2010. How life history and demography promote or inhibit the evolution of helping behaviours. Philosophical Transactions of the Royal Society B 365, 2599–2617.

Levine, D., 1998. Modeling altruism and spitefulness in experiments. Review of Economic Dynamics 1, 593–622.

Luenberger, D.G., 1969. Optimization by Vector Space Methods. John Wiley & Sons, New York.

McNamara, J.M., Gasson, C.E., Houston, A.I., 1999. Incorporating rules for responding in evolutionary games. Nature 401, 368–371.

Maynard Smith, J., 1974. The theory of games and the evolution of animal conflicts. Journal of Theoretical Biology 47, 209–221.

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

Mohlin, E., 2010. Internalized social norms in conflicts: an evolutionary approach. Economics of Governance 11, 169–181.

Nakamaru, M., Levin, S.A., 2004. Spread of two linked social norms on complex interaction networks. Journal of Theoretical Biology 230, 57–64.

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Nowak, M.A., May, R.M., 1992. Evolutionary games and spatial chaos. Nature 359, 826–829.

Nowak, M.A., Sigmund, K., 2000. Games on grids. In: Dieckmann, U, Law, R., Metz, J.A.J. (Eds.), The Geometry of Ecological Interactions. Cambridge University Press, Cambridge.

Nowak, M.A., Tarnita, C.E., Wilson, E.O., 2010. The evolution of eusociality. Nature 466, 1057–1062.

Ok, E.A., Vega-Redondo, F., 2001. On the evolution of individualistic preferences: an incomplete information scenario. Journal of Economic Theory 97, 231–254.

Queller, D.C., 1984. Kin selection and frequency dependence: a game theoretic approach. Biological Journal of the Linnean Society 23, 133–143.

Queller, D.C., 1985. Kinship, reciprocity and synergism in the evolution of social behaviour. Nature 318, 366–367.

Roberts, G., Sherratt, T.N., 1998. Development of cooperative relationships through increasing investment. Nature 394, 175–179.

Robson, A.J., 1990. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. Journal of Theoretical Biology 144, 379–396.

Rousset, F., Billiard, S., 2000. A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. Journal of Evolutionary Biology 13, 814–825.

Rousset, F., 2004. Genetic Structure and Selection in Subdivided Populations. Princeton University Press, Princeton.

Rowthorn, R., 2006. The evolution of altruism between siblings: Hamilton's rule revisited. Journal of Theoretical Biology 241, 774–790.

Schaffer, M.E., 1988. Evolutionarily stable strategies for finite populations and variable contest size. Journal of Theoretical Biology 132, 467–478.

Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. Journal of Economic Theory 97, 273–297.

Tarnita, C.E., Antal, T., Ohtsuki, H., Nowak, M.A., 2009a. Evolutionary dynamics in set structured populations. Proceedings of the National Academy of Sciences 106, 8601–8604.

Tarnita, C.E., Ohtsuki, H., Antal, T., Fu, F., Nowak, M.A., 2009b. Strategy selection in structured populations. Journal of Theoretical Biology 259, 570–581.

Taylor, P.D., 1989. Evolutionary stability in one-parameter models under weak selection. Theoretical Population Biology 36, 125–143.
Taylor, P.D., Day, T., 2004. Stability in negotiation games and the emergence of cooperation. Proceedings of the Royal Society Biology 271, 669–674.
Taylor, P.D., Jonker, L.B., 1978. Evolutionarily stable strategies and game dynamics. Mathematical Biosciences 40, 145–156.
Taylor, P.D., Frank, S.A., 1996. How to make a kin selection model. Journal of Theoretical Biology 180, 27–37.
Taylor, P.D., Wild, G., Gardner, A., 2006. Direct fitness or inclusive fitness: how shall we model kin selection? Journal of Evolutionary Biology 20, 296–304.
Van Veelen, M., 2006. Why kin and group selection models may not be enough to explain human other-regarding behaviour. Journal of Theoretical Biology 242, 790–797.
Van Veelen, M., 2007. Hamilton's missing link. Journal of Theoretical Biology 246, 551–554.
Van Veelen, M., 2009. Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong. Journal of Theoretical Biology 259, 589–600.
Van Veelen, M., 2011a. A rule is not a rule if it changes from case to case (a reply to Marshall's comment). Journal of Theoretical Biology 270, 189–195.
Van Veelen, M., 2011b. The replicator dynamics with n players and population structure. Journal of Theoretical Biology 276, 78–85.
Wahl, L.M., Nowak, M.A., 1999. The continuous prisoner's dilemma: I. Linear reactive strategies. Journal of Theoretical Biology 200, 307–321.
Weibull, J.W., 1995. Evolutionary Game Theory. MIT Press, Cambridge.
Wenseleers, T., 2006. Modelling social evolution: the relative merits and limitations of a Hamilton's rule-based approach. Journal of Evolutionary Biology 19, 1419–1522.
West, S.A., Mouden, C.E., Gardner, A., 16 common misconceptions about the evolution of cooperation in humans. Evolution and Human Behavior, in press. doi:10.1016/j.evolhumbehav.2010.08.001.
Wild, G., Taylor, P., 2004. Fitness and evolutionary stability in game theoretic models of finite populations. Proceedings of the Royal Society B 271, 2345–2349.
Williams, G.C., Williams, D.C., 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.
Wright, S.G., 1921. Systems of mating. Genetics 6, 111–178.
Wright, S.G., 1922. Coefficients of inbreeding and relationship. American Naturalist 56, 330–338.