Evolution leads to Kantian morality*

Ingela Alger† and Jörgen W. Weibull‡

June 3, 2015

Abstract

What preferences or moral values should one expect evolution to favor? We provide a generalized definition of evolutionary stability of heritable traits in arbitrarily large aggregative interactions under random matching that may be assortative. We establish stability results when these traits are strategies in games, and when they are preferences or moral values in games in which each player's preferences or moral values are the player's private information. We show that certain moral preferences, of a kind that exactly reflects the assortativity in the matching process, are evolutionarily stable. In particular, selfishness is evolutionarily unstable as soon as there is any assortativity. We also establish that evolutionarily stable strategies are the same as those played in equilibrium by rational individuals with evolutionarily stable moral preferences. We provide simple operational criteria for evolutionary stability and apply these to canonical examples.

Keywords: Evolutionary stability, assortativity, morality, *homo moralis*, public goods, contests, helping.

JEL codes: C73, D01, D03.

---

*Support by Knut and Alice Wallenberg Research Foundation and by ANR - Labex IAST is gratefully acknowledged. Ingela Alger also thanks Agence Nationale de la Recherche (ANR) for funding (Chaire d'Excellence). We are grateful for comments from seminar participants at EconomiX, WZB, GREQAM, ETH Zürich, Nottingham, OECD, Séminaire Roy (PSE), IGIER Bocconi, Bern, and École Polytechnique, and from participants at the 2nd Toulouse Economics and Biology Workshop, the 2014 EEA meeting, and the 2014 ASSET meeting.

†Toulouse School of Economics (LERNA, CNRS) and Institute for Advanced Study in Toulouse

‡Stockholm School of Economics, KTH Royal Institute of Technology, and Institute for Advanced Study in Toulouse

1 Introduction

Economics provides a rich set of powerful theoretical models of human societies. Since these models feature individuals whose motivations—preferences and/or moral values—are given, their predictive power depends on the assumptions made regarding these motivations. But if preferences are inherited from past generations, the formation of these preferences may itself be studied theoretically. In particular, one may ask what preferences or moral values have a survival value, and thus what preferences and moral values humans should be expected to have from first principles.

Should we expect pure self-interest, altruism (Becker, 1976), warm glow (Andreoni, 1990), reciprocal altruism (Levine, 1998), inequity aversion (Fehr and Schmidt, 1999), self-image concerns (Bénabou and Tirole, 2006), moral motivation (Brekke, Kverndokk, and Nyborg, 2003), or something else? This question is at the heart of the literature on preference evolution initiated by Güth and Yaari (1992). In a recent contribution to this literature, we found that evolution under certain conditions favors a class of preferences that we called *homo moralis* (Alger and Weibull, 2013).[1] We derived this result in a model where individuals interact in pairs. *Homo moralis* then attaches some weight to his material self-interest but also to what is "the right thing to do if others would do what I do". But in real life many interactions involve more than two persons. Can the methods and results for pairwise interactions be generalized, and if so, how? What preferences and/or moral values does evolution lead to then? These are the questions we address in this paper.

Two major issues drive this quest. First, since (to the best of our knowledge) this exploration has not been made before, we simply did not know what preferences to find. Like in Alger and Weibull (2013), we will let the mathematics show us the way to the preferences that evolution favors, but now for groups of arbitrary size. Our finding, arguably not easy to anticipate and expressing a form of social preference-cum-morality that we have not seen before, will be reported and examined here. The second major reason for pursuing this work is to find out whether moral motivation is evolutionarily viable only in small groups.

More precisely, we propose a general model for the study of the evolutionary foundations of human motivation in strategic interactions in arbitrarily large groups. We define evolutionary stability as a property of abstract "traits" or "types" that can be virtually any characteristic of an individual, such as a behavior pattern or strategy, a goal function,

[1] See also the discussion in Bergstrom (1995).

preference, moral value, belief, or cognitive capacity. Individuals live in a infinite population and are randomly matched in groups of size n to play an n-player game. Each player gets a material payoff that depends on his or her own strategy and on some aggregate of the others' strategies; formally, individuals play an aggregative game in material payoffs.[2] Each player's strategy set may be simple, such as in a simultaneous-move game, or very complex, such as in a extensive-form game with many information sets. Strategies may be pure or mixed. A type is evolutionarily stable if it materially outperforms other types, when the latter appear as rare mutants in the population, and a type is evolutionarily unstable if it is materially outperformed by some rare mutant type.

A key assumption in our model is that the random matching may be assortative in the sense that individuals who are of a vanishingly rare ("mutant") type may face a positive probability of being matched with others of their own rare type, even in the limit as the rare type vanishes. While such matching patterns may at first appear counter-intuitive or even impossible, it is not difficult to think of reasons for why they can arise. First, while distance is not explicitly modeled here, geographic, cultural, linguistic and socioeconomic space imposes (literal or metaphoric) transportation costs, which imply that (1) individuals tend to interact more with individuals in their (geographic, cultural, linguistic or socioeconomic) vicinity,[3] and (2) cultural or genetic transmission of types (say, behavior patterns, preferences or moral values) from one generation to the next also has a natural tendency to take place in the vicinity of where the rare type originally appeared. Taken together, these two tendencies imply the assortativity that we here allow for.[4] In the present model we formalize the

[2]The notion of aggregative games is, to the best of our knowledge, due to Dubey, Mas-Colell and Shubik (1980). See also Corchón (1996). The key feature is that the payoff to a player depends only on the players' own strategy and some (symmetric) aggregation of others' strategies. For a recent paper on aggregative games, see Acemoglu and Jensen (2013). For work on aggregative games more related to ours, see Haigh and Cannings (1989) and Koçkesen, Ok and Sethi (2000a,b).

[3]Homophily has been documented by sociologists (e.g., McPherson, Smith-Lovin, and Cook, 2001, and Ruef, Aldrich, and Carter, 2003) and economists (e.g., Currarini, Jackson, and Pin, 2009, 2010, and Bramoullé and Rogers, 2009). In particular, in a study about race and gender-based choice of friends and meeting chances in U.S. high schools, Currarini, Jackson, and Pin (2009) find that there are strong within-group biases not only in the inferred utility from meetings but also in meeting probabilities. This is particularly relevant for us, since we assume away partner choice.

[4]In biology, the concept of assortativity is known as *relatedness*, and the propensity to interact with individuals locally is nicely captured in the infinite island model, originally due to Wright (1931); see also Rousset (2004).

assortativity of a random matching process in terms of what we call the *assortativity profile*; a probability vector for the events that none, some, or all the individuals in a (vanishingly rare) mutant's group also are mutants, thus generalizing Bergstrom's (2003) definition of assortativity from pairwise encounters to n-person encounters.[5]

Our analysis delivers three main results. First, although we impose minimal restrictions on potential preferences or moral values, our analysis, when applied to preference evolution under incomplete information shows that evolution favors a particular class of preferences. An individual with preferences in this class evaluates what would happen to her own material payoff if with some probability others were to do as she does. Such preferences allow a distinct moral interpretation, and accordingly we use the name *homo moralis* for this class of preferences.[6] In particular they generalize Kantian morality in a probabilistic direction. Indeed, in his *Grundlegung zür Metaphysik der Sitten* (1785), Immanuel Kant wrote "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law." Similarly, *homo moralis* can be interpreted to "act according to that maxim whereby you can, at the same time, will that others should do likewise with some probability."[7]

Importantly, *homo moralis* preferences for groups of size above two are qualitatively different from those for groups of size two. For arbitrary group size n, a *homo moralis* individual maximizes a weighted average of n terms, where the k^{th} term, for $k = 1, 2, ..., n$, is the (hypothetical) material payoff that the individual would obtain if $k-1$ other individuals would use the same strategy as the individual uses. For evolutionary stability the weights must exactly reflect the assortativity profile. Furthermore, and this is the second main result, any preferences that lead to equilibrium behaviors that differ from those of *homo moralis* are evolutionarily unstable. In particular, then, our results imply that material self-interest is evolutionarily inviable as soon as the probability is positive that at least one of the individuals with whom a mutant interacts also is a mutant (when mutants are vanishingly rare).

Our third main result is that the equilibrium behaviors among *homo moralis* whose morality profile exactly reflects the assortativity profile are the same as the behaviors selected

[5] See Bergstrom (2013) and Alger and Weibull (2013) for further discussions of assortativity when $n = 2$.

[6] For a discussion of several ethical principles, see Bergstrom (2009).

[7] *Homo moralis* preferences have got nothing to do with the equilibrium concept "Kantian equilibrium" proposed by Roemer (2010).

for under strategy evolution. This result establishes that evolutionarily stable strategies (under uniform or assortative random matching) need not be interpreted only as resulting when individuals are "programmed" to certain strategies, but can also be interpreted as resulting when individual are rational and free to choose whatever strategy they like, but whose preferences have emerged from natural selection. Together with a first- and second-order characterization for games in Euclidean strategy spaces under conditional independence in the assortativity, we obtain operational methods to find the (symmetric) equilibria of n-player games among *homo moralis*, methods we illustrate in various canonical examples.

The general model is described in the next section. This model is then applied to preference evolution under incomplete information (Section 3) and strategy evolution (Section 4). In Section 5 we present a characterization result, which we then apply to several commonly studied games in Section 6. Prior to concluding (Section 8), we review the literature in Section 7.

2 Model

Consider an infinite (continuum) population of individuals who are randomly matched into groups of $n \geq 1$ individuals to interact according to some game given in normal form $\Gamma = \langle N, X^n, \pi \rangle$, where $N = \{1, 2, ..., n\}$ is the set of players, X is the set of *strategies* available to each player and $\pi : X \times X^{n-1} \to \mathbb{R}$ is the *material payoff function*.[8] The material payoff to any player $i \in N$ from using strategy $x_i \in X$ against the strategies $x_j \in X$ $(j \neq i)$ of the others in the group is denoted $\pi(x_i, \boldsymbol{x}_{-i})$. We assume that $\pi(x_i, \boldsymbol{x}_{-i})$ is invariant under permutations of the components of $\boldsymbol{x}_{-i}$, the strategy profile of all other individuals in the group. These games may thus be called *aggregative*.[9] We will assume throughout that the set X is a non-empty, compact and convex set in some topological vector space, and that the function π is continuous.[10] The generality of the strategy set X allows for simultaneous-move

[8]This game will subsequently serve as a "game protocol" in the sense of Weibull (2004), that is, participants will be allowed to have their own personal preferences over strategy profiles, preferences that are not required to be functions of the material payoff outcomes.

[9]More precisely: for any $x_i \in X$ and $\mathbf{x}_{-i} \in X^{n-1}$, and any bijection $h : \{2, 3, ..., n\} \to \{2, 3, ..., n\}$: $\pi\left(x_i, x_{h(2)}, x_{h(3)}, ..., x_{h(n)}\right) = \pi(x_i, \mathbf{x}_{-i})$.

[10]More precisely, it is sufficient for the subsequant analysis that X is a locally convex Hausdorff space, see Aliprantis and Border (2006).

games, games with sequential moves and asymmetric information etc. Indeed, Γ may be any symmetric and finite n-player extensive-form game with perfect recall and X its the set of mixed or behavior strategies.

Each individual has some *type* (or *trait*) $\theta \in \Theta$, which may influence his/her choice of strategy, or *behavior* in the game Γ, where Θ is the set of potential types. Consider a population in which at most two types from Θ are present. For any types θ and τ, and any $\varepsilon \in (0,1)$, let $s = (\theta, \tau, \varepsilon)$ be the *population state* in which the two types are represented in population shares $1-\varepsilon$ and ε, respectively. Let $S = \Theta^2 \times (0,1)$ denote the set of population states. We are particularly interested in states $s = (\theta, \tau, \varepsilon)$ in which ε is small, then calling θ the *resident* type, being predominant in the population, and τ, being rare, the *mutant* type.

In a given population state $s \in S$, the behavioral outcomes, or, more precisely, strategy profiles used, may, but need not, be uniquely determined. For each population state s, let $V(s) \subset \mathbb{R}^2$ be the set of (average) material-payoff pairs that *can* arise in population state s, where, for any $v = (v_1, v_2) \in V(\theta, \tau, \varepsilon)$, the first component, v_1, is the average material payoff to individuals of type θ, and the second component, v_2, that to individuals of type τ. We assume that $V(s)$ is non-empty and compact for all states $s = (\theta, \tau, \varepsilon)$. Then

$$\varphi(\theta, \tau, \varepsilon) = \min_{v \in V(\theta,\tau,\varepsilon)} (v_1 - v_2) \tag{1}$$

is well-defined. In words, $\varphi(\theta, \tau, \varepsilon)$, is the material payoff difference between residents and mutants, in the residents' worst possible outcome as compared with mutants (in terms of material payoffs), across all behavioral outcomes that are possible in state $s = (\theta, \tau, \varepsilon)$. In particular, $\varphi(s) > 0$ if and only if the residents earn a (strictly) higher (average) material payoff than the mutants in all possible outcomes in that state.[11]

The following definitions of evolutionary stability and instability are generalizations of the definitions in Alger and Weibull (2013), from $n = 2$ to $n \geq 2$, and from preferences to arbitrary types.

Definition 1 *A type θ is **evolutionarily stable against a type** τ if $\varphi(\theta, \tau, \varepsilon) > 0$ for all $\varepsilon > 0$ sufficiently small. A type θ is **evolutionarily stable** if it is evolutionarily stable against all types $\tau \neq \theta$. A type θ is **evolutionarily unstable** if there exists a type τ such that $\varphi(\theta, \tau, \varepsilon) < 0$ for arbitrarily small $\varepsilon > 0$.*

[11] The function φ is a generalization of the so-called score function in evolutionary game theory, see, e.g., Bomze and Pötscher (1989).

Our requirement for stability is demanding; the residents should earn a higher material payoff in all behavioral outcomes for all sufficiently small population shares of the mutant type. By contrast, the requirement for instability is relatively weak; it suffices to find one type that would earn a higher material payoff in some behavioral outcome in some population state with arbitrarily few mutants.[12] Clearly, by these definitions no type is both evolutionarily stable and unstable, and there may, in general, exist types that are neither stable nor unstable.

2.1 Matching

The matching process is exogenous. In any population state $s = (\theta, \tau, \varepsilon) \in S$, the number of mutants—individuals of type τ—in a group that is about to play game $\Gamma = \langle N, X^n, \pi \rangle$, is a random variable that we will denote T. For any *resident* drawn at random from the population let $p_m(\varepsilon)$ be the conditional probability $\Pr[T = m \mid \theta, s]$ that the total number of mutants in the resident's group is m, for $m = 0, 1, .., n-1$.[13] Likewise, for any mutant, also drawn at random from the population, let $q_m(\varepsilon)$ be the conditional probability $\Pr[T = m \mid \tau, s]$ that the total number of mutants in his or her group is m, for $m = 1, .., n$. We assume that each function p_m and each function q_m is continuous and has a limit as $\varepsilon \to 0$, which we denote p_m^0 and q_m^0, respectively.

In order to get a grip on these limiting probabilities, we use the *algebra of assortative encounters* developed by Bergstrom (2003) for pairwise interactions. For a given population state $s = (\theta, \tau, \varepsilon)$, let $\Pr[\theta|\theta, \varepsilon]$ denote the conditional probability for an individual of type θ that another, uniformly randomly drawn member of his or her group also is of type θ. Likewise, let $\Pr[\theta|\tau, \varepsilon]$ denote the conditional probability for an individual of type τ that any other uniformly randomly drawn member of his or her group has type θ. Let $\phi(\varepsilon)$ be the difference between the two probabilities:

$$\phi(\varepsilon) = \Pr[\theta|\theta, \varepsilon] - \Pr[\theta|\tau, \varepsilon]. \tag{2}$$

[12] More precisely, for any given $\varepsilon > 0$ there should exist some $\varepsilon' \in (0, \varepsilon)$ such that $\varphi(\theta, \tau, \varepsilon') < 0$.

[13] The first random draw cannot, technically, be uniform, in an infinite population. The reasoning in this section is concerned with matchings in finite populations in the limit as the total population size goes to infinity. We refer the reader to the appendix for a detailed example.

This defines the *assortment function* $\phi : (0,1) \to [-1,1]$. We assume that

$$\lim_{\varepsilon \to 0} \phi(\varepsilon) = \sigma, \tag{3}$$

for some $\sigma \in \mathbb{R}$, the *index of assortativity* of the matching process (Bergstrom, 2003). Moreover, by setting $\phi(0) = \sigma$ we henceforth extend the domain of ϕ from $(0,1)$ to $[0,1)$.

The following equation is a necessary balancing condition:

$$(1-\varepsilon) \cdot [1 - \Pr[\theta|\theta,\varepsilon]] = \varepsilon \cdot \Pr[\theta|\tau,\varepsilon]. \tag{4}$$

Each side of the equation equals the probability for the following event: draw at random an individual from the population at large and then draw at random another individual from the first individual's group, and observe that these two individuals are of different types. Equations (2) and (4) together give

$$\begin{cases} \Pr[\theta|\theta,\varepsilon] = \phi(\varepsilon) + (1-\varepsilon)[1-\phi(\varepsilon)] \\ \Pr[\theta|\tau,\varepsilon] = (1-\varepsilon)[1-\phi(\varepsilon)]. \end{cases} \tag{5}$$

Now let $\varepsilon \to 0$. Then, from (4), $\Pr[\theta|\theta,\varepsilon] \to 1$, and hence $p_0(\varepsilon) \to 1$. In other words, residents virtually never meet mutants when the latter are vanishingly rare. Without loss of generality we may thus uniquely extend the domain of p_m from $(0,1)$ to $[0,1)$, while preserving its continuity, by setting $p_0(0) = 1$. We also note that together with (5), this property implies that $\sigma \in [0,1]$.[14]

Turning now to the limit of $q_m(\varepsilon)$ as ε tends to zero (for $m = 1, ..., n$), we first note that in the special case $n = 2$,

$$q_2^0 = \lim_{\varepsilon \to 0} \Pr[\tau|\tau,\varepsilon] = 1 - \lim_{\varepsilon \to 0} \Pr[\theta|\tau,\varepsilon] = 1 - \left[1 - \lim_{\varepsilon \to 0} \phi(\varepsilon)\right] = \sigma.$$

However, for $n > 2$ there remains a statistical issue, namely whether or not, for a given mutant, the types of any two *other* members in her group are statistically dependent or not (in the given population state). We will not make any specific assumption about this in the general analysis, and we will refer to the vector $\boldsymbol{q}^0 = (q_1^0, ..., q_n^0)$ as the *assortativity profile* of the matching process.

[14] This contrasts with the case of a finite population, where negative assortativity can arise for population states with few mutants (see Schaffer, 1988).

2.2 Homo moralis

Prior to turning to the analysis, we define *homo moralis* preferences. We write utility functions in the same form as the material payoff function, that is, with the player's own strategy as the first argument and the profile of others' strategies as the second (vector) argument; thus, for any player $i \in N$ and any strategy profile $\boldsymbol{x} \in X^n$, the utility of individual i is written as a function of $(x_i, \boldsymbol{x}_{-i})$. An individual with *homo moralis* preferences evaluates $(x_i, \boldsymbol{x}_{-i})$ by maximizing a weighted sum of the material payoffs that she would obtain if all, some, or none of the others would choose the same strategy as herself. To formally describe all the possible hypothetical strategy profiles that she thus ponders, we define a vector-valued random variable.

Let Δ^n be the unit simplex of probability vectors in $\mathbb{R}^n$; $\Delta^n = \left\{\boldsymbol{\mu} \in \mathbb{R}^n_+ : \Sigma^n_{m=1}\mu_m = 1\right\}$. For any $\boldsymbol{\mu} \in \Delta^n$, any player $i \in N$, and any strategy profile $\boldsymbol{x} \in X^n$, let $\tilde{\boldsymbol{x}}_{-i} : \Omega \to X^{n-1}$ be a vector-valued random variable such that with probability μ_m (for $m = 1, .., n$) exactly $m-1$ of the $n-1$ components in $\boldsymbol{x}_{-i}$ are replaced by x_i, with equal probability for each subset of $m-1$ replaced components, while the remaining components keep their original value.[15] For each $m = 1, ..., n$, there are $\binom{n-1}{m-1}$ elements in the associated subset of components of $\boldsymbol{x}_{-i}$, so the probability for a particular such subset of components is $\mu_m / \binom{n-1}{m-1}$.

Definition 2 *Player i is a* homo moralis *if his or her utility function $u_{\boldsymbol{\mu}} : X^n \to \mathbb{R}$ satisfies*

$$u_{\boldsymbol{\mu}}(x_i, \boldsymbol{x}_{-i}) = \mathbb{E}_{\boldsymbol{\mu}}\left[\pi(x_i, \tilde{\boldsymbol{x}}_{-i}) \mid \boldsymbol{x}\right] \quad \forall \boldsymbol{x} \in X^n \tag{6}$$

for some $\boldsymbol{\mu} \in \Delta^n$. The vector $\boldsymbol{\mu}$ is the player's morality profile.

Three extreme cases are noteworthy. First, the utility function $u_{\boldsymbol{\mu}}(x_i, \boldsymbol{x}_{-i})$ would take the value $\pi(x_i, \boldsymbol{x}_{-i})$ if $\mu_1 = 1$. In this case, the individual's goal is to choose a strategy x_i that maximizes her own material payoff, given the strategy profile $\boldsymbol{x}_{-i}$ for all other participants. Second, at the opposite extreme the utility function $u_{\boldsymbol{\mu}}(x_i, \boldsymbol{x}_{-i})$ would take the value $\pi(x_i, x_i, .., x_i)$ if $\mu_n = 1$; in this case, her goal is "to do the right thing" according to Kant's categorical imperative applied to material payoffs. In other words, she would then choose a strategy x_i that maximizes her material payoff if all others were to choose that same strategy. We refer to the first case as *homo oeconomicus* and the second as *homo kantientis*. Third, if $\mu_1 + \mu_n = 1$, the individual maximizes a convex combination of own material

[15] If it happens that $x_j = x_i$ for some $j \neq i$, then the replacement has no effect on that component j.

payoff and the material payoff that would arise should all players use the same strategy x_i: $u_{\boldsymbol{\mu}}(x_i, \boldsymbol{x}_{-i}) = \mu_1 \cdot \pi(x_i, \boldsymbol{x}_{-i}) + \mu_n \cdot \pi(x_i, x_i, .., x_i)$. In particular, if $n = 2$ the equality $\mu_1 + \mu_n = 1$ always holds and one then obtains $u_{\boldsymbol{\mu}}(x, y) = (1 - \kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x)$ for $\kappa = \mu_2$, the same expression as in Alger and Weibull (2013). However, for $n > 2$ a *homo moralis* may also attach a positive weight to the material payoff that she would obtain if some but not all others were to use the same strategy as herself. Morality still has a distinct Kantian flavor, since the individual evaluates what would happen to her material payoff if others were to behave as she does.

For any *homo moralis* and $\boldsymbol{\mu} \in \Delta^n$, let $\beta_{\boldsymbol{\mu}} : X \rightrightarrows X$ be defined by

$$\beta_{\boldsymbol{\mu}}(x) = \arg\max_{y \in X} u_{\boldsymbol{\mu}}\left(y, \mathbf{x}^{(n-1)}\right), \tag{7}$$

where $\mathbf{x}^{(n-1)}$ is the $(n-1)$-dimensional vector whose all components equal $x \in X$. The set of symmetric Nash-equilibrium strategies in a game played by n *homo moralis* with the same morality profile $\boldsymbol{\mu}$ is

$$X_{\boldsymbol{\mu}} = \left\{x \in X : x \in \beta_{\boldsymbol{\mu}}(x)\right\}. \tag{8}$$

Thanks to permutation invariance, a strategy x belongs to this fixed-point set $X_{\boldsymbol{\mu}}$ if and only if

$$x \in \arg\max_{y \in X} \sum_{m=1}^{n} \mu_m \cdot \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right), \tag{9}$$

where $\mathbf{y}^{(m-1)}$ is the $(m-1)$-dimensional vector whose components equal y, and $\mathbf{x}^{(n-m)}$ is the $(n-m)$-dimensional vector whose components equal x.

3 Preference evolution under incomplete information

From now on, let Θ be the set of all continuous aggregative utility functions, i.e., each type $\theta \in \Theta$ uniquely determines a continuous function $u_\theta : X \times X^{n-1} \to \mathbb{R}$ that its "host" strives to maximize, and for every continuous aggregative utility function u there exists a type $\theta \in \Theta$ that has u as its goal function. In line with the notation introduced above, we will write $\theta = \boldsymbol{\mu}$ to denote *homo moralis* with morality profile $\boldsymbol{\mu} \in \Delta^n$.

We focus on the case when each individual's utility function is his or her private information. Then an individual's behavior cannot be conditioned on the types of the others with whom (s)he has been matched. However, individual behavior may be adapted to the population state at hand (that is, the types present in the population, and their population

shares). Arguably, Bayesian Nash equilibrium is a natural criterion to delineate the set $V(s)$ of (average) material-payoff pairs that can arise in a population state s.[16]

More precisely, in any given state $s = (\theta, \tau, \varepsilon) \in \Theta^2 \times (0,1)$, a (type-homogenous Bayesian) Nash equilibrium is a pair of strategies, one for each type, such that each strategy is a best reply for any player of that type in the given population state. In other words, all players of the same type use the same strategy, and each individual player finds his or her strategy optimal, given his or her utility function.

Definition 3 *In any state $s = (\theta, \tau, \varepsilon) \in \Theta^2 \times (0,1)$, a strategy pair $(\hat{x}, \hat{y}) \in X^2$ is a (type-homogenous Bayesian)* ***Nash Equilibrium*** *if*

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot u_\theta\left(x, \hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}}^{(n-m-1)}\right) \\ \hat{y} \in \arg\max_{y \in X} \ \sum_{m=1}^{n} q_m(\varepsilon) \cdot u_\tau\left(y, \hat{\mathbf{y}}^{(m-1)}, \hat{\mathbf{x}}^{(n-m)}\right). \end{cases} \tag{10}$$

Let $B^{NE}(s) \subseteq X^2$ denote the set of (type-homogenous Bayesian) Nash equilibria in state $s = (\theta, \tau, \varepsilon)$, that is, all solutions $(\hat{x}, \hat{y})$ of (10). For given types θ and τ, this defines an equilibrium correspondence $B^{NE}(\theta, \tau, \cdot) : (0,1) \rightrightarrows X^2$ that maps mutant population shares ε to the associated set of equilibria. As discussed above, under the assumption that all probabilities in (10) are continuous in ε and converge as $\varepsilon \to 0$, the domain of these probabilities was continuously extended to $[0,1)$. This allows us to likewise extend the domain of $B^{NE}(\theta, \tau, \cdot)$ to include $\varepsilon = 0$, where $(\hat{x}, \hat{y}) \in B^{NE}(\theta, \tau, 0)$ if and only if

$$\begin{cases} \hat{x} \in \arg\max_{x \in X} \ \sum_{m=0}^{n-1} \left[\lim_{\varepsilon \to 0} p_m(\varepsilon)\right] \cdot u_\theta\left(x, \hat{\mathbf{y}}^{(m)}, \hat{\mathbf{x}}^{(n-m-1)}\right) \\ \hat{y} \in \arg\max_{y \in X} \ \sum_{m=1}^{n} \left[\lim_{\varepsilon \to 0} q_m(\varepsilon)\right] \cdot u_\tau\left(y, \hat{\mathbf{y}}^{(m-1)}, \hat{\mathbf{x}}^{(n-m)}\right). \end{cases} \tag{11}$$

We note, in particular, that the first equation in (11) is equivalent with (symmetric) Nash equilibrium play among the residents themselves, and is hence independent of the mutant type τ.

By a slight generalization of the arguments in the proof of Lemma 1 in Alger and Weibull (2013) one obtains that the set $B^{NE}(\theta, \tau, \varepsilon)$ is compact for each $(\theta, \tau, \varepsilon) \in \Theta^2 \times [0,1)$, and the correspondence $B^{NE}(\theta, \tau, \cdot) : [0,1) \rightrightarrows X^2$ is upper hemi-continuous. Moreover, $B^{NE}(\theta, \tau, \varepsilon) \neq \varnothing$ if u_θ and u_τ are concave in their first arguments. We will henceforth focus on types θ and τ such that $B^{NE}(\theta, \tau, \varepsilon)$ is non-empty for all $\varepsilon \in [0,1)$. This holds, for example, if all functions u_θ are concave in their first argument, the player's own strategy.

[16] This can be interpreted as an adiabatic process in which preferences change on a slower time scale than actions, see Sandholm (2001).

Given a population state $s = (\theta, \tau, \varepsilon)$ and some Nash equilibrium $(\hat{x}, \hat{y}) \in B^{NE}(s)$, the average equilibrium material payoffs to residents and mutants, respectively, equal $F(\hat{x}, \hat{y}, \varepsilon)$ and $G(\hat{x}, \hat{y}, \varepsilon)$, where $F, G : X^2 \times [0, 1) \to \mathbb{R}$ are defined by

$$F(x, y, \varepsilon) = \sum_{m=0}^{n-1} p_m(\varepsilon) \cdot \pi\left(x, \mathbf{x}^{(n-m-1)}, \mathbf{y}^{(m)}\right), \tag{12}$$

and

$$G(x, y, \varepsilon) = \sum_{m=1}^{n} q_m(\varepsilon) \cdot \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right), \tag{13}$$

where $\mathbf{x}^{(n-m-1)}$ is the $(n-m-1)$-dimensional vector whose components all equal x, $\mathbf{y}^{(m)}$ the m-dimensional vector whose components all equal y, and likewise for $\mathbf{y}^{(m-1)}$ and $\mathbf{x}^{(n-m)}$. Both F and G are continuous by virtue of the assumed continuity of the material payoff function and the matching probabilities.

For each type $\theta \in \Theta$ let $\beta_\theta : X \rightrightarrows X$ denote the best-reply correspondence,

$$\beta_\theta(y) = \arg\max_{x \in X} u_\theta\left(x, \mathbf{y}^{(n-1)}\right) \quad \forall y \in X,$$

and $X_\theta \subseteq X$ its set of fixed points,

$$X_\theta = \{x \in X : x \in \beta_\theta(x)\}.$$

Given the unrestricted nature of the set of potential types, for any resident type there may be other types such that, if appearing in rare mutants, would give rise to the same behavior as that of the residents. We define the *behavioral alikes* to a type θ as those types that, as vanishingly rare mutants among residents of type θ, behave just as a resident could rationally do, in some equilibrium. Formally, for any given type $\theta \in \Theta$, this is the subset[17]

$$\tilde{\Theta}(\theta) = \left\{\tau \in \Theta : (x^*, y^*) \in B^{NE}(\theta, \tau, 0) \text{ for some } x^* \in X_\theta \text{ and } y^* \in \beta_\theta(x^*)\right\}. \tag{14}$$

Examples of such behavioral alikes are individuals with utility functions that are positive affine transformations of the utility function of the residents, and also individuals for whom some strategy in X_θ is dominant.[18]

[17] This definition labels a slightly wider range of types as behavioral alikes than according to our definition in Alger and Weibull (2013); $\Theta_\theta \subseteq \tilde{\Theta}(\theta)$.

[18] For example, if $x^* \in X_\theta$, let $u(x_i, \mathbf{x}_{-i}) \equiv -(x_i - x^*)^2$.

We are now in a position to state and prove our main result, namely, that *homo moralis* with morality profile that reflects the assortativity profile of the matching process is evolutionarily stable against all types that are not its behavioral alikes, and any type that does not behave like this particular variety of *homo moralis* when resident is unstable:

Theorem 1 Homo moralis *with morality profile* $\boldsymbol{\mu} = \boldsymbol{q}^o$ *is evolutionarily stable against all types* $\tau \notin \tilde{\Theta}(\boldsymbol{q}^o)$*. A type* $\theta \in \Theta$ *is evolutionarily unstable if* $X_\theta \cap X_{\boldsymbol{q}^o} = \varnothing$**.**

Proof: Since π is continuous, and all functions p_m and q_m (given $\theta, \tau \in \Theta$) are continuous in ε by hypothesis, also the two functions $F, G : X^2 \times [0,1) \to \mathbb{R}$ (given $\theta, \tau \in \Theta$) are continuous.

For the first claim, let $\theta = \boldsymbol{\mu}$ for $\boldsymbol{\mu} = \boldsymbol{q}^o$ and $\tau \notin \tilde{\Theta}(\boldsymbol{q}^o)$, and suppose that $(x,y) \in B^{NE}(\boldsymbol{q}^o, \tau, 0)$. Then $x \in X_{\boldsymbol{q}^o}$ so $u_{\boldsymbol{q}^o}\left(x, \mathbf{x}^{(n-1)}\right) \geq u_{\boldsymbol{q}^o}\left(y, \mathbf{x}^{(n-1)}\right)$. Since $\tau \notin \tilde{\Theta}(\boldsymbol{q}^o)$: $y \notin \beta_{\boldsymbol{q}^o}(x)$. Hence, $u_{\boldsymbol{q}^o}\left(x, \mathbf{x}^{(n-1)}\right) > u_{\boldsymbol{q}^o}\left(y, \mathbf{x}^{(n-1)}\right)$, or, equivalently, $F(x,y,0) > G(x,y,0)$. Let $D : X^2 \to \mathbb{R}$ be defined by $D(x,y) = F(x,y,0) - G(x,y,0)$. By continuity of F and G, also D is continuous. Since $B^{NE}(\boldsymbol{q}^o, \tau, 0)$ is compact and $D(x,y) > 0$ on $B^{NE}(\boldsymbol{q}^o, \tau, 0)$, we have $\min_{(x,y) \in B^{NE}(\boldsymbol{q}^o, \tau, 0)} D(x,y) = \delta$ for some $\delta > 0$. Again by continuity of F and G, there exists a neighborhood $U \subseteq X^2 \times [0,1)$ of the compact set $B^{NE}(\boldsymbol{q}^o, \tau, 0) \times \{0\}$ such that $F(x,y,\varepsilon) - G(x,y,\varepsilon) > \delta/2$ for all $(x,y,\varepsilon) \in U$. Since $B^{NE}(\boldsymbol{q}^o, \tau, \cdot) : [0,1) \rightrightarrows X^2$ is compact-valued and upper hemi-continuous, there exists an $\bar{\varepsilon} > 0$ such that $B^{NE}(\boldsymbol{q}^o, \tau, \varepsilon) \times [0, \varepsilon] \subset U$ for all $\varepsilon \in [0, \bar{\varepsilon})$. It follows that $F(x,y,\varepsilon) - G(x,y,\varepsilon) > \delta/2$ for all $\varepsilon \in [0, \bar{\varepsilon})$ and all $(x,y) \in B^{NE}(\boldsymbol{q}^o, \tau, \varepsilon)$. For $V(\boldsymbol{q}^o, \tau, \varepsilon)$ defined as the set of vectors $v = (v_1, v_2) \in \mathbb{R}^2$ such that $v_1 = F(\hat{x}, \hat{y}, \varepsilon)$ and $v_2 = G(\hat{x}, \hat{y}, \varepsilon)$ for some $(\hat{x}, \hat{y}) \in B^{NE}(\boldsymbol{q}^o, \tau, \varepsilon)$, we thus have $\varphi(\boldsymbol{q}^o, \tau, \varepsilon) > \delta/2$ for all $\varepsilon \in [0, \bar{\varepsilon})$. This establishes the first claim.

For the second claim, let $\theta \in \Theta$ be such that $X_\theta \cap X_{\boldsymbol{q}^o} = \varnothing$ and let $x_\theta \in X_\theta$. Then $u_{\boldsymbol{q}^o}(\hat{x}, \mathbf{x}_\theta^{(n-1)}) > u_{\boldsymbol{q}^o}(x_\theta, \mathbf{x}_\theta^{(n-1)})$ for some $\hat{x} \in X$. Since Θ is the set of all continuous (aggregative) functions, there exists a type $\tau \in \Theta$ for which $\hat{x}$ is a strictly dominant strategy (for example $u_\tau\left(x, \mathbf{x}^{(n-1)}\right) \equiv -(x - \hat{x})^2$), so individuals of that type will always play $\hat{x}$. By definition of $u_{\boldsymbol{q}^o}$,

$$G(x_\theta, \hat{x}, 0) = u_{\boldsymbol{q}^o}(\hat{x}, \mathbf{x}_\theta^{(n-1)}) > u_{\boldsymbol{q}^o}(x_\theta, \mathbf{x}_\theta^{(n-1)}) = F(x_\theta, \hat{x}, 0).$$

Let $\langle \varepsilon_t \rangle_{t \in \mathbb{N}}$ be any sequence from $(0,1)$ such that $\varepsilon_t \to 0$. By upper hemi-continuity of $B^{NE}(\theta, \tau, \cdot)$ there exists a sequence $\langle x_t, y_t \rangle_{t \in \mathbb{N}}$ from X^2 such that $x_t \to x_\theta \in X_\theta$ and $(x_t, y_t) \in B^{NE}(\theta, \tau, \varepsilon_t)$ for all $t \in \mathbb{N}$. By definition of type τ, $y_t = \hat{x}$ for all $t \in \mathbb{N}$. Since F and G are

continuous, there exists a $T > 0$ such that $G(x_t, \hat{x}, \varepsilon_t) > F(x_t, \hat{x}, \varepsilon_t)$ for all $t > T$, and thus $\varphi(\theta, \tau, \varepsilon_t) < 0$ for all such t. Hence, for any given $\bar{\varepsilon} > 0$ there exist infinitely many $\varepsilon \in (0, \bar{\varepsilon})$ such that $\varphi(\theta, \tau, \varepsilon) < 0$. **Q.E.D.**

The theorem establishes that as long as there is some assortativity, in the sense that $q_1^0 \neq 1$, evolutionary stability requires *homo moralis* preferences of a morality profile that precisely reflects this assortativity (or any preferences that would give rise to precisely the same behavior). In particular, then, this result provides a novel insight about a question of particular interest for economists, namely, whether the common assumption of selfishness has an evolutionary justification. In a nutshell, the theorem says that if preferences are unobservable and individuals play some Bayesian Nash equilibrium, selfishness (individuals with $\pi(x_i, \boldsymbol{x}_{-i})$ as their utility function) is evolutionarily stable (modulo behavioral alikes) if and only if there is no assortativity at all in the matching process, i.e., $q_1^0 = 1$**.**

The intuition for this result is that in a population that consists almost solely of *homo moralis* with the "right" morality profile, individuals play a strategy that would maximize the average material payoff to a vanishingly rare mutant in this population. In a sense, thus, a population consisting of such *homo moralis preempts* entry by rare mutants, rather than doing what would be best (in terms of material payoff) for the residents if there were no mutants around.[19]

4 Strategy evolution

Here we adopt the assumption that was used for the original formulation of evolutionary stability (Maynard Smith and Price, 1973), namely, that an individual's type is a strategy that she always uses. A question of particular interest is whether strategy evolution gives guidance to the behaviors that result under preference evolution.

Formally, let the set of potential types be $\Theta = X$, the strategy set for the game $\Gamma = \langle N, X^n, \pi \rangle$. Thus, in a population where some types $\theta = x$ and $\tau = y$ are present, x is always played the residents and y is always played by the mutants. The material payoff to a resident who belongs to a group with m mutants can be written $\pi\left(x, \mathbf{x}^{(n-m-1)}, \mathbf{y}^{(m)}\right)$, where $\mathbf{x}^{(n-m-1)}$ is the $(n-m-1)$-dimensional vector whose components all equal x, and $\mathbf{y}^{(m)}$ is

[19] See also Alger and Weibull (2013) and Robson and Szentes (2014) for a similar observation. Importantly, this logic is very different from that of group selection.

the m-dimensional vector whose components all equal y. Likewise, the material payoff to a mutant who belongs to such a group is $\pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right)$. Hence, given any pair of strategies (x, y), for each ε the average material payoff to a resident is $F(x, y, \varepsilon)$ and the average material payoff to a mutant is $G(x, y, \varepsilon)$, see (12) and (13).

Under strategy evolution, then, the set of (average) material-payoff pairs that can arise in population state s, $V(s) \subset \mathbb{R}^2$, is a singleton for all population states $s \in S = X^2 \times (0,1)$, and

$$\varphi(x, y, \varepsilon) = F(x, y, \varepsilon) - G(x, y, \varepsilon).$$

Furthermore, for any $x, y \in X$, $\varphi(x, y, \varepsilon)$ converges (to some real number) as ε tends to zero.

A *necessary* condition for x to be an evolutionarily stable strategy is

$$\lim_{\varepsilon \to 0} \varphi(x, y, \varepsilon) \geq 0 \qquad \forall y \in X. \tag{15}$$

In other words, it is necessary that the residents on average do not earn a lower material payoff than the mutants when the latter are virtually absent from the population. Likewise, a *sufficient* condition for evolutionary stability is that this inequality holds strictly for all strategies $y \neq x$.

Let $H : X^2 \to \mathbb{R}$ be the function defined by

$$H(y, x) = \lim_{\varepsilon \to 0} G(x, y, \varepsilon). \tag{16}$$

The function value $H(y, x)$ is the average material payoff to a mutant with strategy y in a population where the resident strategy is x and where the population share of mutants is vanishingly small. Since $H(x, x) = \lim_{\varepsilon \to 0} G(x, x, \varepsilon) = \lim_{\varepsilon \to 0} F(x, x, \varepsilon) = \lim_{\varepsilon \to 0} F(x, y, \varepsilon)$, the necessary condition (15) for a strategy x to be evolutionarily stable may be written

$$H(x, x) \geq H(y, x) \qquad \forall y \in X, \tag{17}$$

or, equivalently,

$$x \in \arg\max_{y \in X} H(y, x). \tag{18}$$

This condition says that for a strategy x to be evolutionarily stable, its users have to earn the same average material payoff as the "the most threatening mutants", those with the highest average material payoff that any vanishingly rare mutant can obtain against the resident. As under preference evolution, then, an evolutionarily stable type *preempts* entry by rare mutants.

A sufficient condition for a strategy x to be evolutionarily stable is that

$$H(x,x) > H(y,x) \tag{19}$$

for all $y \neq x$. Interestingly, then, irrespective of n, evolutionarily stable types may be interpreted as Nash equilibrium strategies in a derived two-player game, where "nature" plays strategies against each other:

Proposition 1 *Let $\Theta = X$. If x is an evolutionarily stable strategy in $\Gamma = \langle N, X^n, \pi \rangle$, then (x,x) is a Nash equilibrium of the symmetric two-player game in which the strategy set is X and the payoff function is H. If (x,x) is a strict Nash equilibrium of the latter game, then x is an evolutionarily stable strategy in $\Gamma = \langle N, X^n, \pi \rangle$, while if (x,x) is a not a Nash equilibrium, then x is evolutionarily unstable.*

This proposition allows us to make a first connection between strategy evolution and *homo moralis* preferences. Indeed, while under strategy evolution each individual mechanistically plays a certain strategy—is "programmed" to execute a certain strategy—we will now see that any evolutionarily stable strategy may be viewed as if emerging from individuals' free choice, as if they were striving to maximize a specific utility function. To see this, note that thanks to permutation invariance, $H(y,x)$ writes

$$H(y,x) = \sum_{m=1}^{n} q_m^0 \cdot \pi\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right). \tag{20}$$

Combining this observation with Proposition 1 and the fixed-point equation (18), we obtain the following proposition:

Corollary 1 *Let $\Theta = X$ (strategy evolution). If x is an evolutionarily stable strategy, then it belongs to $X_{\boldsymbol{q}^o}$. Every strategy $x \in X_{\boldsymbol{q}^o}$ for which $\beta_{\boldsymbol{q}^o}(x)$ is a singleton is evolutionarily stable. Every strategy $x \notin X_{\boldsymbol{q}^o}$ is evolutionarily unstable.*

This corollary establishes that the behavior induced under strategy evolution is *as if* individuals were equipped with *homo moralis* preferences with a morality profile that exactly reflects the assortativity profile. More formally, in games $\Gamma = \langle N, X^n, \pi \rangle$ where *homo moralis* of morality profile $\boldsymbol{q}^o$ has a unique best reply to each strategy in $X_{\boldsymbol{q}^o}$, preference evolution under incomplete information induces the same behaviors as strategy evolution. This establishes a second connection between strategy evolution and *homo moralis* preferences;

evolutionarily stable strategies may be viewed as emerging from preference evolution when individuals are not programmed to strategies but instead are (game-theoretically) rational and play equilibria under incomplete information.

5 Conditional independence and differentiability

How does *homo moralis* behave in comparison with *homo oeconomicus*? In this section we focus on conditionally independent random matching. For this class of matching processes we determine the set of equilibrium strategies among *homo moralis* with the same morality profile for aggregative games in Euclidean spaces.

5.1 Conditional independence

By conditional independence we here mean that the matching process is such that, for a given mutant, the types of any two *other* members in her group are statistically independent (in the given population state). Then,

$$\begin{aligned} q_m^0 &= \lim_{\varepsilon \to 0} \binom{n-1}{m-1} \left(\Pr\left[\tau|\tau,\varepsilon\right]\right)^{m-1} \left(1-\Pr\left[\tau|\tau,\varepsilon\right]\right)^{n-m} \\ &= \binom{n-1}{m-1} \sigma^{m-1} \left(1-\sigma\right)^{n-m} \end{aligned} \tag{21}$$

for any $n \geq 2$ and all $m \in \{1,..,n\}$, and where σ was defined in (3). In the appendix we present a matching process with the conditional statistical independence property.

Under conditional independence, evolution favors *homo moralis* preferences of a particularly simple morality profile, one that can be described with a single parameter, $\sigma \in [0,1]$**.** Indeed, the goal of evolutionarily stable *homo moralis* preferences is then to maximize her expected material payoff if others were to choose the same strategy as she does with probability σ and statistically independently of each other. From a mathematical viewpoint, *homo moralis* then defines a homotopy (see e.g. Munkres, 1975), parametrized by σ, between selfishness $\sigma = 0$ and Kantian morality, $\sigma = 1$. In this case, we refer to σ as the individual's *degree of morality*.

5.2 Differentiability

Suppose that X is a non-empty subset of $\mathbb{R}^k$ for some $k \in \mathbb{N}$. We will say that x is *strictly evolutionarily stable* (SES) if (19) holds for all $y \neq x$, and we will call a strategy $x \in X$ *locally strictly evolutionarily stable* (LSES) if (19) holds for all $y \neq x$ in some neighborhood of x. If, moreover, $\pi : X^n \to \mathbb{R}$ is differentiable, then so is $H : X^2 \to \mathbb{R}$, and standard calculus can be used to find evolutionarily stable strategies. Let $\nabla_y H(y, x)$ be the gradient of H with respect to y. We call this the *evolution gradient*; it is the gradient of the (average) material payoff to a mutant strategy y in a population state with residents playing x, and vanishingly few mutants. Writing "$\cdot$" for the inner product and boldface $\mathbf{0}$ for the origin, the following result follows from standard calculus:[20]

Proposition 2 *Let $X \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$, and let $x \in int(X)$. If $H : X^2 \to \mathbb{R}$ is continuously differentiable on a neighborhood of $(x, x) \in X^2$, then condition (i) below is necessary for x to be LSES, and conditions (i) and (ii) are together sufficient for x to be LSES. Furthermore, any strategy x for which condition (i) is violated is evolutionarily unstable.*

(i) $\nabla_y H(y, x)_{|y=x} = \mathbf{0}$,

(ii) $(x - y) \cdot \nabla_y H(y, x) > 0$ *for all* $y \neq x$ *in some neighborhood of* x.

The first condition says that there should be no direction of marginal improvement in material payoff for a rare mutant at the resident type. The second condition ensures that if some nearby rare mutant $y \neq x$ were to arise in a vanishingly small population share, then the mutant's material payoff would be increasing in the direction leading back to the resident type, x.

Conditions (i) and (ii) in Proposition 2 can be used to obtain remarkably simple and operational and conditions for evolutionarily stable strategies if the strategy set X is one-dimensional ($k = 1$) and π is continuously differentiable. Writing π_j for the partial derivative of π with respect to its j^{th} argument, one obtains:[21]

---

[20] See, e.g., Theorem 2 in Section 7.4 of Luenberger (1969), which also shows that Proposition 2 in fact holds when the gradient is the Gateaux derivative in general vector spaces

[21] Symmetry of π implies that $\pi_n(\hat{\mathbf{x}}) = \pi_j(\hat{\mathbf{x}})$ for all $j > 1$.

Proposition 3 *Assume conditionally independent matching with index of assortativity σ, and suppose that π is continuously differentiable on a neighborhood of $\hat{\mathbf{x}} \in X^n$, where $X \subseteq \mathbb{R}$. If $\hat{x} \in int\,(X)$ is evolutionarily stable, then*

$$\pi_1(\hat{\mathbf{x}}) + \sigma \cdot (n-1) \cdot \pi_n(\hat{\mathbf{x}}) = 0, \tag{22}$$

where $\hat{\mathbf{x}}$ is the n-dimensional vector whose components all equal $\hat{x}$. If $\hat{x} \in int\,(X)$ does not satisfy (22), then $\hat{x}$ is evolutionarily unstable.

Proof: If π is continuously differentiable, H is continuously differentiable. Hence, if $x \in int\,(X)$, Proposition 2 holds, and the following condition is necessary for x to be an evolutionarily stable strategy:

$$\nabla H_y(y,x)_{|y=x} = \sum_{m=1}^{n} \binom{n-1}{m-1} \sigma^{m-1} (1-\sigma)^{n-m} \left[\sum_{j=1}^{m} \pi_j\left(y, \mathbf{y}^{(m-1)}, \mathbf{x}^{(n-m)}\right) \right]_{|y=x} = 0.$$

Since π is aggregative, this equation may be written

$$\sum_{m=1}^{n} \binom{n-1}{m-1} \sigma^{m-1} (1-\sigma)^{n-m} \left[\pi_1(\mathbf{x}) + (m-1)\,\pi_n(\mathbf{x})\right] = 0, \tag{23}$$

where $\mathbf{x}$ is the n-dimensional vector whose components all equal x. Since

$$\sum_{m=1}^{n} \binom{n-1}{m-1} \sigma^{m-1} (1-\sigma)^{n-m} (m-1) = (n-1) \cdot \sigma,$$

the expression in (23) simplifies to $\pi_1(\mathbf{x}) + (n-1) \cdot \sigma \cdot \pi_n(\mathbf{x}) = 0$. **Q.E.D.**

Let $\theta = \sigma$ denote *homo moralis* with degree of morality σ. Together with Corollaries 1 and **??**, the preceding proposition implies:

Corollary 2 *Suppose that $X \subset \mathbb{R}$ is an open set, that π is continuously differentiable, and that $\beta_\sigma(x)$ is a singleton for each $x \in X_\sigma$. Then the set X_σ coincides with the set of evolutionarily stable strategies. Moreover, each $x \in X_\sigma$ must satisfy (22).*

6 Examples

6.1 Public goods

Consider a game in which each individual makes a contribution (or exerts an effort) at some personal cost, and where the sum of all contributions give rise to a benefit to all. More

specifically, letting $x_i \geq 0$ denote the contribution of individual i, $\mathbf{x}_{-i}$ the vector of others' contributions, and with $X = (0, +\infty)$, let

$$\pi(x_i, \mathbf{x}_{-i}) = B\left(\sum\nolimits_{j=1}^{n} x_j\right) - C(x_i)$$

for some continuous (benefit and cost) functions $B, C : (0, +\infty) \to \mathbb{R}_+$ that are twice differentiable with $B', C' > 0$, $B'' \leq 0$ and $C'' \geq 0$, with at least one of the two last inequalities strict. Under conditionally independent assortativity, the associated function H (see (20)) is strictly concave, implying that (22) is both necessary and sufficient for an individual contribution $\hat{x} > 0$ to be evolutionarily stable. The relevant partial derivatives are

$$\pi_1(\hat{\mathbf{x}}) = B'(n\hat{x}) - C'(\hat{x}) \text{ and } \pi_n(\hat{\mathbf{x}}) = B'(n\hat{x}),$$

so a contribution $\hat{x} > 0$ is evolutionarily stable if and only if

$$[1 + (n-1)\sigma] \cdot B'(n\hat{x}) = C'(\hat{x}). \tag{24}$$

This equation has at most one solution, and it has a unique solution $\hat{x} > 0$ if $[1 + (n-1)\sigma] \cdot B'(0) > C'(0)$, an arguably natural condition in many applications, and which we henceforth assume to be met.[22] Under this condition, the unique evolutionarily stable contribution is increasing in the index of assortativity.

Since the conditions stated in Corollary 2 are satisfied, $X_\sigma = \{\hat{x}\}$. For $\sigma = 0$, equation (24) is nothing but the standard formula according to which "own marginal benefit" equals "own marginal cost"; $\hat{x}$ then corresponds to what *homo oeconomicus* would do when playing against other *homo oeconomicus*. At the other extreme, for $\sigma = 1$, the benevolent social planner's solution obtains; then $\hat{x}$ solves $\max_{x \in X} [B(nx) - C(x)]$. For intermediary values of σ, intermediary values of $\hat{x}$ obtain, and this may or may not be decreasing in group size n.

To see this, consider the case when both B and C are power functions; let $B(x) \equiv x^b$ and $C(x) \equiv x^c$ for some $b \in (0, 1]$ and $c \geq 1$ such that $b < c$. Then the unique evolutionarily stable individual contribution is

$$\hat{x} = \left(\frac{b}{c} \cdot \left[(1-\sigma) n^{b-1} + \sigma n^b\right]\right)^{1/(c-b)}.$$

[22] We also note that this holds true even if B would be linear, granted $C'' > 0$. For although others' contributions are then strategically irrelevant for the individual player, a positive index of assortativity makes the individual willing to contribute more than under uniform random matching.

This contribution is decreasing (increasing) in group size n in the extreme cases when the index of assortativity is zero (one). However, the individual contribution is not monotonic for all σ. See diagram below, showing $\hat{x}$ as a function of n for $\sigma = 0$, 0.25, 0.5, 0.75 and 1 (higher curves for higher σ).[23]

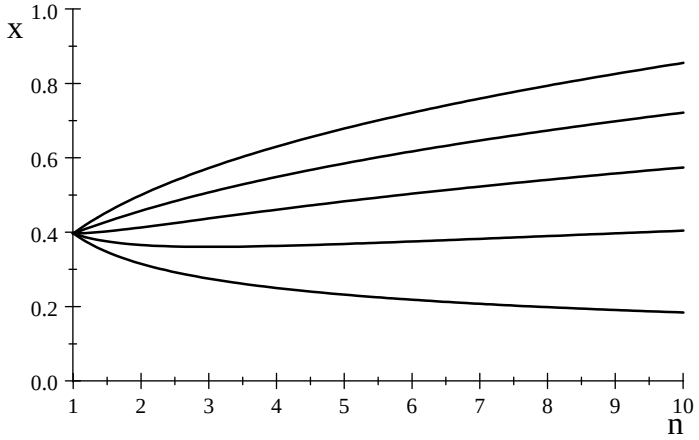

Figure 1: The evolutionarily stable individual contribution in the public-goods game.

To see why this non-monotonicity arises, we study $d\hat{x}/dn$, which has the same sign as $\sigma bn - (1-\sigma)(1-b)$. Consider first the special case of a linear benefit function ($b = 1$). Then the evolutionarily stable contribution is increasing in group size n for any $\sigma > 0$. The reason is that an increase in group size then additively increases the marginal benefit from increasing one's own contribution, by simply increasing the number of individuals who benefit from the public good. This effect is also present for non-linear benefit functions ($b < 1$), but then there is also an opposite effect, namely that the marginal benefit of an individual's contribution decreases as the others' contributions increase. For b small enough, this second effect outweighs the first effect; to see this, note that as b tends to zero, the evolutionarily stable contribution tends to be strictly decreasing in n for all n (granted $0 < \sigma < 1$).

Finally, there is a third effect at work, an effect which helps explain why in some cases the evolutionarily stable individual contribution may be decreasing in n when n is small (see diagram), but increasing in n for large values of n. Indeed, we see that $d\hat{x}/dn$ changes sign as n reaches the value $(1-\sigma)(1-b)/(\sigma b)$. The intuition for this is that, beyond the effect of group size n on the mean value of *homo moralis'* hypothetical number of

[23] The diagram has been drawn for $b = 0.5$ and $c = 2$.

others who contribute likewise, $(n-1)\,\sigma$, there is an effect on the variance of this number, $(n-1)\,\sigma\,(1-\sigma)$, and hence on the risk that others might not contribute much. Indeed, a vanishingly rare mutant faces considerable uncertainty as to the contributions his opponents will make, when n is small. For $n=2$, the uncertainty is hefty; a mutant's opponent either makes the same contribution or the resident contribution. As n increases, the mutant's uncertainty becomes less hefty, since then the variance of the share of other contributors tends to zero. For n small, the riskiness may strong enough to reduce homo moralis' incentive to contribute more when when the group is larger.

Remark 1 *The public goods interaction described here is symmetric. However, as noted before, our general model also applies to asymmetric interactions as long as these are* ex ante symmetric*, i.e. such that each individual at the outset is just as likely to be cast in either player role (as, for instance, in a laboratory experiment). To illustrate, suppose that only some individuals are free to give a contribution. More precisely, let* $\tilde{A} \subset \{1,...,n\}$ *denote the random set of active players. Suppose further that* ex ante*, each individual faces the same probability* $p \in (0,1)$ *to get an active player role, that is, to be in the set* $\tilde{A}$*. A player's strategy, is then a contribution to make if called upon to be active (without being told who else is active). Let* x_i *denote player* i*'s strategy so defined. We may then write the ex ante payoff function of any player* i *in the symmetric form*

$$\pi\left(x_i, \mathbf{x}_{-i}\right) \;=\; p\cdot\mathbb{E}\left[B\left(\sum\nolimits_{j\in\tilde{A}} x_j\right) \mid i\in\tilde{A}\right] - p\cdot C\left(x_i\right) + \;(1-p)\cdot\mathbb{E}\left[B\left(\sum\nolimits_{j\in\tilde{A}} x_j\right) \mid i\notin\tilde{A}\right],$$

where the expectation is taken with respect to the random draw of the subset $\tilde{A}$*.*

6.2 Team work

Suppose instead that the jointly produced good in the previous example is a private good, split evenly between the members of the group or team. The same analysis applies, with the only difference that the individual benefit be divided by n. One then obtains the following necessary and sufficient condition for the evolutionarily stable individual contribution:

$$\left[1+(n-1)\,\sigma\right]\cdot B'\left(n\hat{x}\right) = n\cdot C'\left(\hat{x}\right).$$

Comparing this with the public goods case (equation (24)), we note that the evolutionarily stable individual contribution now is smaller, that it is still increasing in the index of assortativity, and that it is now necessarily decreasing in group size.

6.3 Contests

Many real interactions involve competing for a prize. Examples include competition between job seekers for a vacancy, between firms for a contract, between employees for promotion, etc. Such interactions may be modeled as a contest in which each participant makes a nonnegative effort at some personal cost, and where each participant's effort probabilistically translates to a "result," and the participant with the "best" result wins the prize. More specifically, let $x_i \geq 0$ be participant i's effort, $\mathbf{x}_{-i}$ the vector of efforts of the others, and let $\tilde{y}_i = x_i + \varepsilon_i$ be participant i's result (as valued by the "umpire"). With absolutely continuously distributed random terms, ties occur with probability zero. For quadratic costs of effort, the material payoff to participant i is:

$$\pi\left(x_i, \mathbf{x}_{-i}\right) = b \cdot \Pr\left[\tilde{y}_i > \tilde{y}_j \ \forall j \neq i\right] - \frac{1}{2}x_i^2 \tag{25}$$

where $b > 0$ is the value of the prize in question. This defines a continuously and (infinitely) differentiable function on $X^n = \mathbb{R}_+^n$. For Gumbel distributed random terms, the winning probability for each participant i satisfies[24]

$$\Pr\left[\tilde{y}_i > \tilde{y}_j \ \forall j \neq i\right] = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \quad \forall x \in X^n.$$

From this it is easily verified that a necessary condition (22) for an effort level $\hat{x} > 0$ to be evolutionarily stable boils down to

$$\hat{x} = \frac{1-\sigma}{n} \cdot \left(1 - \frac{1}{n}\right) \cdot b. \tag{26}$$

The evolutionarily stable individual effort is proportional to the value b of the price, linearly decreasing (towards zero) in the index of assortativity, σ, and decreasing in n (for all $n \geq 2$). Aggregate effort, however, is increasing in n.

6.4 Helping others

People often help others, also when no reward or reciprocation is expected. To model such behaviors, consider a group of n *ex ante* identical individuals, and suppose that with some exogenous probability $p \in (0,1)$ exactly one individual loses one unit of wealth, with equal probability for all individuals when this happens. The $n-1$ others observe this event, and

[24] This is a standard result in random utility theory, see, e.g., Anderson et al. (1992).

each of them may then help the unfortunate individual by transferring some personal wealth. These decisions are voluntary and simultaneous. For any individual level of wealth $w \geq 0$, let $U(w)$ be the individual's indirect utility from consumption, where U meets the usual Inada conditions.

We model this as a game where initial wealth is normalized to unity:

$$\pi(x_i, \mathbf{x}_{-i}) = (1-p) \cdot U(1) + p \cdot \left[\left(1 - \frac{1}{n}\right) U(1 - x_i) + \frac{1}{n} U\left(\sum_{j \neq i} x_j \right) \right]$$

Here $x_i \geq 0$ is i's voluntary transfer in case another individual is hit by the wealth loss. Applying equation (24), for an individual transfer $\hat{x} \in (0,1)$ to be evolutionarily stable, it must satisfy

$$U'(1 - \hat{x}) = \sigma \cdot U'[(n-1)\hat{x}]. \tag{27}$$

This equation uniquely determines $\hat{x} \in (0,1)$, since the left-hand side is continuously and strictly increasing in $\hat{x}$, from $U'(1)$ towards plus infinity, and the right-hand side is continuously and strictly decreasing in $\hat{x}$, from plus infinity to $U'(n-1)$. It follows immediately from (27) that this transfer is an increasing function of the index of assortativity σ and a decreasing function of group size n. Both effects are intuitively expected; higher assortativity makes helpfulness more worthwhile and more individuals watching the wealth-loss makes free-riding among them the more severe. In the special case when indirect utility is a power function, $U(w) \equiv w^a$ for some $a \in (0,1)$, one obtains

$$\hat{x} = \frac{\sigma^{1/(1-a)}}{n - 1 + \sigma^{1/(1-a)}}.$$

While no transfers are given under uniform random matching ($\sigma = 0$), post-transfer wealth levels are equalized when $\sigma = 1$, so full insurance then holds, while partial insurance obtains for intermediate values of σ. Furthermore, it is easy to verify that the aggregate transfer, $n\hat{x}$, is increasing in n and converges to $\sigma^{1/(1-a)}$ as $n \to \infty$.

7 Literature

When introduced by Maynard Smith and Price (1973) the concept of evolutionary stability was defined as a property of *mixed strategies* in *finite and symmetric two-player games* played under *uniform random matching* in an *infinite population*, where uniform random matching means that the probability for an opponent's strategy does not depend on one's own strategy. Broom, Cannings and Vickers (1997) generalized Maynard Smith's and Price's original

definition to *finite and symmetric n-player games*, for $n \geq 2$ arbitrary, while maintaining the assumption of uniform random matching in an infinite population.[25] They noted the combinatorial complexity entailed by this generalization, and reported some new phenomena that can arise when interactions involve more than two parties. Evolutionary stability and asymptotic stability in the replicator dynamic, in the same setting, was further analyzed by Bukowski and Miekisz (2004). Schaffer (1988) extended the definition of Maynard Smith and Price to the case of uniform random matching in *finite populations*, and also considered interactions involving all individuals in the population ("playing the field"). Grafen (1979) and Hines and Maynard Smith (1979) generalized the definition of Maynard Smith and Price from uniform random matching to the kind of *assortative matching* that arises when strategies are genetically inherited and games are played among kin.

Our model generalizes most of the above work within a unified framework. To see this, note first how our definition of evolutionarily stability relates to Maynard Smith's and Price's (1973) original definition of an evolutionarily stable (mixed) strategy in a symmetric and finite two-player game under uniform random matching. Suppose thus that X is the unit simplex of mixed strategies in such a game and let $\Theta = X$, that is, let a type be a mixed strategy (as if individuals were "programmed" to strategies). For any population state $s = (x, y, \varepsilon) \in X^2 \times (0, 1)$, the set $V(s)$ of possible material-payoff pairs is then a singleton. Its unique element $v \in V(s)$ has components $v_1 = (1-\varepsilon)\pi(x, x) + \varepsilon\pi(x, y) = \pi[x, (1-\varepsilon)x + \varepsilon y]$ and $v_2 = (1-\varepsilon)\pi(y, x) + \varepsilon\pi(y, y) = \pi[yx, (1-\varepsilon)x + \varepsilon y]$. In other words, v_1 (resp. v_2) is the "post-entry" expected material payoff to strategy x (resp. y). By Definition 1, x is evolutionarily stable against y if $\varphi(x, y, \varepsilon) > 0$ for all $\varepsilon > 0$ sufficiently small, which is equivalent with being evolutionarily stable in the sense of Maynard Smith and Price (1973). Suppose a strategy x is unstable in the sense of Definition 1. Since φ is here continuous, there then exists a strategy $y \neq x$ such that $\varphi(x, y, \varepsilon) < 0$ for all $\varepsilon > 0$ sufficiently small, that is, such that this mutant's post-entry expected material payoff exceeds that of the resident strategy x whenever the mutant appears in sufficiently small population shares. Second, note that the functions F and G (see (12) and (13)) are generalizations, from uniform to assortative matching, of the functions used by Broom, Cannings and Vickers (1997) in their definition of an evolutionarily stable strategy in symmetric and finite n-player games (here x and y may be mixed strategies in a finite game).

[25] Precursors to their work are Haigh and Cannings (1989), Cannings and Whittaker (1995) and Broom, Cannings and Vickers (1996).

In a pioneering study, Güth and Yaari (1992) defined evolutionary stability for parametrized *utility functions*, assuming uniform random matching and complete information.[26] This approach is often referred to as "indirect evolution." The literature on preference evolution now falls into four broad classes, depending on whether the focus is on interactions where information is complete[27] or incomplete[28], and whether non-uniform random matching is considered.[29] Few models deal with interactions involving more than two individuals. Like here, the articles in this category focus exclusively on interactions that are symmetric in material payoffs, the payoffs that drive evolution. Unlike us, they restrict attention to uniform random matching. Koçkesen, Ok, and Sethi (2000a,b) show that under complete information about opponents' preferences, players with a specific kind of interdependent preferences fare better materially than players who seek to maximize their material payoff. Sethi and Somanathan (2001) go one step further and characterize sufficient conditions for a population of individuals with the same degree of reciprocity to withstand the invasion of selfish individuals, again in a complete information framework. By contrast, Ok and Vega-Redondo (2001) analyze the case of incomplete information. They identify sufficient conditions for a population of selfish individuals to withstand the invasion by non-selfish individuals, and for selfish individuals to be able to invade a population of identical non-selfish individuals.

8 Conclusion

To understand human societies it is necessary to understand human motivation. In this paper we build on a large literature in biology and in economics, initiated by Maynard Smith and Price (1973), to propose a theoretical framework within which one may study the evolution of human motivational types by way of natural selection. The framework is based upon a

[26] See also Frank (1987).

[27] See Robson (1990), Güth and Yaari (1992), Ockenfels (1993), Huck and Oechssler (1996), Ellingsen (1997), Bester and Güth (1998), Fershtman and Judd (1987), Fershtman and Weiss (1998), Koçkesen, Ok and Sethi (2000a,b), Bolle (2000), Possajennikov (2000), Sethi and Somanathan (2001), Heifetz, Shannon and Spiegel (2007a,b), Akçay et al. (2009), Alger (2010), and Alger and Weibull (2010, 2012).

[28] See Ok and Vega-Redondo (2001), Dekel, Ely and Yilankaya (2007), and Alger and Weibull (2013).

[29] In the literature cited in the preceding two footnotes, only Alger (2010), Alger and Weibull (2010, 2012, 2013) allow for non-uniform random matching. Bergstrom (1995, 2003) also allows for such assortative matching, but he restricts attention to strategy rather than preference evolution.

general definition of an evolutionarily stable type, where an individual's type guides his or her behavior in interactions in groups of any size. The framework may be applied to interactions where others' preferences are known or unknown, and it allows for assortativity in the process by which individuals are matched together to interact. Since our analysis focuses on whether a homogenous population may withstand a small-scale invasion of individuals of a different type, a key factor is the probability with which mutants are matched with other mutants when these are vanishingly rare. In two-player interactions, such assortativity is simply the probability that the individual with whom a mutant interacts also is a mutant (the index of assortativity; Bergstrom, 2003). We generalize this notion to n-player interactions by defining the *assortativity profile* of an n-party matching process, for which the assortativity profile is a vector that provides the probabilities that none, some, or all the individuals with whom a mutant interacts also are mutants, in the limit as the share of mutants in the population tends to zero. There is some assortativity as soon as the probability that at least one of the individuals with whom a mutant interacts also is a mutant is positive.

We apply the framework to preference evolution when an individual's preferences are his or her private information. The set of potential preferences is taken to be the set of all continuous and aggregative preferences over strategy profiles. Our analysis shows that a particular preference comes out as a clear winner in the evolutionary race. This preference belongs to the class of *homo moralis* preferences, according to which an individual maximizes a weighted sum of the material payoff that she would obtain if none, some, or all the individuals with whom she interacts would do as she does; the weights represent the individual's morality profile. More precisely, we find that (a) *homo moralis* preferences with a morality profile that reflects the assortativity in the matching process are evolutionarily stable, and (b) under quite weak assumptions, any preferences that lead to different behaviors from that of this *homo moralis* are evolutionarily unstable. Furthermore, equilibrium behavior in a homogeneous population consisting of *homo moralis* with this type of morality is the same as under strategy evolution.

Interestingly, then, our analysis shows that group size has no effect on what class of preferences is favored by evolution when preferences are the interacting individuals' private information; *homo moralis* preferences with a morality profile equal to the assortativity profile stand out as the clear winner in the evolutionary race, independent of group size and of the (material) game played. By contrast, as shown in the examples, group size does affect equilibrium behavior, in groups consisting of identical *homo moralis*. Assuming conditional independence in the matching process, we found that, for any positive index of assortativity

σ, the evolutionarily stable variety of *homo moralis* contributes more than *homo oeconomicus* in public-goods games and also when in team work. By contrast, she exerts less effort in contests and supplies less output in Cournot markets. She is also helpful to others who have been exposed to an exogenous hazard. Moreover, these effects do not generally vanish as group size n increases. This is because *homo moralis* behaves as if she, roughly speaking thought "what would happen if the share σ of the other group members would do like me?" when contemplating her strategy choice.[30]

Although quite general, our model relies on a number of simplifying assumptions. Relaxation of these is a task that has to be left for future research. Moreover, we only apply our general definition of evolutionary stability to two cases, strategy evolution and preference evolution when preferences are private information. Applications to complete or partially incomplete information are called for, in particular in settings where the random matching is not exogenous, as here, but at least partly endogenous. This is a major analytical challenge, however, opening the door to signalling and mimicry, a very rich, important and exciting research area. Yet another challenge would be to investigate evolutionary neutrality, setwise evolutionary stability and/or evolutionary stability properties of heterogenous population states.

For the past twenty years or so economists have proposed varieties of pro-social or other-regarding preference in order to explain certain observed behaviors, mostly in laboratory experiments but sometimes in the field, that are at odds with maximization of one's own material payoff. Our research has so far delivered two results of relevance for behavioral economics. First, the result that natural selection selects preferences with a distinct Kantian flavor; it is as if individuals in their strategy choice attach some importance to "what would happen if others did what I do?" *Homo oeconomicus* is an extreme case; to place no importance at all to this Kantian morality aspect. Second, we have the result that the importance that individuals attach to this Kantian morality aspect depends on the assortativity in the matching process, and is independent of the interaction in question. Since historically, assortativity arguably has varied between populations and over time (depending on geography, technology and social structure), this second result suggests that one should expect the importance of morality to differ accordingly. Likewise, our results suggest that if in a population

[30]We here invoke the law of large numbers, which holds under conditional independence, but arguing heuristically, as if the expected value of the average of a function has the same qualitative features as the function evaluated at the average point.

individuals interact in several different games, and assortativity differs between the games (e.g., sharing with relatives, and engaging in market interactions with strangers), then one should expect different levels of morality in the different games. We hope that our theoretical results, combined with empirical and experimental work, will enhance the understanding of human behavior and motivation.

9 Appendix: A class of matching processes

Let n, I and P be integers greater than one, and imagine a finite population consisting of P individuals. The population is divided into "islands," each island consisting of $I > n$ individuals, and P is some multiple of I. Initially all individuals are of type θ. Suddenly a mutation to another type τ occurs in one of the islands, and only there. Each individual on that island has probability ρ of mutating, and individual mutations are statistically independent. Hence, the random number M of mutants is binomially distributed $M \sim Bin(I, \rho)$. We note that in this mutation process the random number M is also the total number of mutants in the population at large, so the population share M/P of mutants is a random variable with expectation $\varepsilon = \mathbb{E}[M/P] = \rho I/P$. A group of size n is now formed to play a game $\Gamma = \langle N, X^n, \pi \rangle$ (as described in Section 2) as follows, and this is an event that is statistically independent of the above-mentioned mutation. First, one of the islands is selected, with equal probability for each island. Secondly, n individuals from the selected island are recruited to form the group, drawn as a random sample without replacement from amongst the I islanders and with equal probability for each islander to be sampled.

Consider an individual who has been recruited to the group. Let $X \in \{\theta, \tau\}$ denote the individual's type. If $X = \tau$, it is necessary that $M > 0$ and that the individual is from the island where the mutation occurred, so the random number of *other* mutants in her group is binomially distributed, $Bin(n-1, \rho)$. With T denoting the total number of mutants in her group, we have, for $m = 1, 2, ..., n$:

$$\Pr[T = m \mid X_i = \tau] = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \rho^{m-1} (1-\rho)^{n-m}. \tag{28}$$

If instead $X = \theta$, then $M = 0$ is possible and she may well be from another island than where the mutation occurred. We thus have

$$\Pr[T = m \mid X_i = \theta] \leq \frac{I}{P} \cdot \begin{pmatrix} n-1 \\ m \end{pmatrix} \rho^{m} (1-\rho)^{n-m-1}$$

for all $m > 0$. Moreover, for any two group members i and j:

$$\Pr[X_j = \tau \mid X_i = \tau] = \rho \quad \text{and} \quad \Pr[X_j = \tau \mid X_i = \theta] \leq \rho I/P.$$

Keeping ρ, n and I constant, we may write $\Pr[\theta|\theta, \varepsilon]$ for $\Pr[X_j = \theta \mid X_i = \theta]$ and $\Pr[\theta|\tau, \varepsilon]$ for $\Pr[X_j = \theta \mid X_i = \tau]$, and these conditional probabilities are continuous functions of $\varepsilon = \rho I/P$. In addition, we have $1 - \varepsilon \leq \Pr[\theta|\theta, \varepsilon] \leq 1$ and $\Pr[\theta|\tau, \varepsilon] = 1 - \rho$. Letting $P \to \infty$, we obtain $\varepsilon \to 0$ and $\Pr[\theta|\theta, \varepsilon] \to 1$. Hence, $\lim_{\varepsilon \to 0} \phi(\varepsilon) = \rho$, so in this example the index of assortativity is $\sigma = \rho$.

References

Acemoglu, D. and M.K. Jensen (2013) "Aggregate comparative statics," *Games and Economic Behavior*, 81, 27 - 49.

Akçay, Erol, Jeremy Van Cleve, Marcus W. Feldman, and Joan Roughgarden (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. and J. Weibull (2010): "Kinship, Incentives and Evolution," *American Economic Review*, 100, 1725-1758.

Alger, I. and J. Weibull (2012): "A Generalization of Hamilton's Rule—Love Others How Much?" *Journal of Theoretical Biology*, 299, 42-54.

Alger, I. and J. Weibull (2013): "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching," *Econometrica*, 81:2269-2302.

Aliprantis, C.D. and K.C. Border (2006): *Infinite Dimensional Analysis*, 3rd ed. New York: Springer.

Anderson, S.P., A. de Palma, and J.-F. Thisse (1992): *Discrete Choice Theory of Product Differentiation.* VCambridge (USA): MIT Press.

Andreoni, J. (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 100, 464-477.

Becker, G. (1976): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology," *Journal of Economic Literature*, 14, 817–826.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652-1678.

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

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

Bergstrom, T. (2009): "Ethics, Evolution, and Games among Neighbors," Working Paper, UCSB.

Bergstrom, T. (2013): "Measures of Assortativity," *Biological Theory,* 8, 133-141.

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

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.

Bomze, I., and B. Pötscher (1989): *Game Theoretical Foundations of Evolutionary Stability.* New York: Springer.

Bramoullé, Y., and B. Rogers (2009): "Diversity and Popularity in Social Networks," Discussion Papers 1475, Northwestern University, Center for Mathematical Studies in Economics and Management Science.

Brekke, K.A., S. Kverndokk, and K. Nyborg (2003): "An economic model of moral motivation," *Journal of Public Economics*, 87, 1967-1983.

Broom, M., C. Cannings and G.T. Vickers (1996): "Choosing a Nest Site: Contests and Catalysts", *Amer. Nat.* 147, 1108-1114.

Broom, M., C. Cannings and G.T. Vickers (1997): "Multi-Player Matrix Games", *Bulletin of Mathematical Biology* 59, 931-952.

Bukowski, M., and J. Miękisz (2004): "Evolutionary and asymptotic stability in symmetric multi-player games", *International Journal of Game Theory* 33, 41-54.

Cannings, C., and J.C. Whittaker (1995): "The Finite Horizon War of Attrition", *Games and Economic Behavio*r 11, 193-236.

Corchón, L. (1996): *Theories of Imperfectly Competitive Markets.* Berlin: Springer Verlag.

Currarini, S., M.O. Jackson, and P. Pin (2009): "An Economic Model of Friendship: Homophily, Minorities and Segregation," *Econometrica,* 77, 1003–1045.

Currarini, S., M.O. Jackson, and P. Pin (2010): "Identifying the Roles of Race-Based Choice and Chance in High School Friendship Network Formation," *Proceedings of the National Academy of Sciences,* 107, 4857–4861.

Day, T., and P.D. Taylor (1998): "Unifying Genetic and Game Theoretic Models of Kin Selection for Continuous types," *Journal of Theoretical Biology*, 194, 391-407.

Dekel, E., J.C. Ely, and O. Yilankaya (2007): "Evolution of Preferences," *Review of Economic Studies*, 74, 685-704.

Dubey, P., A. Mas-Colell, and M. Shubik (1980): "Efficiency Properties of Strategic Market Games", *Journal of Economic Theory* 22, 339-362.

Duffie, D. and Y. Sun (2012): "The Exact Law of Large Numbers for Independent Random Matching", *Journal of Economic Theory* 147, 1105-1139.

Ellingsen, T. (1997): "The Evolution of Bargaining Behavior," *Quarterly Journal of Economics*, 112, 581-602.

Fehr, E., and K. Schmidt (1999): "A theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817-868.

Fershtman, C. and K. Judd (1987): "Equilibrium Incentives in Oligopoly," *American Economic Review*, 77, 927–940.

Fershtman, C., and Y. Weiss (1998): "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics*, 70, 53-73.

Frank, R.H. (1987): "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *American Economic Review*, 77, 593-604.

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

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

Güth, W., and M. Yaari (1992): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," in U. Witt. *Explaining Process and Change – Approaches to Evolutionary Economics.* Ann Arbor: University of Michigan Press.

Haigh, J., and C. Cannings (1989): "The n-Person War of Attrition", *Acta Applic. Math.* 14, 59-74.

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.

Heifetz, A., C. Shannon, and Y. Spiegel (2007a): "The Dynamic Evolution of Preferences," *Economic Theory*, 32, 251-286.

Heifetz, A., C. Shannon, and Y. Spiegel (2007b): "What to Maximize if You Must," *Journal of Economic Theory*, 133, 31-57.

Hines, W.G.S., and J. Maynard Smith (1979): "Games between Relatives," *Journal of Theoretical Biology*, 79, 19-30.

Huck, S., and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior,* 28, 13–24.

Jackson, M.O., and A. Watts (2010): "Social Games: Matching and the Play of Finitely Repeated Games," *Games and Economic Behavior,* 70, 170-191.

Koçkesen, L., E.A. Ok, and R. Sethi (2000a): "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory,* 92, 274-299.

Koçkesen, L., E.A. Ok, and R. Sethi (2000b): "Evolution of Interdependent Preferences in Aggregative Games," *Games and Economic Behavior* 31, 303-310.

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

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

Maynard Smith, J., and G.R. Price (1973): "The Logic of Animal Conflict," *Nature,* 246:15-18.

McPherson, M., L. Smith-Lovin, and J.M. Cook (2001): "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415-444.

Munkres, James (1975): *Topology, a First Course.* London: Prentice Hall.

Ockenfels, P. (1993): "Cooperation in Prisoners' Dilemma—An Evolutionary Approach",

European Journal of Political Economy, 9, 567-579.

Ok, E.A., and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory,* 97, 231-254.

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

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

Robson, A.J., and B. Szentes (2014): "A Biological Theory of Social Discounting," forthcoming, *American Economic Review.*

Roemer, J.E. (2010): "Kantian equilibrium," *Scandinavian Journal of Economics*, 112, 1-24.

Ruef, M., H.E. Aldrich, and N.M. Carter (2003): "The Structure of Founding Teams: Homophily, Strong Ties, and Isolation among U.S. Entrepreneurs," *American Sociological Review*, 68, 195-222.

Sandholm, W. (2001): "Preference Evolution, Two-Speed Dynamics, and Rapid Social Change," *Review of Economic Dynamics,* 4, 637-679.

Schaffer, M.E. (1988): "Evolutionarily Stable Strategies for Finite Populations and Variable Contest Size," *Journal of Theoretical Biology*, 132, 467-478.

Schelling, T. (1960): *The Strategy of Conflict.* Cambridge: Harvard University Press.

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

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

Weibull, J.W. (2004): "Testing Game Theory", in S. Huck (ed.): *Advances in Understanding Strategic Behaviour: Game Theory, Experiments, and Bounded Rationality - Essays in Honor of Werner Güth.* Palgrave MacMillan.