Evolution of preferences in group-structured populations: genes, guns, and culture

Ingela Alger*, Jörgen W. Weibull†, Laurent Lehmann‡

September 4, 2018.§

Abstract

During human evolution, individuals interacted in groups connected by limited migration and sometimes conflicts. If the spread of preferences, from one generation to the next, depends on their material success, which preferences will prevail? Building on population biology models of spatially structured populations, and assuming preferences to be private information, we characterize which preferences, if any, cannot be displaced, once established. We find that such uninvadable preferences represent different motives when expressed in terms of fitness than when expressed in terms of material payoffs. At the fitness level, individuals appear be driven by a mix of self-interest and a Kantian motive**,** which involves evaluating one's behavior in light of what own fitness would be if others were to choose the same behavior. This Kantian motive is borne out from kin selection (be it genetic or cultural). At the material payoff level, individuals appear to be driven by these two motives, but also in part by an other-regarding motive towards own group members. This motive represents within-group spite or altruism. We show how population structure—group size, migration rates, probability of group conflicts, cultural loyalty towards parents—shape the relative importance of these motives.

Keywords: strategic interactions, preference evolution, evolution by natural selection, cultural transmission, pro-sociality, altruism, Kantian moral concerns, spite.

JEL codes: A12, A13, B52, C73, D01, D63, D64, D91.

*Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France, and Institute for Advanced Study in Toulouse. ingela.alger@tse-fr.eu

†Stockholm School of Economics, and Institute for Advanced Study in Toulouse. jorgen.weibull@hhs.se

‡Department of Ecology and Evolution, University of Lausanne, Switzerland. laurent.lehmann@unil.ch

§All authors together conceived the model, I.A. and L.L. derived the main results, and I.A. wrote the manuscript with input from all authors. We thank Lee Dinetan, Gustav Karreskog, Jonathan Newton, Jorge Peña, Peter Wikman, and seminar audiences at University of Gothenburg, Université Catholique de Louvain, as well as participants at the conference "Neuroeconomics and the Biological Basis of Preferences and Strategic Behavior" at Simon Fraser University for helpful comments. Support by Knut and Alice Wallenberg Research Foundation and by ANR-Labex IAST is gratefully acknowledged. We also thank Agence Nationale de la Recherche for funding (Chaire d'Excellence ANR-12-CHEX-0012-01 for I. Alger, and Chaire IDEX ANR-11-IDEX-0002-02 for J. Weibull).

1 Introduction

Preferences are fundamental to economics, and are usually treated as primitives.[1] Behavioral economics has proposed and studied a rich array of human behavioral motivations, notably altruism (Becker, 1976), warm glow (Andreoni, 1990), a concern for fairness (Rabin, 1993), reciprocal altruism (Levine, 1998), inequity aversion (Fehr and Schmidt, 1999), identity concerns (Akerlof and Kranton, 2000), moral motivation (Brekke, Kverndokk, and Nyborg, 2003), self-image concerns (Bénabou and Tirole, 2006), and a concern for honesty (Alger and Renault, 2006). However, little is yet known about their evolutionary background. What preferences should individuals be expected to have if they are transmitted across generations, and preferences that increase individual survival and reproduction spread at the expense of other preferences? Answering this question is essential to understanding what are the ultimate drivers of human behavior in social and economic interactions (e.g., Hirshleifer, 1977, Alexander, 1979, Bergstrom, 1996, Robson, 2001).

Suppose that an individual's material returns or payoffs in interactions with others enhance the individual's reproductive success (survival and number of offspring). It may then appear plausible that biological evolution of genetically transmitted traits would favor individuals who strive to maximize their own material returns or payoffs. In formal models, it has been established that the following three conditions are sufficient for this conjecture to be valid: *(i)* the population is very large and homogeneous (no subdivision by sex, age, size, etc.) and reproduction is clonal, *(ii)* interacting individuals do not know each other's preferences or goal functions, and *(iii)* interactions are uniformly random in the population, in the sense that each encounter is just as likely (see Ok and Vega-Redondo, 2001, Dekel, Ely, and Yilankaya, 2007). Except for knife-edge cases, these conditions are also necessary. However, they are rarely, if ever, met, even as approximations. Indeed, all natural populations (human or otherwise) are structured into groups, connected to each other by limited migration and in some species there are also occasional group conflicts. Limited migration causes limited genetic and/or cultural mixing in the population (Cavalli-Sforza and Bodmer, 1971, Hartl and Clark, 2007), and this in turn has profound consequences for the evolution of traits.

Spatial population structure imposes constraints on the ways in which traits—including strategies and preferences—can spread. For example, consider a genetically transmitted trait and suppose that in an initially homogenous group (or location) suddenly a new trait

[1] Throughout this paper we use concepts and terminology that are standard in economics, and model behavior as a choice of action (or stream of actions) from a set of feasible actions, where this choice is guided by a striving to maximize some goal (utility) function. The utility function together with the information and the constraints imposed by the environment are thus what biologists would call the proximate causes driving behavior. Furthermore, by contrast to the evolutionary biology literature where the terms "altruism" and "spite" are used to refer to the fitness consequences of a behavior on the actor and others, in economics they are used to describe the proximate causes behind behaviors. Thus, in economics, an individual who has a utility function which puts a positive weight on another individual's material well-being is altruistic; and an individual who has a utility function which puts a negative weight on another individual's material well-being is spiteful. For further discussion of the meaning of these terms in different academic disciplines, we refer to West et al. (2007) and Bshary and Bergmüller (2008).

appears in one individual. Since children acquire their traits from their parents, in the second generation some interactions between carriers of the new trait may occur between siblings and in the third among cousins. Hence, locally interacting individuals are more likely to share the same new trait (through their common ancestor), even when rare in the population at large, than are individuals sampled at random from the whole population. Such assortative matching between interacting individuals tends to favor behavior that promotes the survival and/or reproductive success of others in their group or location, since such behavior is more likely to benefit individuals carrying the new trait than when matching is uniformly random (Hamilton, 1964, 1971, Grafen, 1985, Frank, 1998, Rousset 2004). This is the so-called mechanism of *kin selection* in evolutionary biology (Maynard-Smith 1964). In this literature, assortative matching between pairs of individuals is usually quantified by the *coefficient of relatedness*–the fraction of their genes inherited from a local common ancestor– a quantity depending on features of the population structure, such as the migration rate and local group size.

By the same token, however, individuals who share a local common ancestor are also more likely to be competing for the same local resources than are individuals drawn at random from the population at large. In the above-mentioned genetic inheritance example, siblings are likely to compete for the same resources. Since interacting groups are finite, such local competition tends to favor behavior that hinders the survival and/or reproductive success of others in the group or location, since achieving an edge over neighbors then translates into evolutionary success (Hamilton, 1971, Schaffer, 1988, Frank, 1998, Rousset 2004). As assortative matching and local competition can, in general, not be separated, their joint effects need to be taken into account in order to understand the evolutionary success of traits under limited dispersal, a question that has received much attention in the evolutionary biology literature (see e.g. Hamilton 1967 and Taylor 1992 for pioneering and paradigmatic examples, and Frank 1998 and Rousset, 2004, for general theoretical treatments).

While clearly relevant for the understanding of the evolution of motivational factors in social interactions, this literature is yet of limited direct value for economists, because in the bulk of these analyses (a) the focus is on the evolution of strategies and not of preferences; (b) predictions are derived at the level of basic fitness components, such as reproduction and survival, and not at the level of the material payoffs generated by interactions; (c) transmission is assumed to be genetic instead of cultural, although the latter is highly relevant to the social and behavioral sciences, including economics.

Here we propose a framework more in line with those used by economists, and we focus on preference evolution. Specifically, we model the following thought experiment that takes place in a large population over an infinite sequence of discrete time periods. The population is structured into a large number of groups of equal size. Within each group, individuals engage in a strategic interaction in which all parties' strategy choices may affect the material payoffs to all participants. The material payoffs so realized in turn determine the expected fitness of each individual in the population, where an individual's fitness is the number of those individuals in the following time period who have acquired their trait from him or her. If transmission is genetic, an individual's fitness is the number of his surviving offspring and the individual himself if he survives. If transmission is cultural, an individual's fitness

is the number of individuals in the next time period who acquire their trait by copying this individual. Offspring may migrate to another group or stay in the natal group. Many different transmission scenarios are covered by this setting. For instance, generations may or may not be overlapping, groups may wage wars against each other, traits may be transmitted culturally from parent to child or by imitation of materially successful individuals. In all our scenarios, genetic and cultural, the population is initially homogenous; all individuals are *ex ante* identical. Suddenly, a different, mutant trait, spontaneously appears in exactly one individual. The original, resident, trait is uninvadable if there exists no mutant trait, such that the initial mutant produces enough descendants for its trait to be maintained in the population in the long run.

Within this setting we study preference evolution. The traits in question are utility functions that guide the carriers's choice of strategy in the group interaction. We evaluate a utility function's fitness consequences for its carriers in terms of the material payoffs that result when groups play Bayesian Nash equilibria under incomplete information. We ask if there exist utility functions that are uninvadable in the sense that any mutant utility function does worse, in terms of material payoffs, than the resident function in *all* equilibria. This analysis allows to establish a link between, on the one hand, the environment in which a population evolves—represented by the material game, the population structure, and the transmission scenario—and, on the other hand, preferences that motivate individuals in their choice of strategy. The following three main results emerge from our analysis.

First, under general transmission scenarios, allowing for both genetic and cultural evolution, when material payoffs have marginal effects on fitnesses, uninvadable preferences generically involve a mix of self-interest, a Kantian moral concern, which involves evaluating one's behavior in light of what own material payoff would be if others were to choose the same behavior, and, in addition, a concern for other group members' material payoffs. The weight on the Kantian motive is proportional to the coefficient of relatedness. The weight on other group members' material payoffs may be negative (spite) or positive (altruism), and it depends on the *coefficient of local competitiveness*, which measures the fitness benefit that an individual garners relative to his neighbors by diminishing or enhancing their material payoffs.

Second, we provide sufficient conditions for preferences of a particularly simple form, namely, a convex combination of own material payoff and the material payoff that would arise should all others choose the same behavior, to be uninvadable. Under these conditions, the weight given to the second, Kantian, component is given by the *coefficient of scaled relatedness*, a coefficient that combines the (standard) coefficient of relatedness with the coefficient of local competitiveness. This weight allows to determine whether, on balance, equilibrium behaviors are pro- or anti-social, in the sense that equilibrium material payoffs are higher or lower than under selfishness. We show that, depending on the specifics of the transmission process, this coefficient may be negative or positive, and we identify conditions under which it exceeds the coefficient of relatedness. To illustrate these results, we investigate a specific but canonical genetic scenario with overlapping generations, allowing for warfare between groups, and a cultural transmission scenario. The latter turns out to lead to a negative coefficient of scaled relatedness.

Third, our model allows to establish a clear distinction between preferences at the fitness level and preferences at the material payoff level. Specifically, we identify sufficient conditions for preferences of the same form as above, but whereby individuals trade off own *fitness* against the *fitness* that would arise should all others choose the same behavior, to be uninvadable, instead of trading off the material payoffs. By contrast to the preferences at the material payoff level, however, the weight attached to the second, Kantian, component concern at the fitness level is the usual coefficient of relatedness (which is always positive). Hence, uninvadability may be consistent with a Kantian concern both at the fitness level and at the material payoff level, but the former may be stronger or a weaker than the latter.

The most closely related literature is the following (see also Section 5 for a more detailed discussion of the biology literature). First, in Lehmann, Alger, and Weibull (2015) we asked under what conditions, if any, genetic trait evolution in a group structured population is compatible with individual maximizing behavior. Specifically, we established conditions, notably when traits have only marginal effects on fitness, under which evolving strategies can be interpreted as chosen by rational individuals endowed with certain preferences. These preferences correspond to the ones we here show to be uninvadable, as described above. The value added of the present paper is that we here (a) analyze preference evolution rather than strategy evolution, (b) derive novel results on the coefficients of local competitiveness and of scaled relatedness, and (c) consider more general transmission scenarios.

Second, Akçay and van Cleve (2012), investigated the evolutionary stability of behavioral response functions as heritable traits, where different behavioral response functions may represent different other-regarding preferences. In addition to focusing on complete rather than incomplete information, their model differs from ours in two broad respects. First, they focused only on the effects of traits on reproduction (fecundity) under genetic transmission. Second, they only considered necessary first-order condition for the evolutionary stability of such traits. Their first-order condition expresses how many offspring (units of reproduction, fecundity) an individual is willing to forgo, as expressed by the individual's preference, in order to increase the reproduction (number of offspring) of any other given group member. They show how this trade-off depends on the coefficient of scaled relatedness under complete information, thus generalizing previous first-order condition from preference evolution under uniform random matching (Heifetz, Shannon and Spiegel, 2007a-b) to spatially structured populations.

Finally, Alger and Weibull (2013, 2016) investigated the evolutionary stability of preferences under incomplete information, in an abstract model of assortative matching which did not explicitly account for the demographics and population dynamics.[2] They found

[2]By contrast to the present model, assortativity was there modeled as an abstract function that maps the distribution of traits in the population to probabilities governing the matching of interacting individuals. This formalization of assortativity was pioneered in economics by Bergstrom (1995, 2003), who focused on strategy evolution; see also Bowles and Gintis (1998), as well as Alger and Weibull (2010, 2012) for analyses of preference evolution under complete information. This formalization of assortativity, which implicitly assumes marginal effects of traits on fitness, goes back to Hamilton (1971) and Michod and Hamilton (1980) discuss how different formalizations of assortativity are equivalent to each other. It should further be noted that Rogers (1994) studied the evolution of time preference in an age-structured population; a setting that allows for kin selection but not kin competition.

that preferences expressing a certain combination of self-interest and a Kantian concern are evolutionarily stable, and that preferences that are behaviorally distinct from these are evolutionarily unstable. They also showed how the weight given to the Kantian concern depends on the assortativity in group formation. While assortativity in those models is treated as an abstract primitive, it here arises explicitly from the population structure; group size, rates of survival, migration, and conflicts together determine the probability that rare mutants get to interact with each other. The present model thus contributes to this strand of literature by explicitly modeling the population structure and how it gives rise to assortativity. The model makes it clear that relatedness—the probability that individuals who share a common ancestor get to interact—must go hand in hand with local competitiveness, a force which does not appear in Alger and Weibull (2013, 2016). We here also show how relatedness and local competitiveness can be formally traced back to population structure.

In sum, the existing literature displays how some features of the population structure and transmission process translate into features of uninvadable (or stable) "as if" preferences (Lehmann, Alger, and Weibull, 2015) or preferences (Akçay and van Cleve, 2012, Alger and Weibull, 2013, 2016). Compared with the earlier literature on preference evolution, our contribution is thus to explicitly analyze the effects of spatial and socioeconomic group structure and limited migration between groups upon behavior and preferences; and compared to the earlier biology literature, our contribution is to deliver predictions on preference evolution at the level of material payoffs under incomplete information and clarify the connections between the coefficients of relatedness, scaled relatedness, and local competitiveness.

The paper is organized as follows. Section 2 describes the model and provides a characterization of an uninvadable trait. Section 3 provides the main results. In Section 4 we illustrate these results in several specific evolutionary scenarios. Section 5 describes the related literatures, and Section 6 concludes. All the proofs are in the Appendix.

2 Model

2.1 Group structure, life-cycle, and fitness

In this section we present the building blocks of our analysis: the population structure, individuals' life-cycles, and demography. We consider a countably infinite population, divided into infinitely many identical *islands* (groups, locations, or villages). All islands have the same number $n > 1$ of adults at each point in time. From birth, each individual is endowed with a heritable type or *trait* $\theta \in \Theta$, which is fixed throughout his or her life. Time is divided into demographic time periods, each consisting of two phases. In the first phase, the adults in each island engage in a social or economic interaction with each other, to be called *the material game*, the same on all islands and at all times. This is a symmetric non-cooperative normal-form game in which each player has the same set of strategies, X, where X is a non-empty compact set in some normed vector space. The interaction results in material payoff $\pi(x_i, \boldsymbol{x}_{-i}) \in \mathbb{R}$ to each individual $i = 1, ..., n$ in the island, where $x_i \in X$ is the strategy used by the individual at hand and $\boldsymbol{x}_{-i} \in X^{n-1}$ is the vector of strategies used by the others on

i's island.[3] The material payoff-function $\pi : X^n \to \mathbb{R}$ is assumed to be *aggregative* in the sense that the payoff $\pi(x_i, \boldsymbol{x}_{-i})$ to any individual i is invariant under permutation of the components of the vector $\boldsymbol{x}_{-i} \in X^{n-1}$.[4] (This invariance property holds if, for example, all that matters for i's material payoff is the sum, product, maximum or minimum of all his or her neighbors' strategies.) In the main analysis an individual's trait is a utility function which guides his or her choice of strategy, and the set of traits Θ is the set F of aggregative functions $u : X^n \to \mathbb{R}$, but we will also examine the case where an individual's trait is a strategy to play in the material game, i.e., $\Theta = X$.

In the second phase of each demographic time period, *transmission events* occur. These events determine, for each individual, the probabilities that her trait will be transmitted to one or more adults, in her island or in another island, in the next demographic time period. An individual's *immediate descendants* consist of her surviving offspring as well as herself if she survives, where "offspring" may be biological or cultural, and traits are transmitted faithfully to offspring (i.e., "asexual" transmission).[5] The expected number of immediate descendants will be called the individual's *fitness*. This fitness is assumed to be determined by (a) the individual's own material payoff, (b) the material payoffs to the other individuals in her island, and (c) on the average material payoff in the population at large. This dependency is represented by a continuously differentiable *function* $w : \mathbb{R}^{n+1} \to \mathbb{R}$ that maps material payoff vectors $(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ to the fitness $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$, of any individual i whose material payoff is $\pi_i \in \mathbb{R}$ when her island neighbors earn payoffs $\boldsymbol{\pi}_{-i} \in \mathbb{R}^{n-1}$, and $\pi^* \in \mathbb{R}$ is the average material payoff in the population at large. We note that if all individuals in the population achieve the same material payoff, they also obtain the same individual fitness, and this must equal one, by virtue of the assumption that the total population size is constant over time.

Our objective is to draw general conclusions under as weak assumptions as possible. Concerning the nature of the transmission events, the only restriction we impose is that all offspring migrate between groups with a exogenously fixed positive probability. Concerning how fitness depends on material payoffs, the individual-fitness function w is assumed to be invariant under permutation of the material payoffs to the individual's island neighbors, and to be strictly increasing in own material payoff, strictly decreasing in the average material payoff in the population at large, and to be (weakly) less sensitive to other group members'

---

[3] Our assumption that there is but one strategy set and material payoff-function π in each demographic time period and on each island may appear unduly restrictive. However, our approach allows for the possibility that π represents many interactions that may arise simultaneously or randomly, and that may involve any subset of the inhabitants in an island, granted all individuals are equally likely to be involved, that each interaction is aggregative-symmetric, and that individuals have the perceptive and cognitive capacity to know which interaction is at hand.

[4] By *aggregative* we mean that for any $x_i \in X$ and $\boldsymbol{x}_{-i} \in X^{n-1}$, and any bijection $h : \{2, 3, ..., n\} \to \{2, 3, ..., n\}$: $\pi\left(x_i, x_{h(2)}, x_{h(3)}, ..., x_{h(n)}\right) = \pi(x_i, \boldsymbol{x}_{-i})$.

[5] This is a standard assumption in the literature on preference evolution in economics. For notable exceptions in the literature on strategy evolution, see Waldman (1994) and Bergstrom (1995).

material payoffs than to own material payoff.[6] More precisely, the following assumption will be maintained throughout:

[M] *(i)* $\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)/\partial \pi_i > 0$, *(ii)* $\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)/\partial \pi_j \leq \partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)/\partial \pi_i$ *for all* $j \neq i$, *(iii)* $\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)/\partial \pi^* < 0$.

Part of our analysis will be restricted to one-dimensional strategy sets and differentiable payoff functions, a case that we will refer to as the *strategy-differentiable setting*. Formally:[7]

[D] *(i)* $X = \mathbb{R}$, *and (ii)* $\pi : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable.*

We will sometimes use the composite function $\tilde{w} = w \circ \pi$, which gives the fitness of any individual i who plays strategy x_i when the others on his or her island play $\boldsymbol{x}_{-i}$ and strategy x is played by all individuals on all other islands:

$$\tilde{w}\left(x_i, \boldsymbol{x}_{-i}, x\right) = w\left(\pi\left(x_i, \boldsymbol{x}_{-i}\right), \left(\pi\left(x_j, \boldsymbol{x}_{-j}\right)\right)_{j \neq i}, \pi^*\left(x\right)\right), \tag{1}$$

We note that $\tilde{w} : X^{n+1} \to \mathbb{R}$ is continuously differentiable in the strategy-differentiable setting.

Part of our analysis will assume that material payoffs affect fitness by a scaling factor $\delta \geq 0$, the *intensity of selection*. Specifically, for each $x \in X$ and $\boldsymbol{y} \in X^{n-1}$ let

$$\bar{\pi}\left(x, \boldsymbol{y}\right) = \delta \cdot \pi\left(x, \boldsymbol{y}\right). \tag{2}$$

[S] For any payoff vector $\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right) \in \mathbb{R}^{n+1}$, the individual fitness of individual i is $w\left(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*\right)$.

In biology *weak selection* refers to situations in which the fitness effects of heritable traits are small and can be represented in terms of their first-order effects (see, e.g., Nagylaki, 1992, Rousset, 2004). We here implement weak selection by assuming **[S]** and considering the limit as δ tends towards 0.[8]

2.2 Uninvadable traits

What matters for the long-run evolutionary success of any heritable trait is whether individuals with that trait on average have more descendants in future generations than individuals with other traits. We ask whether a monomorphic population, i.e., a population in which all individuals have the same trait, may be invaded by another trait that initially appears in a single individual. Formally, consider any material game $\langle n, X, \pi \rangle$, any transmission scenario

[6] These properties are fully in line with the literature on preference evolution, in which the material payoff is taken to represent fitness.

[7] The uni-dimensionality assumption is inessential. All analysis can be carried out in terms of gradients, but this is avoided in order not to clog the notation.

[8] This formalization of weak selection corresponds to what Wild and Traulsen (2007) call w-weak selection.

whereby traits are transmitted from one demographic time period to the next according to some fitness function $w : \mathbb{R}^{n+1} \to \mathbb{R}$ satisfying Assumption **[M]**, any resident trait $\theta \in \Theta$, and any trait $\tau \in \Theta$ that initially appears in a single individual. The trait θ is *uninvadable* if the random number of demographic time periods during which any given mutant trait $\tau \neq \theta$ remains in the population is finite with probability one. In other words, a trait is called *uninvadable* if any other (mutant) trait that initially appears in a single individual, is bound to disappear from the population.

To get a hold on what traits, if any, are uninvadable in this sense, we use the notion of an individual's *lineage*, defined as the collection of this individual's all descendants: her immediate descendants, defined as her offspring and also herself if she survives, the immediate descendants of her immediate descendants, etc. *ad infinitum*. Suppose that initially everybody in the population has trait θ, and that suddenly one individual on an island, a mutant, switches to another trait, τ. Individuals from the initial's mutants lineage may colonize (via migration) new islands that are occupied by the resident trait, and may do so as singletons or possibly along with other mutants. Irrespective of whether an island of residents is colonized by a single or several mutants, the random time of "first extinction" of the *local lineage* of mutants—defined as the collection of all descendants of mutants who initially jointly colonized that island—is finite with probability one, since the migration rate by assumption is positive and constant.[9]

It turns out that *local lineage fitness* of the mutant trait τ can be used to characterize uninvadability. This is defined as the average fitness of a mutant's local lineage members, the average being taken over all demographic time periods until the first period in which the local lineage is extinct and over all possible initial conditions of that local lineage (single or multiple simultaneous mutants), under the following assumptions: (1) in any island with k mutants, all the mutants achieve the same material payoff, and all the residents achieve the same material payoff; (2) these material payoffs are the same in all islands with k mutants. In general traits may be associated with multiple material payoffs and matching probabilities. While we show how this is done below, for now we proceed under the hypothesis that all pairs of traits uniquely determine material payoffs and matching probabilities for all individuals.

Formally, the local lineage fitness of a mutant trait τ in a population where the resident trait is θ is defined as

$$W(\tau, \theta) = \sum_{k=0}^{n-1} p_k(\tau, \theta) \cdot w\left(\pi(\tau|k), \left\langle \pi(\tau|k), \pi(\theta|k) \right\rangle, \pi^*(\theta)\right), \tag{3}$$

where, for each $k = 0, ..., n-1$, $p_k(\tau, \theta)$ is the probability for a uniformly drawn descendant of the initial mutant that $k = 0, 1, ..., n-1$ of his neighbors are also descendants of the initial mutant. Such *identity-by-descent* depends on demographic features (migration rate, group size, fecundity, survival, etc.) In the fitness term, $\pi(\tau|k)$ denotes the material payoff to the mutant at hand when there are k other mutants among his neighbors, $\langle \pi(\tau|k), \pi(\theta|k) \rangle \in$

[9]Note that even if locally extinct, lineage members may still live on other islands, and some of them may even move to the mutant's native island. However, this last event has probability zero in the limit when the number of islands is infinite.

$\mathbb{R}^{n-1}$ denotes the vector of material payoffs to the other $n-1$ members of i's island (among whom k have the mutant trait τ and the others have the resident trait θ), and $\pi^*(\theta)$ is the material payoff earned by individuals in the rest of the population, where all have the resident trait θ.[10]

With local lineage fitness so defined, uninvadability can be succinctly characterized as follows (Lehmann et al. 2016):[11] a trait $\theta \in \Theta$ is *uninvadable* if and only if for every trait $\tau \in \Theta$,

$$W(\tau, \theta) \leq W(\theta, \theta) \quad \forall \tau \in \Theta. \tag{4}$$

This characterization compares the local lineage fitness of a single initial τ-mutant, $W(\tau, \theta)$, to the local lineage fitness, $W(\theta, \theta)$, of any resident in this population. Equivalently, the criterion can be written

$$\theta \in \arg\max_{\tau \in \Theta} W(\tau, \theta). \tag{5}$$

An uninvadable trait thus preempts entry into the population by earning the maximal expected fitness that members of the local lineage of any original mutant can obtain in a population where the resident trait is θ. In other words, a trait is uninvadable if and only if it is a best reply to itself in terms of local lineage fitness.

The usefulness of this observation is limited by the fact that the weights $p_k(\tau, \theta)$ used in the definition of W may depend in a non-trivial way on the traits θ and τ. However, as the analysis below shows, this difficulty may be circumvented by focusing on (a) the differentiable setting, and (b) when material payoffs have infinitesimal effects on fitness (weak selection). In both cases we will rely on the fact that while it is in general impossible to calculate analytically the distribution of matching probabilities, $\boldsymbol{p}(\tau, \theta) = (p_0(\tau, \theta), ..., p_{n-1}(\tau, \theta))$, it is possible to calculate the *coefficient of pairwise relatedness*

$$r(\tau, \theta) = \sum_{k=0}^{n-1} \frac{k}{n-1} \cdot p_k(\tau, \theta). \tag{6}$$

This is a summary statistic of identity-by-descent, which for descendants of the initial mutant measures the average share of their neighbors who are also descendants of the initial mutant.

2.2.1 Uninvadable strategies

Prior to applying our model to the case when the traits are utility functions, consider briefly the case when the heritable traits are (pure or mixed) strategies in the material game, i.e., $\Theta = X$. Then all payoffs and matching probabilities are uniquely determined, and a strategy $x \in X$ is uninvadable if and only if

$$W(y, x) \leq 1 \quad \forall y \in X. \tag{7}$$

[10] Likewise, one can define $W(\theta, \theta)$ as the local lineage fitness of any given individual with trait θ when the resident trait is θ.

[11] In Lehmann, Alger, and Weibull (2015) we proved this result for scenarios where new islands can be colonized only by singleton mutants. Lehmann et al. (2016, eqs. (14)-(16)) extended that result to allow for scenarios in which multiple offspring from the same group can reproduce in the same non-natal island.

Let $\hat{\mathbf{x}}^{(n-1)}$ denote the $(n-1)$-dimensional vector whose components all equal $\hat{x}$, and $r(\hat{x},\hat{x}) \in [0,1]$ the coefficient of pairwise relatedness in a monomorphic population where everybody plays $\hat{x}$. This relatedness is evaluated under what biologists refer to as *neutral drift* or the *neutral process*, i.e., when every individual in the population face the same prospects of reproduction and survival (the same distribution for the random number of descendants, which obtains if the population is monomorphic), see, e.g., Crow and Kimura (1970) or Rousset (2004). If Assumption **[D]** holds and $\hat{x} \in X$ is uninvadable, then

$$\left.\frac{\partial W(y,\hat{x})}{\partial y}\right|_{y=\hat{x}} = \tilde{w}_1\big(\hat{x},\hat{\mathbf{x}}^{(n-1)},\hat{x}\big) + r(\hat{x},\hat{x})\sum_{j=2}^{n}\tilde{w}_j\big(\hat{x},\hat{\mathbf{x}}^{(n-1)},\hat{x}\big) \quad = \quad 0, \tag{8}$$

where an index i on $\tilde{w}$ denotes the partial derivative with respect to the i-th argument. For any setting where the coefficient of relatedness $r(\hat{x},\hat{x})$ is strictly positive, it can be interpreted as a marginal substitution rate, since it measures the number of units of own fitness that any given individual is willing to forgo to increase the fitness of each neighbor by one unit. Indeed, the first term and the second (the sum) in (8) must have opposite signs; an uninvadable strategy inflicts a marginal fitness cost on the individual (the first term), which equals the sum of the marginal fitness benefits conferred on others (the sum) weighted by the coefficient of pairwise relatedness. Equation (8) is thus nothing but the necessary first-order condition of Hamilton's rule (Hamilton, 1964) for an (interior) strategy to be uninvadable (see equation (3) in Taylor and Frank 1996, or equation (7.5) in Rousset 2004).[12] Such first-order conditions are standard in the biology literature, but for the sake of completeness we provide a proof in the appendix. Note that, besides the recipient-centered interpretation provided above, expression in (8) can also be given an actor-centered interpretation. In this second interpretation, the sum is viewed as the individual's own fitness, should all the neighbors play the same strategy as the individual himself. Both interpretations are found in the literature (see, e.g., Figure 7.1 in Rousset, 2004).

2.2.2 Uninvadable utility functions

Henceforth, we take the heritable traits to be utility functions, i.e., $\Theta = F$. Individuals are assumed to be fully rational, know the set of available strategies X, and strive to maximize the mathematical expectation of their personal utility function. To define uninvadability of utility functions, suppose that initially everybody in the population is endowed with the same utility function $f \in F$, and that suddenly one individual, a *mutant*, is given another utility function $g \in F$. We focus on the case where each individual's utility function, or type, is his or her private information. Then an individual's behavior cannot be conditioned on the types of those with whom (s)he has been matched. However, individual behavior may be adapted to the population state at hand—that is, the current type distribution in the population. We

[12] First-order conditions like equation (8) apply more generally to traits if lineage fitness and individual fitness are differentiable in trait values. The aforementioned evolutionary dynamics literature focuses on the evolution of phenotypes—the composite of an organism's characteristics—thus subsuming virtually any heritable trait.

evaluate utility functions in terms of their expected material payoff consequences for their carriers in all (Bayesian) Nash equilibria under incomplete information.[13]

What matters for an individual's choice of strategy is his or her probabilistic belief about her matching profile, that is, her subjective probability distribution of the numbers of residents and mutants in her own island. Let $\tilde{\boldsymbol{q}} = (\tilde{q}_0, \tilde{q}_1, ..., \tilde{q}_{n-1})$ be a resident's probabilistic belief, where $\tilde{q}_k$ is the (subjective) probability that there are (precisely) k mutants in any given resident's island. Likewise, let $\tilde{\boldsymbol{p}} = (\tilde{p}_0, \tilde{p}_1, ..., \tilde{p}_{n-1})$ be a mutant's probabilistic belief, where $\tilde{p}_k$ is the (subjective) probability that there are (precisely) k *other* mutants in her island. Let $\tilde{\mathbf{y}}^{(j)}$ denote the j-dimensional vector whose components all equal $\tilde{y}$, and $\tilde{\mathbf{x}}^{(j)}$ the j-dimensional vector whose components all equal $\tilde{x}$ (for $j = 1, ..., n-1$). Given a pair of utility functions $(f, g) \in F^2$, a (type-homogenous Bayesian) *Nash equilibrium* under a matching profile $\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}} \rangle$ is a pair of strategies, $(\tilde{x}, \tilde{y})$, one for each preference type, such that each strategy is a best reply for any player of that type, given the (subjective) matching probabilities $\langle \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{p}} \rangle$:[14]

$$\begin{cases} \tilde{x} \in \arg\max_{x \in X} \ \sum_{k=0}^{n-1} \tilde{q}_k \cdot f\left(x, \tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{x}}^{(n-k-1)}\right) \\ \tilde{y} \in \arg\max_{y \in X} \ \sum_{k=0}^{n-1} \tilde{p}_k \cdot g\left(y, \tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{x}}^{(n-k-1)}\right). \end{cases} \tag{9}$$

In particular, if $\tilde{q}_k = 0$ for all $k > 0$, then the first equation in (9) defines symmetric Nash equilibrium play among n individuals of the resident type f (and is hence independent of the mutant type g), and can thus be written more concisely as

$$\tilde{x} \in \arg\max_{x \in X} \ f\left(x, \tilde{\mathbf{x}}^{(n-1)}\right). \tag{10}$$

A utility function will be called *uninvadable* if any other (mutant) utility function, which initially appears in a single individual, is bound to disappear from the population. Our definition uses Nash equilibrium as a testing ground. It requires that if the residents were playing some Nash equilibrium among themselves (thus having mutually adjusted their individual behaviors to best replies, given their preferences), and if residents' and mutants' probabilistic beliefs were correct, then there would not exist any mutant who could enter the population and earn a higher expected material payoff than the residents by acting optimally according to his or her own preferences.

Recall that the population is infinite while any mutant's lineage population, before its first local extinction, is finite. Hence, to determine the viability of mutant preferences it is sufficient to consider (type-homogeneous) Nash equilibria $(\tilde{x}, \tilde{y})$ under matching profiles $\langle \boldsymbol{q}^0, \boldsymbol{p} \rangle$, where $\boldsymbol{q}^0$ is the unit vector $(1, 0, 0, ..., 0)$ and $\boldsymbol{p}$ is a possible vector of mutants' true matching probabilities, p_0, p_1, etc. in the given transmission scenario, when the resident

[13] Alternatives to Nash equilibrium for evolutionary analyses have been considered, see e.g. Curry and Roemer (2012) and Newton (2017).

[14] By Berge's maximum theorem and Kakutani's fixed-point theorem, there exists such a strategy pair if *(i)* the strategy set X is non-empty, finite-dimensional, compact and convex, *(ii)* the function u is continuous, *(iii)* the first maximand in (9) is quasi-concave in $x \in X$ and the second maximand in (9) is quasi-concave in $y \in X$. For generalizations to infinite-dimensional spaces, see e.g. Aliprantis and Border (2006).

utility function is $f \in F$ and the mutant utility function is $g \in F$. In general, any such probability distribution $\boldsymbol{p}$ may depend on both the resident's strategy $\tilde{x}$ and the mutant's strategy $\tilde{y}$. However, if the $\boldsymbol{p}$ distribution depends on the mutant's strategy, then the distribution over group profiles faced by a mutant is determined by its own strategy choice (see (9)). This precludes a meaningful representation of individuals as choosing their strategies, since then the number of island neighbors using the same strategy as the mutant will depend on what strategy the mutant chooses.[15] However, as will be shown below, under either assumption **[D]** or **[S]** (in addition to **[M]**), the true $\boldsymbol{p}$ distribution depends at most on the residents' strategy $\tilde{x}$**.** We will then write $\boldsymbol{p}(\tilde{x})$ for a mutant's unique matching probability vector when residents use strategy $\tilde{x}$.

In force of the fact that a strategy $\hat{x} \in X$ is uninvadable if and only if $W(y, \hat{x}) \leq 1$ for all strategies $y \in X$, we formally define a utility function $f \in F$ to be *uninvadable* if $W(\tilde{y}, \tilde{x}) \leq 1$ for all utility functions $g \in F$ and all $\tilde{x} \in X^2$ satisfying (10) and

$$\tilde{y} \in \arg\max_{y \in X} \sum_{k=0}^{n-1} p_k(\tilde{x}) \cdot g\left(y, \tilde{\mathbf{y}}^{(k)}, \tilde{\mathbf{x}}^{(n-k-1)}\right). \tag{11}$$

Any utility function that is not uninvadable will be called *invadable*.

3 Results

We begin by considering necessary first-order conditions under assumptions **[M]** and **[D]**, then illustrate the results in a base-line evolutionary scenario, and finally consider weak selection under assumptions **[M]** and **[S]**.

3.1 The strategy-differentiable setting

For any given reference strategy $x \in X$, define the associated utility function $u_x : X^n \to \mathbb{R}$ by

$$u_x(x_i, \boldsymbol{x}_{-i}) = [1 - r(x,x)] \cdot \tilde{w}(x_i, \boldsymbol{x}_{-i}, x) \; + \; r(x,x) \cdot \tilde{w}\left(x_i, \mathbf{x}_{-i}^{(n-1)}, x\right), \tag{12}$$

where $\mathbf{x}_{-i}^{(n-1)} \in X^{n-1}$ is the strategy vector whose all components all equal x_i (that is, when all neighbors use i's strategy).[16] An individual equipped with the goal function $u_{\hat{x}}$ evaluates her strategy, x_i, both in terms of how it affects her own fitness, given the neighbors' strategies and the strategy played in the population at large, reflected in the first term, and how it would affect her neighbors' fitnesses should they also play her strategy x_i, reflected in the

[15] This ultimately ows to the fact that under the (biological or cultural) transmission process of traits, the $\tilde{p}$ distribution is determined in the ancestral state of a group, and hence is not a choice variable of a current group member.

[16] It is easily verified that $u_{\hat{x}}(x_i, \boldsymbol{x}_{-i})$ is invariant under permuation of the strategies in $\boldsymbol{x}_{-i}$. Hence, $u_{\hat{x}} \in F$.

second term. The latter has a flavor of Kant's (1785) categorical imperative. The weights attached to these two fitness effects are $1 - r(x,x)$ and $r(x,x)$, respectively.

The following proposition identifies conditions under which the utility function u_x is uninvadable.

Proposition 1 *Suppose that **[D]** holds and that there exists a unique strategy $\hat{x}$ that satisfies (8). If $\hat{x}$ is uninvadable and also is the unique Nash equilibrium strategy in the game in which all players have utility function $u_{\hat{x}}$, then $u_{\hat{x}}$ is uninvadable.*

The proposition underlines the challenge of characterizing uninvadable preferences without making additional assumptions, given our stringent definition of an uninvadable utility function. Indeed, this definition in effect requires all the Nash equilibrium strategies among residents, the set of strategies satisfying (10), to be uninvadable under strategy evolution. But no utility function whereby its carrier attaches a uniquely defined weight to his neighbor's fitnesses could encompass the fact that in case of multiple uninvadable strategies the coefficient of relatedness typically differs between them.

While Proposition 1 thus shows that the utility function u_x is uninvadable in a restricted set of cases, it establishes that, at least in such cases, preferences whereby individuals attach a positive weight $r(\hat{x},\hat{x})$ to the Kantian concern about fitness, where $\hat{x}$ is a candidate uninvadable strategy, may be viable. Does this also mean that Kantian concerns at the level of material payoffs may prevail? The next two propositions provide answers to this question.

For any given reference strategy $x \in X$, define the goal function $v_x \in F$ by

$$v_x(x_i, \boldsymbol{x}_{-i}) = [1 - \kappa(x)] \cdot \pi(x_i, \boldsymbol{x}_{-i}) \; + \; \kappa(x) \cdot \pi\left(x_i, \mathbf{x}_{-i}^{(n-1)}\right) \tag{13}$$

where

$$\kappa(x) = \frac{r(x,x) - \lambda(x)\left[\frac{1}{n-1} + \left(\frac{n-2}{n-1}\right) r(x,x)\right]}{1 - \lambda(x)\, r(x,x)}, \tag{14}$$

is the *coefficient of scaled relatedness,* and $\lambda(x)$ is the *coefficient of local competitiveness,* a coefficient that measures the marginal effect of neighbors' material payoffs on own fitness, relative to the marginal effect of own material payoff on own fitness, in a population in which all individuals play x**:**

$$\lambda(x) = -\left(\sum_{j \neq i} \frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\right) / \left(\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\right). \tag{15}$$

A positive coefficient $\lambda(x)$ can be interpreted as there being competition for local resources: an increase in the material payoffs to neighbors then reduces an individual's fitness. A negative coefficient $\lambda(x)$ means that there is a positive externality at the level of material payoffs between neighbors: an increase in the material payoffs to neighbors then increases an individual's fitness. Like u_x, the function v_x can be interpreted as involving a Kantian concern, but at the level of material payoffs and not at the level of fitnesses. Indeed, an individual equipped with this utility function evaluates her strategy, x_i, by pondering both

how it affects her own material payoff, given the neighbors' strategies, reflected in the first term, and how it would affect her neighbors' fitnesses should they also play her strategy x_i, reflected in the second term. The weight attached to the Kantian concern, however, differs from that attached to the Kantian concern in u_x, whenever $r(x,x) \neq \kappa(x)$.

The following result obtains:

Lemma 1 *If **[D]** holds and $\hat{x} \in X$ is uninvadable, then*

$$[1-\kappa(\hat{x})] \cdot \pi_1\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right) \; + \; \kappa(\hat{x}) \cdot \sum_{j=1}^{n} \pi_j\left(\hat{x}, \hat{\mathbf{x}}^{(n-1)}\right) \quad = \quad 0. \tag{16}$$

In words, any uninvadable strategy $\hat{x}$ must meet the first-order condition for it to be a candidate symmetric Nash equilibrium strategy of a game in which all players have utility function $v_{\hat{x}}$. The following proposition identifies conditions under which the utility function $v_{\hat{x}}$ is uninvadable:

Proposition 2 *Suppose that **[D]** holds and that there exists a unique strategy $\hat{x} \in X$ that satisfies (16). If $\hat{x}$ is uninvadable and also the unique Nash equilibrium strategy in the game in which all players have utility function $v_{\hat{x}}$, then $v_{\hat{x}}$ is uninvadable.*

Taken together, Propositions 1 and 2 imply that in a population in which all individuals play the same unique uninvadable strategy $\hat{x}$, these individuals may be perceived as having a Kantian concern at the fitness level as well as at the material payoff level. Moreover, the strength of the Kantian concern at the fitness level, measured by $r(\hat{x},\hat{x})$, may differ from the strength of the Kantian concern at the material payoff level, measured by $\kappa(\hat{x})$. The next proposition shows that $\kappa(\hat{x})$ can be smaller of larger than $r(\hat{x},\hat{x})$, and that it is sufficient to know whether neighbors exert positive or negative externalities on each other's material payoffs to know whether $\kappa(\hat{x})$ or $r(\hat{x},\hat{x})$ is largest. Finally, the proposition shows that $\kappa(\hat{x})$ may even be negative. Specifically:

Proposition 3 *The weight $\kappa(\hat{x})$ attached to the neighbors' material payoffs in the function $v_{\hat{x}}$ lies in the interval $[-1,1]$. Furthermore, $\kappa(\hat{x}) > r(\hat{x},\hat{x})$ if and only if $\lambda(\hat{x}) < 0$. A necessary and sufficient condition for $\lambda(\hat{x}) < 0$ is that $\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)/\partial \pi_j$, evaluated at $\hat{x}$, is strictly positive.*

The coefficient $\kappa(\hat{x})$ can be interpreted as a marginal substitution rate, since it gives the number of units of own material payoff that any given individual is willing to forgo to increase the material payoff of each neighbor by one unit. If there was no local competition, i.e., if $\lambda(\hat{x}) = 0$, then $\kappa(\hat{x})$ would be equal to $r(\hat{x},\hat{x})$, meaning that the individual would be willing to forgo material payoff at a rate given by relatedness, $r(\hat{x},\hat{x})$. This is intuitive, since (pairwise) relatedness can be thought of as a measure of the recipient's ability, relative to that of the donor, to transmit a given trait (here a strategy) to the next generation (recall

Hamilton's rule above). When $\lambda(\hat{x}) > 0$, however, a transfer of own payoff to neighbors has consequences which also need to be taken into account in the substitution rate.

To see this, let us first consider the case when there is but one neighbor, that is $n = 2$. A payoff transfer to this neighbor increases competition from the neighbor at rate $\lambda(\hat{x})$ (since $\lambda(\hat{x})$ measures the relative increase in competition in the neighborhood of an individual when its payoff is varied, see (15)). In that case, the fitness benefit to the donor from giving the transfer to the neighbor is reduced by $\lambda(\hat{x})$, so that the numerator in (14) becomes $r(\hat{x},\hat{x}) - \lambda(\hat{x})$. Moreover, a transfer of resources to the neighbor alleviates the competition that the neighbor experiences, and the neighbor is related to the donor according to coefficient $r(\hat{x},\hat{x})$. Hence, the cost of the transfer is reduced by $\lambda(\hat{x})\, r(\hat{x},\hat{x})$, which explains the denominator in (14).

Second, when there are multiple neighbors, $n > 2$, a transfer given to one neighbor enhances the competition by $\lambda(\hat{x})/(n-1)$, but also for the $(n-2)$ other neighbors, each of which is related to the donor according to coefficient $r(\hat{x},\hat{x})$. Therefore, the fitness benefit of the transfer to the donor is reduced by $\lambda(\hat{x})/(n-1)$ times the term in square brackets in the numerator; which explains the numerator of $\kappa(\hat{x})$. In the denominator, the cost of the transfer is still reduced by $\lambda(\hat{x})\, r(\hat{x},\hat{x})$, which is the expected alleviation of competition that the transfer induces for the individual's neighbors (recall that $\lambda(\hat{x})$ accounts for all neighbors through the term $(n-1)$).

3.2 A base-line evolutionary scenario

We illustrate these first results in a canonical evolutionary scenario. Suppose that traits are genetically determined. The success of a trait then depends on the number of surviving biological offspring that its carriers are able to produce. We consider the simplest possible scenario that captures this idea. For simplicity, assume that reproduction is asexual, i.e., each offspring inherits the trait of its single parent. There are three stages within each demographic time period, after play of the material game. In Stage 1, each adult produces a Poisson-distributed number of offspring (clones), and then dies. In Stage 2, each offspring either stays in his or her natal island and aspires to replace a deceased adult there, or migrates to another, randomly chosen island, where she aspires to replace a deceased adult. Individual offsprings' migration decisions are statistically independent, and offspring who migrate disperse uniformly to the other islands, with statistical independence between their destinations. In Stage 3, in each island the deceased adults are replaced by (uniformly) randomly drawn aspiring offspring, native and immigrant. The fortunate ones settle and become adults while the unfortunate ones die.[17] In this baseline biological scenario, with non-overlapping generations, the fitness of individual i writes:

$$w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*) = \frac{(1-m)\, n f(\pi_i)}{(1-m)\sum_{j=1}^{n} f(\pi_j) + nm f(\pi^*)} + \frac{m f(\pi_i)}{f(\pi^*)}, \tag{17}$$

[17] We assume that the total number of offspring is large enough for the probability of there being fewer aspiring offspring than there are deceased adults at an island to be negligible.

where $0 < m \leq 1$ is the probability for each newborn to migrate to another island, and $f(\pi_i) > 0$ is i's expected number of offspring, or fecundity. The first term is the expected number of i's offspring who manage to secure a "breeding spot" on the natal island. This term is the product of three factors: (a) the probability for not migrating, $(1-m)$; (b) the number of available spots on the island, n; and, for each available spot, (c) the competition for the spot, among native and migrating offspring from other islands, where $f(\pi^*)$ is the fecundity in the population at large. The second term is the expected number of i's offspring who migrate and manage to secure a breeding spot on another island: each offspring who migrates to another island competes against $nf(\pi^*)$ other individuals for the n available spots.[18]

In the appendix we show that in this scenario, for any given reference strategy x, the coefficient of pairwise relatedness is given by

$$r(x,x) = \frac{(1-m)^2}{n-(n-1)(1-m)^2}, \tag{18}$$

while the coefficient of local competitiveness is given by

$$\lambda(x) = \frac{(n-1)(1-m)^2}{n-(1-m)^2}. \tag{19}$$

Both coefficients turn out to be independent of the reference strategy x. Hence, the utility functions u_x and v_x, defined in equations (12) and (13), are independent of what strategy x is used in the population at large and can, in this evolutionary scenario, be explicitly parametrized in terms of the migration rate m and group size n. Both coefficients, pairwise relatedness and local competitiveness, are strictly positive for all n and all $m \in (0,1)$. By contrast, if $m = 1$, any two migrants from the same island have zero probability of arriving at the same island (since there are infinitely many islands and offspring migrate individually and statistically independently), and hence the probability of interacting with an individual from the same lineage is nil, $r(x,x) = 0$, and, moreover, there is no fitness benefit from out-competing neighbors materially, $\lambda(x) = 0$. Note also that $\kappa(x) = 0$ when $m = 1$. Interestingly, however, it turns out in this example that, by substituting (18) and (19) into (14), we have

$$\kappa(x) = 0 \tag{20}$$

for all n and m. In other words, in this evolutionary scenario, for any migration rate $m \in (0,1)$ any uninvadable strategy must be as if individuals have a Kantian concern at the level of fitnesses ($r(x,x) > 0$), but are purely selfish at the level of material payoffs ($\kappa(x) = 0$). In biology, this result is known as Taylor's (1992a) cancellation result.[19] It is a pivotal result of strategy evolution in spatially structured populations, noticed initially

[18] Since the total number of islands is infinite, the probability is zero for the event that more than one of i's offspring happen to migrate to the same island. For a more detailed derivation of an equation like (17) from the random variables underlying reproduction, see Lehmann and Balloux (2007).

[19] To see why this is the case, consider a focal individual expressing an action producing $B/(n-1)$ units of fecundity (by way of increasing material payoff) for each of her group neighbors. This implies that the

in agent-based simulations by Wilson, Pollock, and Dugatkin (1992), proven formally by Taylor (1992a) for the island model, and then shown to hold for arbitrary migration patterns between groups (e.g., Taylor, 1992b, Rousset, 2004, and Ohtsuki, 2012). This result is a useful yardstick for understanding how changes in the transmission scenario can tip the balance either in the direction of pro-sociality or anti-sociality. Such tipping effects will be investigated in Section 4.

3.3 Weak selection

The key implication of weak selection is that in the limit as the scaling factor δ tends to zero (see equation (2) and Assumption [S]), the matching probabilities $p_k(y,x)$, while still depending on the transmission events, do not depend on the strategies x and y used in the population, and are evaluated under the neutral process induced by the corresponding transmission events. The probability for a randomly drawn descendant of an ancestor, be it a resident or mutant, to coexist in its island with k other descendants of the same ancestor is then solely determined by the vital rates in a population in which everybody uses the same strategy x, no matter which. Let p_k^0 denote the matching probabilities in the neutral process, and write $\boldsymbol{p}^0 = \left(p_0^0, p_1^0, ..., p_{n-1}^0\right)$.

Letting $B_{\text{NE}}^0(f,g) \subseteq X^2$ denote the set of (type-homogenous Bayesian) Nash equilibria under matching profile $\langle \boldsymbol{q}^0, \boldsymbol{p}^0 \rangle$, we call a utility function $f \in F$ *uninvadable* if $W(\tilde{y},\tilde{x}) \leq W(\tilde{x},\tilde{x})$ for all utility functions $g \in F$ and all $(\tilde{x},\tilde{y}) \in B_{\text{NE}}^0(f,g)$.

As a stepping stone towards our result on preference evolution, we begin by deriving a lemma under strategy evolution. For this purpose, we define the *lineage payoff-advantage* of a mutant strategy $y \in X$ in a population of residents using strategy $x \in X$ as

$$\Pi(y,x) = \sum_{k=0}^{n-1} p_k^0 \cdot \tilde{\pi}^{(k)}(y,x), \tag{21}$$

where $\tilde{\pi}^{(k)}(y,x)$ is the mutant's *payoff advantage* when there are k other mutants in her or his island, defined by

$$\tilde{\pi}^{(k)}(y,x) = \pi(y|k) - \lambda_0 \cdot \left[\frac{k}{n-1}\pi(y|k) + \frac{n-1-k}{n-1}\pi(x|k)\right]. \tag{22}$$

The first term in this expression is the payoff of a descendant of the initial mutant who finds herself in an island with k other such descendants. The term in square brackets is the

gains in the number of fitness units to the trait underlying the action is $B \cdot r(x,x)$, since each neighbour carries the type of the focal individual with probability $r(x,x)$. Producing additional offspring, however, increases local competition if these offspring do not migrate to other groups. This implies that the loss in the number of fitness units to the type underlying the action is $B \cdot [1/n + r(x,x)(n-1)/n](1-m)^2$, since each non-dispersing offspring [probability $(1-m)$] of the focal individual and of her related neigbours enters in competition with probability $(1-m)$ with the new created offspring, and where the focal individual's offspring is a fraction of $1/n$ of the group's offspring experiencing the increase in competition. Owing to the fact that $r(x,x) = (1-m)^2[1/n + r(x,x)(n-1)/n]$ (see equation (94) in the appendix), fitness gains and losses exactly cancel out.

average material payoff earned by the other members in the island. Under weak selection, the coefficient of local competitiveness is obtained by evaluating the marginal effect of own material payoff on all neighbors' fitnesses, relative to its effect on own fitness, in the limit as the intensity of selection goes to zero:

$$\lambda_0 = \left.\lambda(x)\right|_{\delta=0}. \tag{23}$$

Likewise, and for further use below:

$$r_0 = \left.r(x,x)\right|_{\delta=0} \tag{24}$$

and

$$\kappa_0 = \frac{r_0 - \lambda_0\left[1/(n-1) + (n-2)\, r_0/(n-1)\right]}{1-\lambda_0 r_0}. \tag{25}$$

The lineage payoff-advantage is the average payoff advantage of the initial mutant's descendants. It reflects two facts implied by population structure. First, it measures the fitness consequences of an individual's ability to out-compete his neighbors, with λ_0 capturing the importance of doing so. Second, it reflects the fact that descendants of the initial mutant who stay in the natal island interact with each other, as captured by the matching probabilities p_k^0.[20]

A strategy is uninvadable under weak selection if and only if there is no strategy that would have a lineage payoff-advantage if appearing as a rare mutant. Formally:[21]

Lemma 2 *A strategy $\hat{x} \in X$ is uninvadable under weak selection if and only if*

$$\Pi(y,\hat{x}) \leq \Pi(\hat{x},\hat{x}) \qquad \forall y \in X. \tag{26}$$

Moreover, $1-n \leq \lambda_0 \leq 1$.

Condition (26) says that a resident strategy $\hat{x}$ is uninvadable if it also maximizes the lineage payoff-advantage that a mutant could hope to get in a population where $\hat{x}$ is the resident strategy. In doing so, such a strategy preempts entry by any mutant strategy. This lemma in turn implies:

Corollary 1 *A utility function $f \in F$ is uninvadable under weak selection if and only if $\Pi(\tilde{y},\tilde{x}) \leq \Pi(\tilde{x},\tilde{x})$ for all utility functions $g \in F$ and all $(\tilde{x},\tilde{y}) \in B^0_{\mathrm{NE}}(f,g)$.*

---

[20] In particular, $\tilde{\pi}^{(k)}(x,x) = (1-\lambda_0)\,\pi(x|k)$, where the factor in parenthesis accounts for competition among lineage members.

[21] This result agrees with a result in Lehmann, Alger, and Weibull (2015); see equation (14) therein. However, in that paper we did not allow for reproductive processes where offspring from the same group can reproduce in the same non-natal group, while we do allow for that here. Furthermore, we did not consider cultural transmission — which we do here — and here we identify the functional form for λ_0, see (23), and we determine a lower bound for λ_0. Finally, the proof of the present, more general result is more detailed and self-contained, than the previous proof.

For any given type $f \in F$, let $X(f)$ be the set of symmetric Nash equilibrium strategies, i.e., which satisfy equation (10). Thanks to Lemma 2 and Corollary 1, we are in a position to establish a general result for uninvadable utility functions under weak selection. In order to express this result, we need some notation and a definition. For any probability vector $\boldsymbol{p} = (p_0, p_1, ..., p_{n-1})$, any player i, and any strategy profile $\boldsymbol{x} \in X^n$, let $\tilde{\boldsymbol{z}}$ be a *random strategy-profile* such that with probability p_k (for each $k = 0, 1, .., n-1$) exactly k of the $n-1$ components in $\boldsymbol{x}_{-i}$ are replaced by x_i, with equal probability for each subset of k replaced components, while the remaining components in $\boldsymbol{x}_{-i}$ keep their original value. Let the utility function v^0 be defined by:

$$v^0(x_i, \boldsymbol{x}_{-i}) = \mathbb{E}_{\boldsymbol{p}^0}\left[\pi(x_i, \tilde{\boldsymbol{z}}_{-i}) - \lambda_0 \cdot \sum_{j \neq i} \pi(\tilde{z}_j,\ \tilde{\boldsymbol{z}}_{-j}) \mid \boldsymbol{x}\right] \quad \forall \boldsymbol{x} \in X^n. \tag{27}$$

We note that in v^0 we have a utility function that does not depend on any reference strategy. We note that if $\lambda_0 = 0$ and $p_0^0 = 1$, v^0 is but that of the familiar *Homo oeconomicus*. By contrast, if $\lambda_0 \neq 0$ and $p_0^0 < 1$, the individual evaluates any strategy profile $(x_i, \boldsymbol{x}_{-i})$ by pondering his expected *material payoff advantage* over his neighbors, $\pi(x_i, \tilde{\boldsymbol{z}}_{-i}) - \lambda_0 \cdot \sum_{j \neq i} \pi(\tilde{z}_j, \tilde{\boldsymbol{z}}_{-j})$, if all, some, or none of the others in her island would use the same strategy as herself (drawn randomly according to $\boldsymbol{p}^0$), instead of playing their strategies, given by $\boldsymbol{x}_{-i}$. A positive weight $\lambda_0 > 0$ expresses a form of *envy* or *spite*. If instead $\lambda_0 < 0$, then it is as if individuals care positively, or *altruistically*, about their neighbors' material payoffs. Because the utility function v^0 expresses both a Kantian concern and a comparison with others' material payoffs, we refer to an individual with such a utility function as a *competitive Kantian*. Our next result establishes that selection favors competitive moralists of the variety represented by the utility function v^0, and that all other utility functions, unless they have an identical best reply to some resident equilibrium, are selected against**:**

Proposition 4 *The utility function v^0 is uninvadable under weak selection. A utility function $f \in F$ is invadable under weak selection if there exists a $\tilde{x} \in X(f)$ such that $\tilde{x} \notin X(v^0)$.*

This result provides a precise, operational, and general answer to our initial "as if"-question, for the case of weak selection.[22] To see more concretely what it means, consider the case $n = 2$. Then $p_1^0 = r_0$, and

$$v^0(x_i, x_j) = (1 - r_0) \cdot [\pi(x_i, x_j) - \lambda_0 \pi(x_j, x_i)] + r_0 \cdot [\pi(x_i, x_i) - \lambda_0 \pi(x_i, x_i)]. \tag{28}$$

[22] The goal function $v^0(x_i, \boldsymbol{x}_{-i})$ is the average payoff advantage (increment in individual fitness) to individual i over the p_k^0 distribution. We note that a representation of this goal function more in line with the biologists way of thinking about maximizing behavior; namely, in terms of individual "inclusive fitness" can be obtained by writing $v^0(x_i, \boldsymbol{x}_{-i}) = -c(x_i, \boldsymbol{x}_{-i}) + r_0 b(x_i, \boldsymbol{x}_{-i})$, where the "cost" $c(x_i, \boldsymbol{x}_{-i})$ and "benefit" $b(x_i, \boldsymbol{x}_{-i})$ functions are obtained as the intercept and the least square regression coefficient, respectively, by minimizing predicted and observed payoff advantage over the p_k^0 distribution (with the source of variation being the frequency $k/(n-1)$ of the neighbours of i expressing strategy x_i). Such regression procedures, whether univariate or bivariate, are used to obtain inclusive fitness representations of fitness measures and are prevalent in evolutionary genetics (Frank, 1998, Lehmann et al., 2016, Fisher, 1930).

Reorganizing the right-hand side in (28) reveals motivational factors familiar from behavioral economics. First, by writing

$$v^0(x_i, x_j) \;=\; (1-\lambda_0)(1-r_0)\,\pi(x_i, x_j) + \;\lambda_0 r_0\left[\pi(x_i, x_j) - \pi(x_j, x_i)\right] + \;(1-\lambda_0)\,r_0\pi(x_i, x_i)\,, \tag{29}$$

the utility function can be interpreted as the sum of three terms, where the first represents "pure self-interest" (own material payoff), the second a "comparison with the Joneses" (the difference between own material payoff and that of the neighbor), and the third a "Kantian" concern (what is the "right thing to do if others in the population act like me"). Second, by instead bundling the terms as follows, the utility function can alternatively be interpreted as a combination of pure self-interest, "net altruism" (altruism minus spite), and a Kantian concern:

$$v^0(x_i, x_j) \;=\; (1-r_0)\,\pi(x_i, x_j) \;-\lambda_0 r_0\pi(x_j, x_i) + (1-\lambda_0)\,r_0\pi(x_i, x_i)\,. \tag{30}$$

Irrespective of how one chooses to bundle the terms, an individual with utility function u^0 takes into account not only his or her material self-interest, but also other-regarding motives involving altruism or spite if $\lambda_0(1-r_0) \neq 0$ and a Kantian concern if $(1-\lambda_0)\,r_0 \neq 0$.

Finally, we note that when the material payoff function π is continuously differentiable, the necessary first-order condition for a strategy $\tilde{x}$ to be a symmetric Nash equilibrium strategy of the game in which all players have utility function v^0 boils down to

$$\pi_1\left(\tilde{x}, \tilde{\mathbf{x}}^{(n-1)}\right) \;+\; (n-1)\,\kappa_0 \cdot \pi_n\left(\tilde{x}, \tilde{\mathbf{x}}^{(n-1)}\right) = 0. \tag{31}$$

The coefficient κ_0 thus allows to determine whether, on balance, uninvadable strategies are pro-social, purely individualistic, or anti-social, for any island size. We finally note that the index of assortativity used in some of the literature on the evolutionary stability of strategies and preferences (Bergstrom, 2003, Alger and Weibull, 2013, 2016) can be interpreted as the parameter κ_0 of the present model, and is hence driven both by relatedness and by local competition.

4 Evolutionary scenarios

In this section we calculate the coefficients of relatedness and of local competitiveness in three evolutionary scenarios. We determine whether, on balance, behaviors are pro- or anti-social at the level of material payoffs, i.e., whether the coefficient of scaled relatedness is positive or negative, and how this depends on the different elements of the scenarios.

4.1 Survival

We start by generalizing the baseline transmission scenario in Section 3.2, by allowing for the possibility that adults survive from one demographic time period to the next, that is, the possibility of over-lapping generations. Specifically, assume that at the end of Stage 1 an

individual i who obtained material payoff π_i in the material game survives with probability $s(\pi_i) \in [0,1]$. In this biological scenario the fitness of individual i then writes:

$$w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*) = s(\pi_i) + m \cdot [1 - s(\pi^*)] n \cdot \frac{f(\pi_i)}{nf(\pi^*)} + (1-m) \cdot \left(n - \sum_{j=1}^{n} s(\pi_j) \right) \cdot \frac{f(\pi_i)}{(1-m)\sum_{j=1}^{n} f(\pi_j) + nmf(\pi^*)} . \tag{32}$$

The effect of survival is readily seen by considering the case where the survival probability is constant, $s(\pi_i) = s_0$. We show in the appendix that the coefficient of relatedness then equals

$$r(x,x) = \frac{(1-m)^2 + s_0(1-m^2)}{n - (n-1)(1-m)^2 + s_0[1 + (n-1)m^2]}, \tag{33}$$

and that the coefficient of local competitiveness equals

$$\lambda(x) = \frac{(n-1)(1-m)^2}{n - (1-m)^2}. \tag{34}$$

Local competitiveness is the same as in the baseline scenario without survival (see (19)), since the survival rate is the same for all individuals, and hence it is still only the effect of material payoffs on the fecundities that matter for local competitiveness. By contrast, relatedness is higher (see (18)), since an individual who survives and does not migrate, faces a positive probability of interacting locally with his or her own descendants. For all intermediate migration probabilities, $m \in (0,1)$, both $r(x,x)$ and $\lambda(x)$ lie strictly between zero and one. Moreover, both are decreasing in the migration rate m and $r(x,x)$ is increasing in the survival rate s_0. Substituting (33) and (34) into (14), we obtain:

$$\kappa(x) = \frac{2(1-m)s_0}{2(1-m)s_0 + n[2 - m(1-s_0)]}, \tag{35}$$

which is strictly positive for any $s_0 > 0$. In this scenario, then, a positive survival probability induces pro-sociality. However, note that the degree of pro-sociality diminishes with group size. In fact, it vanishes as groups tend to become infinitely large. Figure 1 shows how $\kappa(x)$ depends on the migration rate m when $s_0 = 1/n$, for $n = 2$ (black solid) and $n = 10$ (black dashed), and when $s_0 = 0.8$ for $n = 2$ (blue) and $n = 10$ (blue dashed), as well as $s_0 = 0$ (pink). Furthermore, pro-sociality declines with m, for any given island size n.

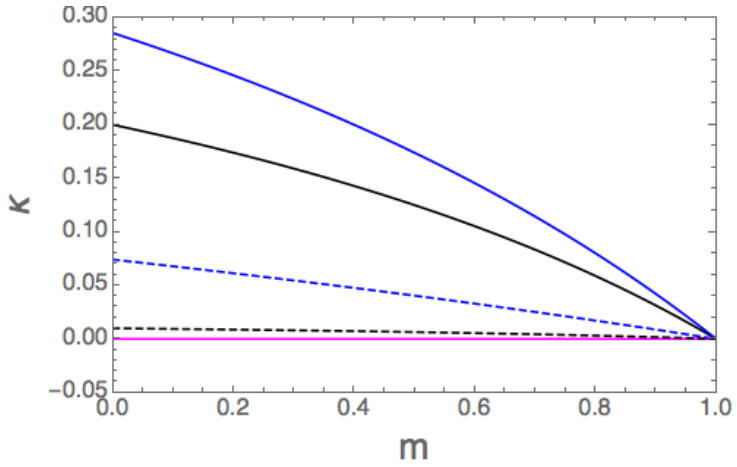

Figure 1: The value of $\kappa(x)$ as a function of the migration rate m.

4.2 Wars

We now revert to the biological scenario with non-overlapping generations, and augment it by introducing wars between groups. Following play of the material game in a demographic time period, but before reproduction, death of the adults, and migration by the offspring, islands are randomly engaged in pairwise wars, under exogenous uniform random matching. In each war, one island wins and the other loses. All individuals in the losing island thus die before they reproduce; the winning island takes over all reproductive resources of the other island and thus doubles its members' fecundity. Technically, the double-sized pool of offspring of the winning island will split in two halves, one for each of the two islands, that they will treat as their "home" island. Let $0 \leq \rho \leq 1$ denote the probability that any given island is drawn into war, the *war risk*, and let $g(\boldsymbol{\pi}, \pi^*)$ denote the conditional probability that an island with material payoff profile $\boldsymbol{\pi} \in \mathbb{R}^n$ wins a war when the average payoff in the rest of the population is π^*, conditional on being drawn into war. Here g is assumed to be increasing and permutation invariant with respect to the material payoffs earned by the inhabitants of the island in question. In other words, g has the properties of standard welfare functions. In this scenario the fitness of individual i writes

$$w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*) = [(1-\rho) + 2\rho g(\boldsymbol{\pi}, \pi^*)] \cdot \left[m \cdot \frac{f(\pi_i)}{f(\pi^*)} + (1-m)\, n \cdot \frac{f(\pi_i)}{(1-m)\sum_{j=1}^{n} f(\pi_j) + nmf(\pi^*)} \right]. \tag{36}$$

The difference with the baseline scenario is the first factor, which contains two terms: the probability that the individual's island will not go to war $(1-\rho)$, and the probability that the island will go to war and win times two $(2\rho g(\boldsymbol{\pi}, \pi^*))$, where the factor two comes from the assumption that a winning island doubles its fecundity and spreads its offspring uniformly over the two islands it now possesses. To see why the second factor is the same as the right-hand side of (17), note that migrants who arrive at any island, irrespective of whether this island has been involved in war or not, come with probability $1-\rho$ from an island that

was not in war, and (recalling that the average probability of winning a war is 1/2) with probability $\rho/2$ from an island that won a war. Moreover, victorious islands send out twice as many migrants as islands that did not go to war. Hence, the expected number of migrants who compete for the breeding spots in any given island is $m\left(1-\rho+2\rho/2\right)\cdot f\left(\pi^*\right)=mf\left(\pi^*\right)$, the same as in the absence of wars.

We show in the appendix that the coefficient of relatedness then coincides with that in the baseline scenario, see (18) in Section 3.2. This is because the only event in which a randomly drawn individual can belong to the same local lineage as a randomly drawn neighbor, is when both belong to an island which did not lose a war, and both stayed in their natal island. Since the risk of losing a war applies to the whole island, while the migration probability applies to the individual, only the latter matters for relatedness.

The coefficient of local competitiveness equals

$$\lambda\left(x\right)=-\frac{\left(n-1\right)\left[2\rho g_n\left(\boldsymbol{\pi}^*,\pi^*\right)-\frac{\left(1-m\right)^2f'\left(\pi^*\right)}{nf\left(\pi^*\right)}\right]}{2\rho g_n\left(\boldsymbol{\pi}^*,\pi^*\right)-\left[n-\left(1-m\right)^2\right]\frac{f'\left(\pi^*\right)}{nf\left(\pi^*\right)}}, \tag{37}$$

where g_n denotes the partial derivative of g with respect to the n-th argument (since g is evaluated in a homogenous population here, and since g is invariant under permutation of the n first arguments, g_n simply captures the marginal effect of an increase in the material payoff of any island member on the probability of winning a war). While the expression is involved, it can readily be seen (by considering a scenario in which $f'\left(\pi^*\right)=0$, for instance) that the effect of material payoffs on the strength in wars can make $\lambda\left(x\right)$ negative, while in the two scenarios without wars studied above, it is always positive. In other words, conflicts between groups reduce spite, and may even reduce it so much that it turns into altruism, i.e., a positive weight attached to the neighbors' material payoffs. Indeed, by substituting (18) and (37) into (14), we obtain:

$$\kappa\left(x\right)=\frac{\rho}{\rho+\frac{\left(2-m\right)m}{2g_n\left(\boldsymbol{\pi}^*,\pi^*\right)}\frac{f'\left(\pi^*\right)}{f\left(\pi^*\right)}}, \tag{38}$$

which is increasing in the marginal effect g_n on the probability of winning wars.

We next turn to weak selection in order to obtain more explicit results on the effects of wars on local competitiveness and scaled relatedness. Recalling the notation under weak selection (see (2)), let each individual's fecundity be exponentially increasing in the individual's material payoff,

$$f\left(\bar{\pi}_i\right)=f_0\cdot\exp\left(\delta_f\cdot\pi_i\right), \tag{39}$$

where $f_0>0$ is baseline fecundity and $\delta_f>0$ represents the intensity of selection with respect to fecundity. Furthermore, assume that the probability of winning a war depends on the two islands' aggregate material payoffs according to

$$g\left(\bar{\boldsymbol{\pi}},\bar{\boldsymbol{\pi}}^*\right)=\frac{\exp\left(V\left(\bar{\boldsymbol{\pi}}\right)\right)}{\exp\left(V\left(\bar{\boldsymbol{\pi}}\right)\right)+\exp\left(V\left(\bar{\boldsymbol{\pi}}^*\right)\right)}, \tag{40}$$

where $\bar{\boldsymbol{\pi}}=\delta_v\cdot\boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}^*=\delta_v\cdot\boldsymbol{\pi}^*$, and $V:\mathbb{R}^n\to\mathbb{R}$ is a strictly increasing symmetric function (like any standard welfare function). Its values $V\left(\bar{\boldsymbol{\pi}}\right)$ and $V\left(\bar{\boldsymbol{\pi}}^*\right)$ represent the "strengths"

of the two islands. This is a logistic version of the Tullock contest function (Tullock, 1980), see Skaperdas (1996). It spans a continuum of cases, from all islands having the same chance to win any war, if the intensity of selection with respect to wars be nil, to the case in which the materially wealthiest island is almost sure to win any war (is the intensity of selection is infinitely large). Letting $\delta_f = \sigma_f \cdot \delta$ in equation (39) and $\delta_v = \sigma_v \cdot \delta$, for non-negative parameters $\sigma_f \geq 0$, $\sigma_v \geq 0$, and $\delta > 0$, we can let both sensitivity parameters tend to zero at proportional rates by focusing on the limit as $\delta \to 0$. Below, however, we let $\sigma_v = \sigma_f$, and thus write δ for δ_v.

Many scenarios can be imagined, of which we consider two. First, if an island's strength is proportional to its total material payoff, i.e., if $V(\bar{\boldsymbol{\pi}}) = \delta \cdot \sum_{i=1}^{n} \pi_i$, then local competitiveness takes the following form (see the appendix):

$$\lambda_0 = \frac{(n-1)(1-m)^2 - \rho(n-1)n/2}{n - (1-m)^2 + \rho n/2}. \tag{41}$$

This changes sign when the risk of war is $\rho^* = 2(1-m)^2/n$; it is positive at lower risks of war and negative at higher risk levels for war. Since in the baseline scenario with non-overlapping generations uninvadability under weak selection requires individuals to be selfish on balance (see Section 3.2), the reduction in local competitiveness that the war risk entails, leads to pro-sociality on balance; indeed, for any $\rho > 0$ we obtain $\kappa_0 > 0$:

$$\kappa_0 = \frac{\rho}{\rho + 2m(2-m)}. \tag{42}$$

Moreover, the threat of war ($\rho > 0$) nourishes pro-sociality: κ_0 is increasing in the risk of war, ρ, and is independent of group size, n.[23] Figure 2 shows κ_0 as a function of the migration rate m, for war risk $\rho = 0$ (the pink curve), $\rho = 0.4$ (the orange curve), and $\rho = 0.8$ (the blue curve).[24]

---

[23] Similar results obtain if groups are instead exposed to the threat of environmental shocks, and if some aggregate measure of the group's material wealth enhances its ability to withstand such shocks (see, e.g., Eshel, 1972, and Aoki, 1982, for models considering such cases under strategy evolution).

[24] The analytical models of Bowles (2006, 2009) for the evolution of "parochial altruism" are also close to our scenario with wars; in particular, the expected number of groups $[1 - \rho + 2\rho\nu(\boldsymbol{\pi}, \bar{\pi}^*)]$ to which a focal group has access for reproduction after warfare also appears in Bowles's formalization. However, since in his model there are no explicit assumptions that allow to close the lifecycle, it is impossible to derive the explicit values of λ_0, r_0, and κ_0 for his model.

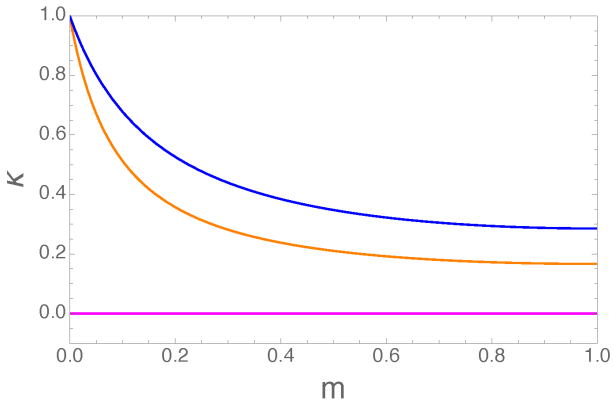

Figure 2: The value of κ_0 as a function of the migration rate m.

Second, while it is arguably a natural bench-mark case to assume that the probability of winning a war depends on the group's total material payoff, sometimes the success or failure in conflicts depends on the strongest or the weakest member of one's group.[25] A general case, that allows for intermediate cases between dependence on the group's total material payoff and its minimal payoff, is obtained by using a CES-functional form. Let

$$V\left(\bar{\boldsymbol{\pi}}\right) = \delta \cdot \left[\sum\nolimits_{i=1}^{n} \pi_i^c\right]^{1/c} \tag{43}$$

for some $c \neq 0$. For $c = 1$ we obtain the previous case, and as $c \to -\infty$, $V\left(\bar{\boldsymbol{\pi}}\right) \to \min_i \{\pi_i\}$ (Leontieff production function). Hence, when c is negative and large in absolute terms, an increase in the poorest group member's material payoff will increase the winning probability, and hence have a positive effect on others' fitness. This suggests a Rawlsian, rather than Benthamite concern for other group members' material well-being. Individuals with medium or high material payoffs may then behave as if they had a particular concern for the poor. Their other-regarding preferences might also be like those proposed in Fehr and Schmidt (1999), according to which people have stronger negative feelings towards inequity when they themselves are poorer than in the opposite case. We finally note that our evolutionary model suggests that a morality term be added to such pro-social concerns, a term that is absent from Fehr's and Schmidt's model. We leave analysis of the implications of such alternative forms for the function V to future research.

4.3 Culture

We now turn to a scenario in which the traits are carried over from one generation to the next by cultural transmission. In every demographic time period, each adult dies and is replaced by exactly one child, who searches a trait to emulate, from its deceased (single) parent, another adult in its island, or an adult in another island. With probability $s\left(\pi_i\right) \in [0,1]$, the loyalty of i's child, the (unique) child of individual i, emulates its parent's trait. With

[25] A host of other hypotheses about group strength could be explored, see, e.g., Konrad (2014) and the references therein.

probability $1-m$ a non-loyal child searches for a trait to emulate among the (now dead) grown-ups in its natal island (including its own parent). With the complementary probability, $m>0$, such a child draws a sample of n grown-ups from the population at large, and emulates the trait of one of them. The probability that an adult on any island is chosen as role model, when compared to others in her island (by a non-loyal child), depends on her trait's attractiveness relative to the attractiveness of the other grown-ups' traits in her island. Likewise, the probability that a child who searches outside its native island will pick a certain island, when looking for a "role model", is assumed to be proportional to the island's relative attractiveness in the world at large.

Let $w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)$ be the expected number of children who emulate their trait from an individual with material payoff π_i when the other island members earn the material payoff vector $\boldsymbol{\pi}_{-i}$, and individuals in all other islands earn material payoff π^*. Then

$$w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right) = s\left(\pi_i\right) + \ m \cdot \left[1 - s\left(\pi^*\right)\right] \cdot \frac{f\left(\pi_i\right)}{f\left(\pi^*\right)} + \ (1-m) \cdot \left(n - \sum_{j=1}^{n} s\left(\pi_j\right)\right) \cdot \frac{f\left(\pi_i\right)}{\sum_{j=1}^{n} f\left(\pi_j\right)}, \tag{44}$$

where, for any individual j in i's island, $f\left(\pi_j\right) > 0$ is the attractiveness of the trait used by j. The first term in (44) is the probability that i's child loyally emulates its parent's trait, without comparison with other adults' traits.[26] The second term concerns the event that children from other islands emulate their trait from one of the parents on i's island. Written more explicitly, this term can be spelled out as

$$mn\left[1 - s\left(\pi^*\right)\right] \cdot \frac{\sum_{j=1}^{n} f\left(\pi_j\right)}{n \cdot f\left(\pi^*\right)} \cdot \frac{f\left(\pi_i\right)}{\sum_{j=1}^{n} f\left(\pi_j\right)}, \tag{46}$$

where the first factor is the expected number of children who search outside their native islands, the second factor is the probability for each such child to decide for i's island, and the third is the conditional probability that it will then choose i as role model. The third term concerns the event that some or all the children in i's island emulate their trait from one among the parents on the island. The product of the first two factors in this term is the expected number of such children and the third factor is the probability, for each such child, that it will choose to imitate individual i. Note that, comparing this scenario to the biological scenario with overlapping generations, loyalty plays a similar role to survival, and

[26] When $m=1$ and payoff affects only attractiveness so that $f\left(\pi_i\right) = f\left(\pi^*\right)$), then (44) writes

$$w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right) = s\left(\pi_i\right) + 1 - s\left(\pi^*\right), \tag{45}$$

which is the individual fitness function under the demographic assumptions of the Bisin and Verdier (2001) model when the mutant frequency vanishes (to see this, set q^i to 0 in equation (3) in Bisin and Verdier, 2001). Comparing (44) to (45) shows the similarity and the difference between the two models from a demographic point of view. By contrast to our model, the population in Bisin and Verdier (2001) is not structured into islands and they assume no effects of traits on f. Furthermore, there is no strategic interaction between individuals in Bisin and Verdier (2001).

attractiveness to fecundity. Moreover, the *cultural import propensity* m plays a similar role to migration. (These observations motivated the notation.)

Then:

$$r(x,x) = \frac{(1-m)^2 + s(\pi(\mathbf{x}))(1-m^2)}{n-(n-1)(1-m)^2 + s(\pi(\mathbf{x}))[1+(n-1)m^2]}, \tag{47}$$

where $\mathbf{x} = (x, ..., x) \in X^n$, and

$$\lambda(x) = \frac{(n-1)(1-m)}{n-1+m}, \tag{48}$$

which leads to

$$\kappa(x) = -\frac{(1-m)[1-s(\pi(\mathbf{x}))]}{2n-[m(n-1)+1][1-s(\pi(\mathbf{x}))]}. \tag{49}$$

Comparison with the biological scenario with overlapping generations reveals that the coefficients of relatedness are the same, but that for any $m < 1$ the coefficient of local competitiveness is larger under cultural transmission. The enhanced competitiveness is strong enough to lead to anti-sociality, since $\kappa(x) < 0$ obtains if and only if $(1-m)[1-s(\pi(\mathbf{x}))] < (2-m[1-s(\pi(\mathbf{x}))])\cdot n$, an inequality which holds for all parameter values. In this example, cultural transmission thus leads to anti-sociality, and anti-sociality is stronger at low values of m. This is because a low cultural import rate enhances local competitiveness. Note that although genetic and cultural transmission here lead to opposite predictions regarding sociality, one qualitative similarity that appears is that like survival under genetic transmission, loyalty under cultural transmission has a positive effect on sociality, $\kappa(x)$. We also note that the negative pro-sociality vanishes as groups tend to become infinitely large: $\kappa(x) \to 0$ as $n \to \infty$.

To illustrate this, Figure 3 shows that $\kappa(x)$ is strictly negative for all $m < 1$, for different loyalty rates and different island sizes: for $s_0 = 0$ and $n = 2$ (the pink curve), $s_0 = 0.4$ and $n = 2$ (the orange curve), $s_0 = 0.8$ and $n = 2$ (the blue curve), $s_0 = 0$ and $n = 10$ (the pink dashed curve), $s_0 = 0.4$ and $n = 10$ (the orange dashed curve), $s_0 = 0.8$ and $n = 10$ (the blue dashed curve).

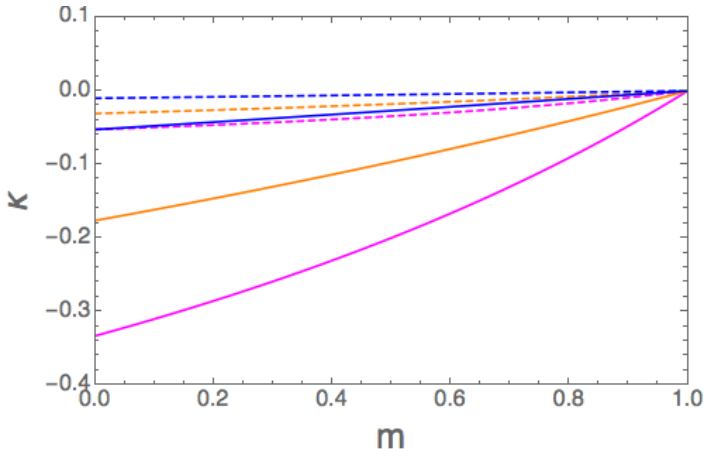

Figure 3: The value of $\kappa(x)$ as a function of the cultural import rate m, for different degrees of background loyalty of offspring towards parents.

4.4 Approximation of the neutral distribution

In the three scenarios above, the coefficients of relatedness and of local competitiveness where obtained in closed form. By contrast, the neutral distribution p_k^0, required to fully determine the v^0 utility function, cannot be obtained in closed form for these scenarios. However, an approximate closed-form expression, in terms of n and r_0, can be obtained. Standard populations genetics results (see e.g., Lessard, 2007, and references therein) suggest that the neutral distribution of types in an island model with constant group size, and with population share of mutants $\varepsilon > 0$, is well approximated by way of the hypergeometric distribution

$$\phi_j(\varepsilon) = \binom{j + \omega\varepsilon - 1}{j} \binom{n - j + \omega(1-\varepsilon) - 1}{n - j} \Big/ \binom{n + \omega - 1}{n}, \tag{50}$$

where $\phi_j(\varepsilon)$ is the probability that there are $j = 0, 1, .., n$ mutants in any given group, and $\omega = r_0/(1 - r_0)$ (see Lessard, 2007, equation (7)). Since $\boldsymbol{p}^0 = \left(p_0^0, ..., p_{n-1}^0\right)$ is the limit distribution when $\varepsilon \to 0$ of the number of *other* mutants in a given mutant's group, we have

$$p_k^0 = \lim_{\varepsilon \to 0} \; (k+1)\, \phi_{k+1}(\varepsilon) \Big/ \left(\sum_{j=1}^{n} j \phi_j(\varepsilon) \right), \tag{51}$$

for $k = 0, 1, ..., n - 1$. Upon rearrangements, this produces

$$p_k^0 = \binom{n}{k+1} \cdot \frac{(k+1)\,\omega}{n} \cdot \frac{\Gamma(k+1)\Gamma(\omega + n - k - 1)}{\Gamma(\omega + n)}, \tag{52}$$

where Γ is the gamma function. This distribution depends only on group size n and pairwise relatedness r_0.

Numerical comparison between this approximation for the above evolutionary scenarios (that can all be subsumed under the relatedness in (47)) and the exact distribution shows that the average total variation between the approximate and exact distributions is quite small. Sampling randomly 10 000 values of s and m when $n = 5$ gives an average total variation of 0.005, a variation that should diminish with n (see Lessard, 2007). It can also be shown that in the special case of a Moran process ($s(\pi_i) = 1/n$) the approximation is in fact exact. (Indeed it can be verified that the expression in (52) then reduces to equation D.6 in Lehmann, Alger, and Weibull, 2015.)

5 Relations to the evolutionary biology literature

Our analysis builds on the *island model*, originally due to Wright (1931), to formally analyze how migration patterns between groups, and the competition for resources within and between groups, may impact the qualitative nature of the preferences that evolutionary forces select for. In evolutionary biology, the island model is the textbook model for understanding the effect of migration and group size on population structure (e.g., Cavalli-Sforza and Bodmer, 1971, Hartl and Clark, 2007; with our equation (18) being the paradigmatic

measure of relatedness derived by Wright 1943), and it has become a work-horse model to analyze conditions favoring pro- and anti-sociality at the level of fitness components (survival and reproduction) in spatially structured populations. The early literature considered traits affecting environmentally induced group extinction (e.g., Eshel, 1972, Aoki, 1982); this is reminiscent of our scenario with wars, but with the probability of surviving a shock depending only on the behaviors in the focal group. This research produced the insight that limited migration tends to favor pro-sociality, at the fitness level. Later work showed that when traits impact individual fecundity, population structure tends to strengthen local competition, ultimately leading to a situation where local competition cancels the effect of relatedness (Wilson, Pollock, and Dugatkin, 1992, Taylor 1992a). These results spurred an extensive theoretical literature seeking to delineate how the assumptions pertaining to demography, life-history, the environment, and the modes of transmission, tip the balance in favor of pro- or anti-sociality at the survival or fecundity level, and qualitatively cover our three transmission scenarios (e.g., Taylor and Irwin, 2000, Gardner and West, 2006, Johnstone and Cant, 2008, Lehmann, Foster, and Feldman, 2008, Lion and Gandon, 2010, Bao and Wild, 2012, and Alberto et al. 2017, for a few relevant examples). In such models with differentiable fitness functions, one generally obtains a first-order condition at the vital rate level which is similar to the equation that we derive at the level of material payoffs (our equation (16)), but with material payoff being replaced by survival or fecundity. In this first-order condition a *scaled relatedness* coefficient aggregates all the consequences of spatial structure. This work is reviewed by Lehmann and Rousset (2010); see also Van Cleve (2015) and Dos-Santos and Peña (2017).

Since we assumed that vital rates are functions of material payoffs, in our setup scaled relatedness is represented by the marginal rate of payoff substitution, $\kappa(x)$. By contrast to the previous literature, our derivation of the first-order condition at the level of material payoffs (equation (16)) allows us to have a single substitution rate, even when strategies affect multiple vital rates. A common feature of the biological literature is also that scaled relatedness is expressed directly in terms of specifics of the transmission process (e.g., migration rate, group size, probability of war, cultural loyalty, see the aforementioned references). By contrast, we decomposed scaled relatedness for constant demography, by showing exactly how it depends on pairwise relatedness on the one hand, and local competitiveness on the other hand. While the pairwise relatedness and scaled relatedness coefficients for the baseline and overlapping-generations scenarios (eq. 33 and eq. 35) have been studied before (see e.g., Taylor 1992, Taylor and Irwin 2000, Lehmann and Rousset, 2010, Akçay and van Cleve 2012), we here derived the associated coefficients of local competitiveness for these two scenarios, which are equivalent and do not depend on survival (i.e., equations (19) and (34)). By contrast, the specifics of our war and cultural transmission scenarios depart from those modelled previously. Hence, the corresponding coefficients of local competitiveness and scaled relatedness coefficients did not appear previously in the literature.[27]

[27] Bowles (2006, 2009) models of the evolution of "parochial altruism" is the closest to our war scenario. In particular, the expected number of groups to which a given group has access for reproduction after warfare, $[(1-\rho)+2\rho g(\boldsymbol{\pi},\pi^*)]$, also appears in Bowles's model. However, since in his model there are no explicit assumptions that allow to close the lifecycle of individuals, it is not possible to derive the explicit values of λ_0, r_0, and κ_0 for his model. For cultural transmission, the closest model to our own scenario is the

The separation of scaled relatedness into coefficient of relatedness and coefficient of local competitivens is valuable for two reasons. First, it allows to disentangle the pathways that affect pro- and anti-sociality. Second, in the uninvadable goal function this decomposition shows how pairwise relatedness and local competitiveness contribute to the weights attached to self-interest (own material well-being), altruism or spite (others' material well-being), and the Kantian concern (the material well-being that would arise should others choose the same strategy).

In sum, our current approach is consistent with, and extends previous formalizations of strategy evolution in the island model. Furthermore, with our evolutionary scenarios we capture, qualitatively and quantitatively, canonical outcomes on the pro- and anti-sociality spectrum, from strong pro-sociality (when competition is mainly between groups) to strong anti-sociality (when competition is mainly within groups).

6 Conclusion

Going back to the advent of life on Earth, we all have a huge number of ancestors, each of whom managed to survive until (s)he successfully reproduced. Our recent ancestors most likely lived in small groups of hunter-gatherers (probably ranging from 5 to 150 grown-ups), extending beyond the nuclear family, for more than two millions of years (Grueter, Chapais, and Zinner, 2012, Malone, Fuentes, and White, 2012, van Schaik, 2016, Layton et al., 2012). This is part of the environment of evolutionary adaptation of the human lineage (e.g., van Schaik, 2016). Heritable behavioral traits should thus be a reflection of the ability to successfully survive and reproduce in such a group-structured environment, in which individuals' behavior certainly had effects on other group members. Our model represents an attempt to formally analyze how migration patterns between groups, and the competition for resources within and between groups, may impact the qualitative nature of the preferences that evolutionary forces select for, when an individual's preferences guide his behavior in interactions within his group under incomplete information. By applying established methods from population biology, we derive such predictions with the aim to enhance insights of value for economics and other social sciences.

Our analysis reveals that uninvadable preferences are different when expressed at the individual fitness level than when expressed at the level of the material payoffs that drive fitness. At both levels, uninvadable preferences combine self-interest with a Kantian concern, but at the material-payoff level there is also an other-regarding component. The Kantian concern depends on relatedness and is driven by kin selection, that is, the fact that in group-structured populations interactions leading to differential reproduction and survival occur between individuals who are more likely to share a common ancestor (be it cultural or genetic) who lived in their group than individuals sampled randomly from the population. The concern for neighbors' material payoffs is driven by two opposing forces. On the one

model of Lehmann, Foster, and Feldman (2008). With $s_0 = 0$, eq. (49) is equivalent to the anti-social scaled relatedness equation (26) found in this previous work, but which was derived for a public goods game.

hand, a negative concern due to local competition, that is, the reproductive benefit from outcompeting neighbors, whose offspring compete locally with own offspring. On the other hand, a positive concern due to global competition, that is, the reproductive benefit from one's group's ability to win conflicts with other groups (or resist exogenous collective shocks, such as natural disaster).

Because both relatedness and local competition arises as soon as individuals have any propensity to stay in their natal group, our analysis implies that the selfish *Homo oeconomicus* survives evolutionary pressure only in the highly unrealistic scenario in which *all* individuals always leave their natal group. Thus, under any realistic transmission scenario and population structure, and when individuals do not know each others' preferences, evolution leads to social preferences that include (a) own material payoff, (b) a Kantian concern similar to the one expressed by *Homo moralis* preferences (Alger and Weibull, 2013, 2016), and (c) a comparison between own and other group members' material payoffs, akin to altruism/spite (Becker, 1976).

Our analysis further reveals whether, on balance, individuals behave pro-socially or anti-socially at the level of material payoffs, in the sense that equilibrium strategies enhance or reduce others' material payoffs. We show that, on balance, pro-sociality obtains if the coefficient of pairwise relatedness exceeds the coefficient of local competitiveness weighted by some factor depending on the transmission process, i.e., if the coefficient of scaled relatedness is positive, and anti-socially if the reverse is true. Whether the former or the latter is true depends on the specifics of the transmission process, as illustrated by three scenarios. In the base-line biological scenario, survival enhances relatedness, and thus pro-sociality. In the scenario with wars, if a group's probability of winning a war depends positively on all group members' material payoffs, it may be fitness-enhancing to exhibit altruism towards one's neighbor, on top of the Kantian concern. If the risk of war is high enough, this pro-social effect may dominate the counter-social effect induced by local competition, and the net result may be altruism. In such transmission scenarios, a combination of self-interest, altruism, and morality obtains. Finally, in the cultural transmission scenario we consider, behavior is, on balance, always anti-social. In a nutshell, under cultural transmission competition is fiercer than under genetic transmission, and competitiveness then always outweighs relatedness. Under these three transmission scenarios, the matching profiles can be (approximately) characterized in terms of only pairwise relatedness and local competitiveness, which thus allows to fully describe preferences in terms of these two population structure coefficients. We have explored these three transmission scenarios in detail, but a host of other scenarios could be considered. For instance, it would be interesting to model explicit age-structure within groups, and intergenerational transfer of resources in order to understand the evolution of social and time preferences in a cooperative breeding context.

We establish these results under weak requirements on individuals' information and cognition. In particular, individuals need not know the material payoffs to others or the preferences of others. Moreover, our formalization allows for the possibility that in fact there are (finitely) many interactions going on simultaneously, or that are randomly selected, and even that each interaction involves only a subset of the inhabitants in an island. What is required is symmetry in the sense that all individuals face the same probabilities of being involved in

any one of the interactions and that the interaction at hand is aggregative and symmetric, as defined in Section 2. The cognitive assumption is then that individuals understand what interaction is at hand.

It is noteworthy that under weak selection the nature of the derived behaviors and preferences are independent of the nature of the strategic interaction within islands. This follows from the fact that the matching profiles are derived from the population structure, without any reference to material payoffs. In a sense, evolution thus leads to social preferences where the weights attached to its components—self-regard, other-regard, and Kantian concerns—are the same for all strategic interactions that fall into the class studied here (symmetrically aggregative material payoff functions).

We hope that the model framework proposed in this paper, and extensions thereof, will prove helpful in understanding the impact of population structure on the evolution of human behavior and preferences.

7 Appendix

7.1 Proof of Equation (8): Hamilton's (marginal) rule

For x to be uninvadable it must be that, given x, $y = x$ is a local maximum of

$$W(y,x) = \sum_{k=0}^{n-1} p_k(y,x)\, \tilde{w}\big(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\big)\,, \tag{53}$$

where $\mathbf{y}^{(k)}$ is the k-dimensional vector whose components all equal y, and $\mathbf{x}^{(n-1-k)}$ is the $(n-1-k)$-dimensional vector whose components all equal x, or $\frac{\partial W(y,x)}{\partial y}\Big|_{y=x} = 0$. Writing $\tilde{w}_j$ for the partial derivative of $\tilde{w}$ with respect to its j-th argument,

$$\begin{aligned}\frac{\partial W(y,x)}{\partial y} = &\sum_{k=0}^{n-1}\left[\frac{\partial p_k(y,x)}{\partial y}\tilde{w}\big(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\big)\right] + \\ &\sum_{k=0}^{n-1}\left[p_k(y,x)\sum_{j=1}^{k+1}\tilde{w}_j\big(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\big)\right].\end{aligned} \tag{54}$$

Noting that for $y = x$, $\tilde{w}\left(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\right) = \tilde{w}\left(x, \mathbf{x}^{(n-1)}, x\right) = 1$, which is independent of k so that it can be factored out in the first term, and that

$$\sum_{k=0}^{n-1}\left(\frac{\partial p_k(y,x)}{\partial y}\Big|_{y=x}\right) = \frac{\partial}{\partial y}\left(\sum_{k=0}^{n-1} p_k(y,x)\right)\Big|_{y=x} = \frac{\partial}{\partial y}(1)\Big|_{y=x} = 0, \tag{55}$$

the expression simplifies to

$$\frac{\partial W(y,x)}{\partial y}\Big|_{y=x} = \sum_{k=0}^{n-1}\left[p_k(y,x)\sum_{j=1}^{k+1}\tilde{w}_j\big(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\big)\right]\Big|_{y=x}. \tag{56}$$

Permutation invariance further implies that for any $j \geq 2$, $\tilde{w}_j\big(x, \mathbf{x}^{(n-1)}, x\big) = \tilde{w}_n\big(x, \mathbf{x}^{(n-1)}, x\big)$ (it's as if the individual whose marginal type change is under consideration were systematically labeled to appear as the last component in the vector $\mathbf{x}^{(n-1)}$). Noticing also that $\sum_{k=0}^{n-1}\left[p_k(y,x)\tilde{w}_1\big(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\big)\right]\big|_{y=x} = \tilde{w}_1\big(x, \mathbf{x}^{(n-1)}, x\big)$, we can write:

$$\begin{aligned}\left.\frac{\partial W(y,x)}{\partial y}\right|_{y=x} &= \tilde{w}_1\big(x, \mathbf{x}^{(n-1)}, x\big) + \sum_{k=1}^{n-1}\left[p_k(y,x)\sum_{j=2}^{k+1}\tilde{w}_j\left(y, \mathbf{y}^{(k)}, \mathbf{x}^{(n-1-k)}, x\right)\right]\Bigg|_{y=x} \\ &= \tilde{w}_1\left(x, \mathbf{x}^{(n-1)}, x\right) + \sum_{k=1}^{n-1}\left[p_k(x,x)k\tilde{w}_n\big(x, \mathbf{x}^{(n-1)}, x\big)\right] \\ &= \tilde{w}_1\left(x, \mathbf{x}^{(n-1)}, x\right) + (n-1)\,\tilde{w}_n\big(x, \mathbf{x}^{(n-1)}, x\big)\sum_{k=1}^{n-1}\left[\frac{kp_k(x,x)}{(n-1)}\right] \\ &= \tilde{w}_1\big(x, \mathbf{x}^{(n-1)}, x\big) + r(x,x)\tilde{w}_n\big(x, \mathbf{x}^{(n-1)}, x\big)\,,\end{aligned} \tag{57}$$

which owing to permutation invariance can also be written

$$\left.\frac{\partial W(y,x)}{\partial y}\right|_{y=x} = \tilde{w}_1\big(x, \mathbf{x}^{(n-1)}, x\big) + r(x,x)\sum_{j=2}^{n}\tilde{w}_j\big(x, \mathbf{x}^{(n-1)}, x\big)\,. \tag{58}$$

7.2 Proof of Proposition 1

Any Nash equilibrium strategy $\tilde{x}$ in the n-player game in which all players have utility function $u_{\hat{x}}$ satisfies the necessary first-order condition

$$\left[1 - r\left(\hat{x},\hat{x}\right)\right]\cdot\tilde{w}_1\left(\tilde{x}, \tilde{\mathbf{x}}^{(n-1)}, \hat{x}\right) \;+\; r\left(\hat{x},\hat{x}\right)\cdot\sum_{j=1}^{n}\tilde{w}_j\left(\tilde{x}, \tilde{\mathbf{x}}^{(n-1)}, \hat{x}\right) \;= 0, \tag{59}$$

or

$$\tilde{w}_1\left(\tilde{x}, \tilde{\mathbf{x}}^{(n-1)}, \hat{x}\right) \;+\; r\left(\hat{x},\hat{x}\right)\cdot\sum_{j=2}^{n}\tilde{w}_j\left(\tilde{x}, \tilde{\mathbf{x}}^{(n-1)}, \hat{x}\right) \;= 0. \tag{60}$$

If there exists a unique strategy $\hat{x}$ that satisfies (8), and if $\hat{x}$ is uninvadable, then the unique solution to the previous equation must be $\tilde{x} = \hat{x}$. Hence, the unique candidate for a Nash equilibrium strategy in the n-player game in which all players have utility function $u_{\hat{x}}$, is $\hat{x}$. Suppose that $\hat{x}$ indeed is such a Nash equilibrium strategy. Suppose further that $u_{\hat{x}}$ is the resident utility function, and consider some mutant utility function $g \neq u_{\hat{x}}$, whose carriers play some $\tilde{y}$ satisfying (see (11))

$$\tilde{y} \in \arg\max_{y\in X}\sum_{k=0}^{n-1}p_k\left(\hat{x}\right)\cdot g\left(y, \tilde{\mathbf{y}}^{(k)}, \hat{\mathbf{x}}^{(n-k-1)}\right). \tag{61}$$

Since $\hat{x}$ is uninvadable, any such strategy $\tilde{y}$ gives rise to a local lineage fitness $W\left(\tilde{y},\hat{x}\right)$ which does not exceed 1. Hence, $u_{\hat{x}}$ is uninvadable.

7.3 Proof of Lemma 1

Recalling that

$$\tilde{w}(x_i, \boldsymbol{x}_{-i}, x) = w\left(\pi\left(x_i, \boldsymbol{x}_{-i}\right), \left(\pi\left(x_j, \boldsymbol{x}_{-j}\right)\right)_{j\neq i}, \pi^*\left(x\right)\right), \tag{62}$$

we obtain

$$\begin{aligned}\tilde{w}_1\big(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right), x\big) = {} & w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + (n-1) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right),\end{aligned} \tag{63}$$

where $\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}$ denotes the $(n-1)$-dimensional vector whose components all equal $\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)$, and

$$\begin{aligned}\tilde{w}_n\big(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right), x\big) = {} & w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + (n-2) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right).\end{aligned} \tag{64}$$

Substituting the last two equations into the last line of (57) produces

$$\begin{aligned}0 = {} & w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + (n-1) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + r\left(x, x\right)(n-1)\, w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + r\left(x, x\right)(n-1)(n-2) \cdot w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) \\ & + r\left(x, x\right)(n-1)\, w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right) \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right).\end{aligned} \tag{65}$$

Noting that with the notation used in this proof, $\lambda\left(x\right)$ writes

$$\lambda\left(x\right) = -\frac{(n-1)\, w_n\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right)}{w_1\left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right), \left(\pi\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right)\right)^{(n-1)}, \pi^*\left(x\right)\right)}, \tag{66}$$

(65) can be written

$$\pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) + (n-1) \cdot \frac{r\left(x, x\right) - \lambda\left(x\right)\left[\frac{1}{n-1} + r\left(x, x\right)\frac{n-2}{n-1}\right]}{1 - \lambda\left(x\right) r\left(x, x\right)} \cdot \pi_n\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) = 0, \tag{67}$$

or

$$\left[1 - \kappa\left(x\right)\right] \cdot \pi_1\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) + \kappa\left(x\right) \cdot \sum_{k=1}^{n} \pi_k\left(x, \left(\mathbf{y}^{(0)}, \mathbf{x}\right)\right) = 0. \tag{68}$$

7.4 Proof of Proposition 2

Any Nash equilibrium strategy $\tilde{x}$ in the n-player game in which all players have utility function $v_{\hat{x}}$ satisfies the necessary first-order condition

$$[1-\kappa(\hat{x})]\cdot\pi_1\left(\tilde{x},\tilde{\mathbf{x}}^{(n-1)}\right) \;+\; \kappa(\hat{x})\cdot\sum_{j=1}^{n}\pi_j\left(\tilde{x},\tilde{\mathbf{x}}^{(n-1)}\right) \;=0. \tag{69}$$

If there exists a unique strategy $\hat{x}$ that satisfies (16), and if $\hat{x}$ is uninvadable, then the unique solution to the previous equation must be $\tilde{x}=\hat{x}$. Hence, the unique candidate for a Nash equilibrium strategy in the n-player game in which all players have utility function $v_{\hat{x}}$, is $\hat{x}$. Suppose that $\hat{x}$ indeed is such a Nash equilibrium strategy. Suppose further that $v_{\hat{x}}$ is the resident utility function, and consider some mutant utility function $g\neq v_{\hat{x}}$, whose carriers play some $\tilde{y}$ satisfying (see (11))

$$\tilde{y}\in\arg\max_{y\in X}\sum_{k=0}^{n-1}p_k(\hat{x})\cdot g\left(y,\tilde{\mathbf{y}}^{(k)},\hat{\mathbf{x}}^{(n-k-1)}\right). \tag{70}$$

Since $\hat{x}$ is uninvadable, any such strategy $\tilde{y}$ gives rise to a local lineage fitness $W(\tilde{y},\hat{x})$ which does not exceed 1. Hence, $v_{\hat{x}}$ is uninvadable.

7.5 Proof of Proposition 3

To show that $\kappa(x)\in[-1,1]$, we begin by studying $\lambda(x)$. Note that the terms that define $\lambda(x)$ are partial derivatives evaluated in a homogenous population. Furthermore, since population size is constant in a homogenous population, each individual's fitness would remain at 1 following a marginal change in the material payoff of all the individuals in the population. Formally:

$$\frac{\partial w(\pi_i,\boldsymbol{\pi}_{-i},\pi^*)}{\partial\pi_i}\bigg|_{\pi_i=\pi_j=\pi^*}+\sum\nolimits_{j=2}^{n}\frac{\partial w(\pi_i,\boldsymbol{\pi}_{-i},\pi^*)}{\partial\pi_j}\bigg|_{\pi_i=\pi_j=\pi^*}+\frac{\partial w(\pi_i,\boldsymbol{\pi}_{-i},\pi^*)}{\partial\pi^*}\bigg|_{\pi_i=\pi_j=\pi^*}=0. \tag{71}$$

By permutation invariance, and using a more compact notation, this writes $w_1(\cdot)+(n-1)w_n(\cdot)+w_{n+1}(\cdot)=0$. Using this and the assumption $w_1(\cdot)>0$,

$$\begin{aligned}\lambda(x)<1 &\Leftrightarrow -(n-1)w_n(\cdot)<w_1(\cdot)\\ &\Leftrightarrow w_1(\cdot)+w_{\pi^*}(\cdot)<w_1(\cdot),\end{aligned} \tag{72}$$

which is true by Assumption **[M]** *(iii)*.

Since $r(x,x) \in [0,1]$ for all x this implies that $\lambda(x) r(x,x) < 1$, and hence

$$\begin{aligned}
\kappa(x) \leq 1 &\Leftrightarrow r(x,x) - \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right] \leq 1 - \lambda(x) r(x,x) \qquad (73)\\
&\Leftrightarrow \lambda(x)\left[r(x,x) - \frac{1}{n-1} - r(x,x)\frac{n-2}{n-1}\right] \leq 1 - r(x,x)\\
&\Leftrightarrow \lambda(x)\left[\frac{r(x,x)-1}{n-1}\right] \leq 1 - r(x,x)\\
&\Leftrightarrow \lambda(x) \geq -(n-1)\\
&\Leftrightarrow -\frac{(n-1) w_n(\cdot)}{w_1(\cdot)} \geq -(n-1)\\
&\Leftrightarrow w_n(\cdot) \leq w_1(\cdot),
\end{aligned}$$

which is true by virtue of Assumption **[M]** *(ii)*.

We now show that $\kappa(x) \geq -1$. For any $\lambda(x) < 1$, $\kappa(x)$ is increasing in $r(x,x)$. Indeed, the partial derivative of the expression for $\kappa(x)$ with respect to $r(x,x)$ has the same sign as (in this expression $r \equiv r(x,x)$ and $\lambda \equiv \lambda(x)$)

$$\begin{aligned}
&[(n-1)(1-\lambda)+\lambda](n-1)(1-\lambda r) + \lambda(n-1)[r(n-1)(1-\lambda) - \lambda(1-r)] \qquad (74)\\
&= (n-1)(1-\lambda)(n-1+\lambda).
\end{aligned}$$

For the inequality $\kappa(x) \geq -1$ to hold, it is thus sufficient that $\kappa(x) \geq -1$ for $r(x,x) = 0$, a condition which reduces to

$$-\lambda(x) \geq -(n-1), \qquad (75)$$

which is true for any $n \geq 2$ since $\lambda(x) < 1$.

Finally,

$$\begin{aligned}
\kappa(x) \leq r(x,x) &\Leftrightarrow \frac{r(x,x) - \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right]}{1 - \lambda(x) r(x,x)} \leq r(x,x) \qquad (76)\\
&\Leftrightarrow r(x,x) - \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right] \leq r(x,x)[1 - \lambda(x) r(x,x)]\\
&\Leftrightarrow \lambda(x)[r(x,x)]^2 \leq \lambda(x)\left[\frac{1}{n-1} + r(x,x)\frac{n-2}{n-1}\right]\\
&\Leftrightarrow \lambda(x) r(x,x)[1 - [1 - r(x,x)](n-1)] \leq \lambda(x).
\end{aligned}$$

This inequality is true if and only if $\lambda(x) \geq 0$ by virtue of the fact that for all $r(x,x) \in [0,1)$ we have $r(x,x)[1 - [1 - r(x,x)](n-1)] \leq 1$. Likewise, it is clear that $\kappa(x) > r(x,x)$ if and only if $\lambda(x) < 0$.

Finally, the last result stated in the proposition is implied by (15) together with Assumption **[M]** *(i)*.

7.6 Proof of Lemma 2

Let $w : \mathbb{R}^{n+1} \to \mathbb{R}$ be any continuously differentiable fitness function, let $b \in \mathbb{R}$, and let $\mathbf{b}$ denote the vector in $\mathbb{R}^{n+1}$ that has all components equal to b. Then, by virtue of (71),

$$w_1(\mathbf{b}) + \sum_{j=2}^{n} w_j(\mathbf{b}) + w_{n+1}(\mathbf{b}) = 0, \tag{77}$$

where an index $k = 1, ..., n+1$ stands for the partial derivative of w with respect to its k-th argument.

Recalling the definition of $\bar{\pi}$ (see (2)), for any given payoff vector $(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*) \in \mathbb{R}^{n+1}$ a first-order Taylor expansion of w with respect to δ evaluated at δ_0 writes

$$\begin{aligned} w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = w(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*) + (\delta - \delta_0) \cdot w_1(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*) \cdot \pi_i \\ + (\delta - \delta_0) \cdot \sum_{j=2}^{n} \left[w_j(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*) \cdot \pi_j \right] \\ + (\delta - \delta_0) \cdot w_{n+1}(\delta_0\pi_i, \delta_0\boldsymbol{\pi}_{-i}, \delta_0\pi^*) \cdot \pi^* + \mathcal{O}(\delta^2). \end{aligned} \tag{78}$$

Evaluated at $\delta_0 = 0$, this expression writes

$$w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = w(\mathbf{0}) + \delta \cdot w_1(\mathbf{0}) \cdot \pi_i + \delta \cdot \sum_{j=2}^{n} w_j(\mathbf{0}) \cdot \pi_j + \delta \cdot w_{n+1}(\mathbf{0}) \cdot \pi^* + \mathcal{O}(\delta^2), \tag{79}$$

where $w(\mathbf{0}) = 1$, and $\mathbf{0}$ is the null vector in $\mathbb{R}^{n+1}$. By permutation invariance of $w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)$ with respect to the components of $\boldsymbol{\pi}_{-i}$, we may for each $j = 2, ..., n$ write $w_n(\mathbf{0})$ instead of $w_j(\mathbf{0})$. Letting $\beta = w_1(\mathbf{0})$ and $\gamma = -(n-1) w_n(\mathbf{0})$, using (77), and rearranging terms, (79) can thus be written

$$w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = 1 + \delta \cdot \left[\beta \cdot (\pi_i - \pi^*) - \frac{\gamma}{n-1} \sum_{j \neq i} (\pi_j - \pi^*) \right] + \mathcal{O}(\delta^2). \tag{80}$$

Letting

$$\lambda_0 = \frac{\gamma}{\beta} = -\frac{(n-1) w_n(\mathbf{0})}{w_1(\mathbf{0})}, \tag{81}$$

and factoring out $\beta > 0$ from (80), and simply omitting to write the factor δ in the fitness function, we conclude that for small $\delta > 0$,

$$w(\delta\pi_i, \delta\boldsymbol{\pi}_{-i}, \delta\pi^*) = 1 + \delta \cdot \beta \left[\pi_i - \lambda_0 \sum_{j \neq i} \frac{\pi_j}{n-1} - (1 - \lambda_0)\pi^* \right] + \mathcal{O}(\delta^2). \tag{82}$$

This shows that λ_0 quantifies local competitiveness among patch members (Lehmann, Alger, and Weibull, 2015; see also Frank, 1998, and Gardner and West, 2004, for a description, but without a formal derivation, of λ_0).

The next step of the proof consists in obtaining an expression for local linage fitness under weak selection. Under weak selection the evolutionary process is what in biology is called *neutral* (Crow and Kimura, 1970, Ewens, 2004, Gillespie, 2004, and, for an explicit example, Rousset, 2004). Formally, this means that we can write

$$p_k(y,\hat{x}) = p_k^0 + \mathcal{O}(\delta) \qquad \forall k, \tag{83}$$

where $\mathcal{O}(\delta)$ accounts for the deviation (relative to the neutral process) of the strategy-profile distribution induced by selection (i.e., $\delta > 0$) that is at most of order δ, where p_k^0 is strategy-independent. Second, recalling the definition of $\tilde{w}$ (see (1)) and letting $\hat{x}$ denote the resident strategy, (82) can be written

$$\tilde{w}\left(y,\left(\mathbf{y}^{(k)},\hat{\mathbf{x}}\right),\hat{x}\right) = 1 + \delta\beta\cdot\left[\tilde{\pi}^{(k)}\left(y,\hat{x}\right) - (1-\lambda_0)\pi(\hat{\mathbf{x}})\right] + \mathcal{O}(\delta^2), \tag{84}$$

where $\left(\mathbf{y}^{(k)},\hat{\mathbf{x}}\right)$ is the $(n-1)$-dimensional vector with k components equal to y and the remaining components equal to $\hat{x}$, and (see equation (22))

$$\tilde{\pi}^{(k)}\left(y,\hat{x}\right) = \pi\left(y|k\right) - \lambda_0\left[\frac{k}{n-1}\pi\left(y|k\right) + \frac{n-1-k}{n-1}\pi\left(\hat{x}|k\right)\right]. \tag{85}$$

Using (83) and (84), local lineage fitness (see (3)) writes

$$\begin{aligned} W(y,\hat{x}) &= \sum_{k=0}^{n-1} p_k(y,\hat{x})\cdot\tilde{w}\left(y,\left(\mathbf{y}^{(k)},\hat{\mathbf{x}}\right)\right) \\ &= 1 + \delta\beta\cdot\sum_{k=0}^{n-1} p_k^0\cdot\left[\tilde{\pi}^{(k)}\left(y,\hat{x}\right) - (1-\lambda_0)\pi(\hat{\mathbf{x}})\right] + \mathcal{O}(\delta^2). \end{aligned} \tag{86}$$

Recalling the definition of lineage payoff-advantage $\Pi(y,\hat{x})$ (see (21)),this can be written as

$$W(y,\hat{x}) = 1 + \delta\beta\cdot\left[\Pi\left(y,\hat{x}\right) - (1-\lambda_0)\pi(\hat{\mathbf{x}})\right] + \mathcal{O}(\delta^2). \tag{87}$$

Neglecting higher order terms in δ in this equation, the condition for uninvadability $[W(y,\hat{x}) \leq W\left(\hat{x},\hat{x}\right)$ for all $y \in X]$ under weak selection is equivalent to the condition $\Pi(y,\hat{x}) \leq \Pi(\hat{x},\hat{x})$ for all $y \in X$.

Finally, we determine the implications of Assumption **[M]** for the bounds on $\lambda_0 = -\left(n-1\right)\cdot w_n\left(\mathbf{0}\right)/w_1\left(\mathbf{0}\right)$, focusing on the non-trivial case $w_n\left(\mathbf{0}\right) \neq 0$. Part *(ii)* of the assumption implies $-\left(n-1\right) \leq \lambda_0$. Moreover, recalling (77) we obtain $\lambda_0 \leq 1$, with strict inequality when either $w_{n+1}\left(\mathbf{0}\right) < 0$ or $w_{n+1}\left(\mathbf{0}\right) = 0$ and $n > 2$.

7.7 Proof of Proposition 4

To establish the first claim, let $v \in \Theta$ and consider any $(\hat{x},\hat{y}) \in B_{\mathrm{NE}}^0\left(u^0,v\right)$. Then, $u^0(\hat{x},\hat{\boldsymbol{x}}) \geq u^0(\hat{y},\hat{\boldsymbol{x}})$. By definition of u^0, this inequality is equivalent to $\Pi(\hat{x},\hat{x}) \geq \Pi(\hat{y},\hat{x})$. Hence, u^0 is uninvadable under weak selection. For the second claim, let $u \neq u^0$. Assume further that $v = u^0$. By hypothesis in the claim, there exists some $(\hat{x},\hat{y}) \in B_{\mathrm{NE}}^0\left(u,u^0\right)$ for which $\hat{x} \notin X\left(u^0\right)$. Hence, there exists $y \in X$ such that $u^0\left(y,\hat{\mathbf{x}}^{(n-1)}\right) > u^0\left(\hat{x},\hat{\mathbf{x}}^{(n-1)}\right)$, or, equivalently, $\Pi(y,\hat{x}) > \Pi(\hat{x},\hat{x})$.

7.8 Calculating the coefficients of local competitiveness and pairwise relatedness

7.8.1 The base-line evolutionary scenario (Section 3.2)

To calculate $\lambda(x)$ we begin by calculating the partial derivatives needed for this purpose. From the individual fitness function (17), we have

$$\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j} = -\frac{n(1-m)^2 f(\pi_i)\left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^{n} f(\pi_j) + nmf(\pi^*)\right]^2} \tag{88}$$

and

$$\begin{aligned}\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i} &= \frac{n(1-m)\left[\partial f(\pi_i)/\partial \pi_i\right]}{(1-m)\sum_{j=1}^{n} f(\pi_j) + nmf(\pi^*)} \\ &- \frac{n(1-m)^2 f(\pi_i)\left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^{n} f(\pi_j) + nmf(\pi^*)\right]^2} \\ &+ \frac{m\left[\partial f(\pi_i)/\partial \pi_i\right]}{f(\pi^*)}\end{aligned} \tag{89}$$

Since

$$\lambda(x) = -\left(\sum_{j\neq i} \frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\right) \Big/ \left(\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\right)\Bigg|_{\pi_i=\pi_j=\pi^*}, \tag{90}$$

and writing $f'(\pi^*)$ for $\frac{\partial f(\pi_j)}{\partial \pi_j}\Big|_{\pi_j=\pi^*}$, we obtain (upon simplification)

$$\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\Bigg|_{\pi_i=\pi_j=\pi^*} = -\frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)} \tag{91}$$

and

$$\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\Bigg|_{\pi_i=\pi_j=\pi^*} = \frac{f'(\pi^*)}{nf(\pi^*)}\left[n-(1-m)^2\right]. \tag{92}$$

Upon simplification, we thus obtain

$$\lambda(x) = \frac{(n-1)(1-m)^2}{n-(1-m)^2}. \tag{93}$$

To calculate $r(x,x)$, one uses a recursion equation (this is a standard technique for calculating probabilities of identity by descent; see Nagylaki, 1992, and Rousset, 2004, for a background). In the scenario at hand, this equation writes

$$r(x,x) = (1-m)^2\left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right]. \tag{94}$$

The left-hand side is the average probability that, in a monomorphic population in which all individuals play x, the neighbor of a randomly drawn member of a certain local lineage is also a member of this local lineage. The terms on the right-hand side details the events in which this happens. The term on the right hand side corresponds to the event in which neither the individual at hand nor the randomly drawn neighbor migrated in from another island. In this case, there is a probability $1/n$ that they have the same parent, in which case they are both members of the same local lineage with certainty; with the complementary probability, they have different parents, in which case the probability that they are both members of the same local lineage equals $r(x,x)$. Solving (94) for $r(x,x)$ yields

$$r(x,x) = \frac{(1-m)^2}{n-(n-1)(1-m)^2}. \tag{95}$$

7.8.2 Survival (Section 4.1)

To calculate $\lambda(x)$ we begin by calculating the partial derivatives needed for this purpose. Here, from the individual fitness function (32):

$$\begin{aligned}\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j} &= -\frac{\partial s(\pi_j)}{\partial \pi_j} \cdot \frac{(1-m) f(\pi_i)}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \\ &- \left(n - \sum_{j=1}^n s(\pi_j)\right) \cdot \frac{(1-m)^2 f(\pi_i)\left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right]^2}\end{aligned} \tag{96}$$

and

$$\begin{aligned}\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i} &= \frac{\partial s(\pi_i)}{\partial \pi_i} - \frac{\partial s(\pi_i)}{\partial \pi_i} \cdot \frac{(1-m) f(\pi_i)}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \\ &+ \left(n - \sum_{j=1}^n s(\pi_j)\right) \cdot \frac{(1-m)\left[\partial f(\pi_i)/\partial \pi_i\right]}{(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)} \\ &- \left(n - \sum_{j=1}^n s(\pi_j)\right) \cdot \frac{(1-m)^2 f(\pi_i)\left[\partial f(\pi_j)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^n f(\pi_j) + nmf(\pi^*)\right]^2} \\ &+ \left[1 - s(\pi^*)\right] \cdot \frac{m\left[\partial f(\pi_i)/\partial \pi_i\right]}{f(\pi^*)}\end{aligned} \tag{97}$$

Writing $s'(\pi^*)$ for $\left.\frac{\partial s(\pi_j)}{\partial \pi_j}\right|_{\pi_j=\pi^*}$ and $f'(\pi^*)$ for $\left.\frac{\partial f(\pi_j)}{\partial \pi_j}\right|_{\pi_j=\pi^*}$, we obtain (upon simplification)

$$\left.\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_j}\right|_{\pi_i=\pi_j=\pi^*} = -s'(\pi^*) \cdot \frac{1-m}{n} - \left[1 - s(\pi^*)\right] \cdot \frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)} \tag{98}$$

and

$$\left.\frac{\partial w(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*)}{\partial \pi_i}\right|_{\pi_i=\pi_j=\pi^*} = \frac{1}{n}(n+m-1)s'(\pi^*) + \left[1 - s(\pi^*)\right]\frac{f'(\pi^*)}{nf(\pi^*)}\left[n - (1-m)^2\right]. \tag{99}$$

Upon simplification, we thus obtain

$$\lambda(x) = \frac{(n-1)(1-m)\left\{(1-m)\left[1-s(\pi^*)\right]\frac{f'(\pi^*)}{f(\pi^*)} + s'(\pi^*)\right\}}{\left[n-(1-m)^2\right]\left[1-s(\pi^*)\right]\frac{f'(\pi^*)}{f(\pi^*)} + (n+m-1)s'(\pi^*)}. \tag{100}$$

The expression in (34) obtains by setting $s'(\pi^*) = 0$.

In this scenario the recursion equation used to calculate $r(x,x)$ writes

$$\begin{aligned} r(x,x) = {} & \left[s(\pi^*)\right]^2 r(x,x) + 2s(\pi^*)\left[1-s(\pi^*)\right](1-m)\left[\frac{1}{n} + \frac{n-1}{n}r(x,x)\right] \\ & + \left[1-s(\pi^*)\right]^2(1-m)^2\left[\frac{1}{n} + \frac{n-1}{n}r(x,x)\right]. \end{aligned} \tag{101}$$

The left-hand side is the average probability that, in a monomorphic population in which all individuals play x, the neighbor of a randomly drawn member of a certain local lineage is also a member of this local lineage. The terms on the right-hand side details the events in which this happens. The first term on the right hand side corresponds to the event that both the individual at hand and the randomly drawn neighbor survived from the previous period. The second term on the right hand side corresponds to the two events in which either the individual at hand or the randomly drawn neighbor survived from the previous period while the other didn't, and the one who didn't survive from the previous period did not migrate in from another island. In this case, there is a probability $1/n$ that one is the offspring of the other, in which case they are both members of the same local lineage with certainty; with the complementary probability, they are not parent and child, in which case the probability that they are both members of the same local lineage equals $r(x,x)$. The third term on the right hand side corresponds to the event in which neither the individual at hand nor the randomly drawn neighbor survived from the previous period, and neither of them migrated in from another island. In this case, there is a probability $1/n$ that they have the same parent, in which case they are both members of the same local lineage with certainty; with the complementary probability, they have different parents, in which case the probability that they are both members of the same local lineage equals $r(x,x)$. Solving (101) for $r(x,x)$ yields

$$r(x,x) = \frac{(1-m)^2 + (1-m^2)s(\pi^*)}{n-(n-1)(1-m)^2 + \left[1+(n-1)m^2\right]s(\pi^*)}. \tag{102}$$

The expression in (33) obtains by replacing $s(\pi^*)$ by s_0.

7.8.3 Wars (Section 4.2)

In the biological scenario with wars, we obtain from the individual fitness function (36):

$$\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j} = 2\rho\left[\partial g\left(\boldsymbol{\pi}, \pi^*\right)/\partial \pi_j\right] \cdot \left[m \cdot \frac{f\left(\pi_i\right)}{f\left(\pi^*\right)} + (1-m)\, n \cdot \frac{f\left(\pi_i\right)}{(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\pi^*\right)}\right] - \left[(1-\rho) + 2\rho g\left(\boldsymbol{\pi}, \pi^*\right)\right] \cdot (1-m)\, n \cdot \frac{(1-m) f\left(\pi_i\right)\left[\partial f\left(\pi_j\right)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\pi^*\right)\right]^2} \tag{103}$$

and

$$\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i} = 2\rho\left[\partial g\left(\boldsymbol{\pi}, \pi^*\right)/\partial \pi_i\right] \cdot \left[m \cdot \frac{f\left(\pi_i\right)}{f\left(\pi^*\right)} + (1-m)\, n \cdot \frac{f\left(\pi_i\right)}{(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\pi^*\right)}\right] + \left[(1-\rho) + 2\rho g\left(\boldsymbol{\pi}, \pi^*\right)\right] \cdot (1-m)\, n \cdot \cdot \frac{\left[(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\pi^*\right)\right] \partial f\left(\pi_i\right)/\partial \pi_i - (1-m) f\left(\pi_i\right)\left[\partial f\left(\pi_j\right)/\partial \pi_j\right]}{\left[(1-m)\sum_{j=1}^{n} f\left(\pi_j\right) + nmf\left(\pi^*\right)\right]^2} \tag{104}$$

$$+ m\frac{\partial f\left(\pi_i\right)/\partial \pi_i}{f\left(\pi^*\right)}. \tag{105}$$

By permutation invariance of g, write $g_n\left(\boldsymbol{\pi}^*, \pi^*\right)$ for $\left.\frac{\partial g(\boldsymbol{\pi}, \pi^*)}{\partial \pi_j}\right|_{\pi_j = \pi^*}$ for all $j = 1, ..., n$. Since, moreover, $g\left(\boldsymbol{\pi}, \pi^*\right) = 1/2$, upon simplification we obtain:

$$\left.\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j}\right|_{\pi_i = \pi_j = \pi^*} = 2\rho g_n\left(\boldsymbol{\pi}^*, \pi^*\right) - \frac{(1-m)^2 f'\left(\pi^*\right)}{nf\left(\pi^*\right)} \tag{106}$$

and

$$\left.\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i}\right|_{\pi_i = \pi_j = \pi^*} = 2\rho g_n\left(\boldsymbol{\pi}^*, \pi^*\right) - \frac{(1-m)^2 f'\left(\pi^*\right)}{nf\left(\pi^*\right)} + \frac{f'\left(\pi^*\right)}{f\left(\pi^*\right)}, \tag{107}$$

so that

$$\lambda\left(x\right) = -\frac{(n-1)\left[2\rho g_n\left(\boldsymbol{\pi}^*, \pi^*\right) - \frac{(1-m)^2 f'(\pi^*)}{nf(\pi^*)}\right]}{2\rho g_n\left(\boldsymbol{\pi}^*, \pi^*\right) + \left[n - (1-m)^2\right]\frac{f'(\pi^*)}{nf(\pi^*)}}. \tag{108}$$

The recursion equation to calculate $r\left(x, x\right)$ writes

$$r\left(x, x\right) = \left(1-m\right)^2\left[\frac{1}{n} + \frac{n-1}{n} r\left(x, x\right)\right]. \tag{109}$$

In this scenario, the only event in which a randomly drawn individual can belong to the same local lineage as a randomly drawn neighbor, is when both stayed in their natal island. In this case, there is a probability $1/n$ that they have the same parent, in which case they belong to the same local lineage with certainty; with the complementary probability, they have different parents, in which case the probability that they belong to the same local lineage is $r(x,x)$. Solving for $r(x,x)$ yields

$$r(x,x) = \frac{(1-m)^2}{n-(n-1)(1-m)^2}. \tag{110}$$

7.8.4 Wars under weak selection (Section 4.2)

Recall that under weak selection we write the individual fitness of individual i as $w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)$, where $(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*) = (\delta \cdot \pi_i, \delta \cdot \boldsymbol{\pi}_{-i}, \delta \cdot \pi^*)$, and $\delta \geq 0$ represents the intensity of selection (see (2)). From (81) in the proof of Lemma 1, we have

$$\lambda_0 = -\frac{\sum_{j\neq i} \partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)/\partial \bar{\pi}_j|_{\delta=0}}{\partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)/\partial \bar{\pi}_i|_{\delta=0}}. \tag{111}$$

Since, for $\delta = 0$, $\bar{\pi}_i = \bar{\pi}_j = \bar{\pi}^*$, we obtain from (106) and (107) that

$$\frac{\partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)}{\partial \bar{\pi}_j}\bigg|_{\delta=0} = 2\rho g_n(\bar{\boldsymbol{\pi}}^*, \bar{\pi}^*) - \frac{(1-m)^2 f'(\bar{\pi}^*)}{n f(\bar{\pi}^*)} \tag{112}$$

and

$$\frac{\partial w(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}, \bar{\pi}^*)}{\partial \bar{\pi}_i}\bigg|_{\delta=0} = 2\rho g_n(\bar{\boldsymbol{\pi}}^*, \bar{\pi}^*) - \frac{(1-m)^2 f'(\bar{\pi}^*)}{n f(\bar{\pi}^*)} + \frac{f'(\bar{\pi}^*)}{f(\bar{\pi}^*)}. \tag{113}$$

With the expressions for f and g given in (39) and (40), and the assumption that $V(\bar{\pi}_i, \bar{\boldsymbol{\pi}}_{-i}) = \delta\left(\pi_i + \sum_{j\neq i}\pi_j\right)$ (note that we assume that the intensity of selection is the same for fecundity and for the probability of winning wars; one can also allow for different selection intensities), we have:

$$\frac{f'(\bar{\pi}^*)}{f(\bar{\pi}^*)} = 1 \tag{114}$$

and

$$g_n(\bar{\boldsymbol{\pi}}^*, \bar{\pi}^*) = \frac{1}{4}. \tag{115}$$

Hence, we get

$$\lambda_0 = -\frac{(n-1)\left[\rho/2 - \frac{(1-m)^2}{n}\right]}{\rho/2 - \frac{(1-m)^2}{n} + 1}, \tag{116}$$

which upon simplification gives the expression in (41). It can then be verified that, together with the fact that r_0 is given by (110), this gives the expression for κ_0 in (42).

7.8.5 Culture (Section 4.3)

In the cultural scenario, we have from (44):

$$\begin{aligned}\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j} &= -\frac{\partial s\left(\pi_j\right)}{\partial \pi_j} \cdot \frac{(1-m) f\left(\pi_i\right)}{\sum_{j=1}^n f\left(\pi_j\right)} \\ &- \left(n - \sum_{j=1}^n s\left(\pi_j\right)\right) \cdot \frac{(1-m) f\left(\pi_i\right)\left[\partial f\left(\pi_j\right) / \partial \pi_j\right]}{\left[\sum_{j=1}^n f\left(\pi_j\right)\right]^2}\end{aligned} \tag{117}$$

and

$$\begin{aligned}\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i} &= \frac{\partial s\left(\pi_i\right)}{\partial \pi_i} - \frac{\partial s\left(\pi_i\right)}{\partial \pi_i} \cdot \frac{(1-m) f\left(\pi_i\right)}{\sum_{j=1}^n f\left(\pi_j\right)} \\ &+ (1-m)\left(n - \sum_{j=1}^n s\left(\pi_j\right)\right) \cdot \frac{\left[\partial f\left(\pi_i\right) / \partial \pi_i\right]\left[\sum_{j \neq i} f\left(\pi_j\right)\right]}{\left[\sum_{j=1}^n f\left(\pi_j\right)\right]^2} \\ &+ \left[1 - s\left(\pi^*\right)\right] \cdot \frac{m\left[\partial f\left(\pi_i\right) / \partial \pi_i\right]}{f\left(\pi^*\right)}.\end{aligned} \tag{118}$$

Upon simplification, we obtain:

$$\left.\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_j}\right|_{\pi_i = \pi_j = \pi^*} = -\frac{(1-m)}{n}\left[s'\left(\pi^*\right) + \left[1 - s\left(\pi^*\right)\right] \cdot \frac{f'\left(\pi^*\right)}{f\left(\pi^*\right)}\right] \tag{119}$$

and

$$\left.\frac{\partial w\left(\pi_i, \boldsymbol{\pi}_{-i}, \pi^*\right)}{\partial \pi_i}\right|_{\pi_i = \pi_j = \pi^*} = \left(\frac{n-1+m}{n}\right)\left[s'\left(\pi^*\right) + \left[1 - s\left(\pi^*\right)\right] \cdot \frac{f'\left(\pi^*\right)}{f\left(\pi^*\right)}\right]. \tag{120}$$

Hence:

$$\lambda(x) = \frac{(n-1)(1-m)}{n-1+m}. \tag{121}$$

For $r(x,x)$ the recursion equation writes

$$\begin{aligned}r(x,x) &= \left[s\left(\pi^*\right)\right]^2 r(x,x) + 2(1-m)\, s\left(\pi^*\right)\left[1 - s\left(\pi^*\right)\right]\left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right] \\ &+ (1-m)^2\left[1 - s\left(\pi^*\right)\right]^2 \cdot \left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right].\end{aligned} \tag{122}$$

The first term on the right-hand side corresponds to the event that both the individual at hand and the randomly drawn neighbor have been loyal to their parents, where the neighbor's parent belongs to the individual's lineage with probability $r(x,x)$. The second term on the right hand side corresponds to the event that either the individual at hand was loyal to its parent but the randomly drawn neighbor was not loyal to its parent, or the other way around. In this case, there is a probability $1/n$ that the non-loyal child acquired its trait from the

loyal child's parent, in which case they both belong to the same local lineage with certainty, while with the complementary probability this did not happen, in which case the probability that the randomly neighbor belongs to the same local lineage is $r(x,x)$. The third term on the right hand side corresponds to the event that neither the individual at hand nor the randomly drawn neighbor were loyal to their parents but both of them acquired their trait from someone in the island. In this case, there is a probability $1/n$ that they acquired their type from the same adult, in which case they belong to the same local lineage with certainty; with the complementary probability they have different cultural parents, in which case the probability that the randomly drawn neighbor belongs to the same local lineage as the individual at hand is $r(x,x)$. We note that the equation simplifies to

$$r(x,x) = [s(\pi^*)]^2 r(x,x) + 2s(\pi^*)[1-s(\pi^*)](1-m)\left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right] \tag{123}$$
$$+ [1-s(\pi^*)]^2 (1-m)^2 \left[\frac{1}{n} + \frac{n-1}{n} r(x,x)\right].$$

The expression in the text obtains upon observing that this equation is identical to the one in (101).

References

[1] Akçay, E., and J. van Cleve (2013): "Behavioral responses in structured populations pave the way to group optimality," *American Naturalist* 179, 257-269.

[2] Akerlof, G. and R. Kranton (2000): "Economics and Identity," *Quarterly Journal of Economics,* 115, 715-753.

[3] Alberto, J. C. Micheletti, D. Ruxton and A. Gardner (2017): "Intrafamily and intragenomic conflicts in human warfare", *Proceedings of the Royal Society of London Series B* (Biological Sciences) 284, issue 1849.

[4] Alexander, R.A. (1979): *Darwinism and Human Affairs*. University of Washington Press, Seattle.

[5] Alger, I. and R. Renault (2006): "Screening Ethics when Honest Agents Care about Fairness," *International Economic Review*, 47, 59-85.

[6] Alger, I. and J. Weibull (2010): "Kinship, Incentives and Evolution," *American Economic Review*, 100, 1725-1758.

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

[8] Alger, I., and J.W. Weibull (2013): "Homo moralis—Preference evolution under incomplete information and assortative matching," *Econometrica* 81, 2269-2302.

[9] Alger, I., and J.W. Weibull (2016): "Evolution and Kantian morality," *Games and Economic Behavior* 98, 56-67.

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

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

[12] Aoki, K. (1982): "A condition for group selection to prevail over counteracting individual selection," *Evolution* 36, 832–842.

[13] Bao, M., and G. Wild (2012): "Reproductive skew can provide a net advantage in both conditional and unconditional social interactions," *Theoretical Population Biology* 82, 200-208.

[14] Becker, G. (1976): "Altruism, egoism, and genetic fitness: economics and sociobiology," *Journal of Economic Literature* 14, 817-826.

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

[16] Bergstrom, T. (1995): "On the evolution of altruistic ethical rules for siblings," *American Economic Review* 85, 58-81.

[17] Bergstrom, T. (1996): "Economics in a family way," *Journal of Economic Literature* 34, 1903-1934.

[18] Bergstrom, T. (2003): "The algebra of assortative encounters and the evolution of cooperation," *International Game Theory Review* 5, 211-228.

[19] Bisin, A., and T. Verdier (2001): "The economics of cultural transmission and the dynamics of preferences", *Journal of Economic Theory* 97: 298-319.

[20] Bowles, S. (2006): "Group competition, reproductive leveling, and the evolution of human altruism," *Science* 314, 1569-1572.

[21] Bowles S. (2009): "Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors?" *Science* 324, 1293-8.

[22] Bowles, S., and H. Gintis (1998): "The moral economy of communities: structured populations and the evolution of pro-social norms," *Evolution and Human Behavior* 19, 3-25.

[23] Brekke, K.A., S. Kverndokk, and K. Nyborg (2003): "An Economic Model of Moral Motivation," *Journal of Public Economics*, 87, 1967–1983.

[24] Bshary, R., and R. Bergmüller (2008): "Distinguishing four fundamental approaches to the evolution of helping," *Journal of Evolutionary Biology* 21, 405-20

[25] Cavalli-Sforza, L.L., and W.F. Bodmer (1971): *The Genetics of Human Populations.* W.H. Freeman, San Francisco.

[26] Crow, J. F. and M. Kimura (1970): *An Introduction to Population Genetics Theory.* Harper and Row, New York.

[27] Curry, P.A., and J.E. Roemer (2012): "Evolutionary stability of Kantian optimization," *Review of Public Economics* 200, 131-146.

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

[29] dos Santos, M., and J. Peña (2017): "Antisocial rewarding in structured populations," *Scientific Reports* 7, 6212.

[30] Eshel, I. (1972): "On the neighbor effect and the evolution of altruistic traits," *Theoretical Population Biology* 11, 258–277.

[31] Ewens, W. J. (2004): *Mathematical Population Genetics.* Springer Verlag, New York.

[32] Fehr, E., and K. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics* 114, 817-868.

[33] Fershtman, C. and K. Judd (1987): "Equilibrium incentives in oligopoly," *American Economic Review* 77, 927–940.

[34] Fershtman, C., and Y. Weiss (1998): "Social rewards, externalities and stable preferences," *Journal of Public Economics* 70, 53-73.

[35] Fisher, R. A. (1930): *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

[36] Frank, S.A. (1998): *Foundations of Social Evolution.* Princeton University Press, Princeton, NJ.

[37] Gardner, A., and S.A. West (2004), "Spite and the scale of competition," *Journal of Evolutionary Biology* 17, 1195-1203.

[38] Gardner, A., and S.A. West (2006): "Demography, altruism, and the benefits of budding," *Journal of Evolutionary Biology* 19, 1707-1716.

[39] Gillespie, J. H. (2004): *Population Genetics: a Concise Guide.* Johns Hopkins, Baltimore & London.

[40] Grafen, A. (1985): "A geometric view of relatedness," in Dawkins, R. and M. Ridley, eds., *Oxford Surveys in Evolutionary Biology.* Oxford University Press, Oxford.

[41] Grueter, C.C., B. Chapais, and D. Zinner (2012): "Evolution of multilevel social systems in nonhuman primates and humans," *International Journal of Primatology* 33, 1002–1037.

[42] Hamilton, W.D. (1964): "The genetical evolution of social behaviour," *Journal of Theoretical Biology* 7, 1-52.

[43] Hamilton, W.D. (1967): "Extraordinary sex ratios: a sex-ratio theory for sex linkage and inbreeding has new implications in cytogenetics and entomology *Science* 156, 477-88.

[44] Hamilton, W.D. (1971): "Selection of selfish and altruistic behavior in some extreme models," in J.F. Eisenberg and W. S.Dillon, eds. *Man and beast: comparative social behavior*. Smithsonian Institution Press, Washington, D.C.

[45] Hartl, D.L., and A.G. Clark (2007): *Principles of Population Genetics* (4th edition). Sinauer and Associates , Sunderland, MA.

[46] Heifetz, A., C. Shannon, and Y. Spiegel (2007a): "What to maximize if you must," *Journal of Economic Theory* 133, 31-57.

[47] Heifetz, A., C. Shannon, and Y. Spiegel (2007b): "The dynamic evolution of preferences," *Economic Theory* 32, 251-286.

[48] Hirshleifer, J. (1977): "Economics from a biological viewpoint," *Journal of Law and Economics* 20, 1-52.

[49] Johnstone, R.A., and M.A. Cant (2008) "Sex differences in dispersal and the evolution of helping and harming," *American Naturalist* 172, 318-30.

[50] Konrad, K.A. (2014) "Strategic aspects of fighting in alliances," in K. Wärneryd, ed. *The Economics of Conflict: Theory and Empirical Evidence*. MIT Press, Cambridge, MA.

[51] Layton, R., S. O'Hara, and A. Bilsborough (2012): "Antiquity and social functions of multilevel social organization among human hunter-gatherers," *International Journal of Primatology* 33, 1215–1245.

[52] Lehmann, L., I. Alger, and J.W. Weibull (2015): "Does evolution lead to maximizing behavior?" *Evolution* 69-7, 1858–1873.

[53] Lehmann, L., and Balloux (2007): "Natural selection on fecundity variance in subdivided populations: kin selection meets bet hedging," *Genetics* 176, 361-377.

[54] Lehmann, L., K., Foster, and F. Feldman (2008): "Cultural transmission can inhibit the evolution of altruistic helping," *American Naturalist* 172, 12-24

[55] Lehmann, L., C. Mullon, E. Akçay, and J. Van Cleve (2016): " Invasion fitness, inclusive fitness, and reproductive numbers in heterogeneous populations," *Evolution* 70, 1689–1702.

[56] Lehmann, L., and F. Rousset (2010): "How life history and demography promote or inhibit the evolution of helping behaviours," *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 2599-2617.

[57] Lessard, S. (2007): "An exact sampling formula for the Wright-Fisher model and a solution to a conjecture about the finite-island model," *Genetics* 177, 1249-1254.

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

[59] Lion, S., and S. Gandon (2010): "Life history, habitat saturation and the evolution of fecundity and survival altruism," *Evolution* 64, 1594-606.

[60] Malone, N., A. Fuentes, and F.J. White (2012): "Variation in the social systems of extant hominoids: comparative insight into the social behavior of early hominins," *International Journal of Primatology* 33, 1251–1277.

[61] Maynard Smith, J. (1964): "Group selection and kin selection," *Nature*, 201, 1145- 1147.

[62] Maynard Smith, J., and G.R. Price (1973): "The logic of animal conflict," *Nature* 246: 15-18.

[63] Michod, R.E., and W.D. Hamilton (1980): "Coefficients of relatedness in sociobiology," *Nature* 288, 694–697.

[64] Nagylaki, T. (1992): *Introduction to Population Genetics*. Springer, Berlin.

[65] Newton, J. (2017): "Shared intentions: The evolution of collaboration," *Games and Economic Behavior* 104, 517-534.

[66] Ohtsuki, H. (2012): "Does synergy rescue the evolution of cooperation? An analysis for homogeneous populations with non-overlapping generations," *Journal of Theoretical Biology,* 307, 20-28.

[67] 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.

[68] Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83, 1281-1302.

[69] Robson, A. (2001): "The Biological Basis of Economic Behavior," *Journal of Economic Literature*, 39, 11-33.

[70] Rogers, A.R. (1994): "Evolution of time preference by natural selection," *American Economic Review*, 84, 460-481.

[71] Rousset, F. (2004): *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

[72] Schaffer, M.E. (1988): "Evolutionarily stable strategies for finite populations and variable contest size," *Journal of Theoretical Biology* 132, 467-478.

[73] Skaperdas, S. (1996): "Contest success functions," *Economic Theory* 7, 283-290.

[74] Taylor, P.D. (1992a) "Altruism in viscous populations - an inclusive fitness model," *Evolutionary Ecology* 6, 352-356.

[75] Taylor, P.D. (1992b) "Inclusive fitness in a homogeneous environment," *Proceedings of the Royal Society B*, 249, 299-302.

[76] Taylor, P.D., and A.J. Irwin (2000) "Overlapping generations can promote altruistic behavior," *Evolution* 54, 1135-41.

[77] Taylor, P.D., and S. Frank (1996) "How to make a kin selection model," *Journal of Theoretical Biology* 180, 27-37.

[78] Tullock, G. (1980): "Efficient rent seeking," in Buchanan, J., Tollison, R. and Tullock, G., Eds., *Toward a Theory of Rent Seeking Society.* Texas A and M University Press, College Station, 97-112.

[79] Van Cleve, J. (2015): "Social evolution and genetic interactions in the short and long term," *Theoretical Population Biology* 2013, 2-26.

[80] Van Schaik, C. P. (2016) *The Primate Origin of Human Behavior.* Wiley-Blackwell, Hoboken, NJ.

[81] Waldman, M. (1994): "Systematic errors and the theory of natural selection," *American Economic Review*, 84, 482-497.

[82] West, S.A., A. Griffin, and A. Gardner (2007): "Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection," *Journal of Evolutionary Biology* 20, 415-32.

[83] Wild, G., and A. Traulsen (2007): "The different limits of weak selection and the evolutionary dynamics of finite populations," *Journal of Theoretical Biology* 247, 382-390.

[84] Wilson, D.S., G.B. Pollock, and L.A. Dugatkin (1992): "Can altruism evolve in purely viscous populations?," *Evolutionary Ecology* 6, 331-341.

[85] Wright, S. (1931): "Evolution in Mendelian populations," *Genetics* 16, 97–159.

[86] Wright, S. (1943): "Isolation by distance," *Genetics* 28, 114–138.