A Heterogeneous Evolutionarily Stable Population of Moral and Selfish Individuals: Exploring the Diversity of Preferences

Charles Ayoubi[a], Boris Thurm[b]

[a] *Chaire en économie et management de l'innovation, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

[b] *LEURE Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Abstract

Why do individuals take different decisions when confronted with similar choices? This paper investigates whether the answer lies in an evolutionary process. Our analysis builds on recent work in evolutionary game theory showing the superiority of a given type of preferences, *homo moralis*, in fitness games with assortative matching. We adapt the classical definition of evolutionary stability to the case where individuals with distinct preferences coexist in a population. This approach allows us to establish the characteristics of an *evolutionarily stable population*. Then, introducing an assortment matrix for assortatively matched interactions, we prove the existence of a heterogeneous *evolutionarily stable population* in 2×2 symmetric fitness games under constant assortment, and we identify the conditions for its existence. Conversely to the classical setting, we find that the favored preferences in a heterogeneous *evolutionarily stable population* are context-dependent. As an illustration, we discuss when and how an *evolutionarily stable population* made of both selfish and moral individuals exists in a prisoner's dilemma. These findings offer a theoretical foundation for the empirically observed diversity of preferences among individuals.

Keywords: Social Preferences, Homo moralis, Preference evolution, Evolutionary Game Theory, Assortative matching, Homophily
JEL classification: C71, C73

1. Introduction

Although commonly used in the economic literature, the hypothesis of rational agents all pursuing their self-interest fails to explain the diversity in human behavior (Henrich et al., 2001). Empirical evidence shows that individuals exhibit a large heterogeneity in their preferences (Falk et al., 2018). This diversity has been observed in various contexts such as voting behavior (Piketty, 1995), altruism (Andreoni and Miller, 2002), environmental consciousness (Schlegelmilch et al., 1996), and risk aversion (Burks et al., 2009), suggesting the existence of distinct preferences among individuals. The findings of (Van Leeuwen et al., 2012), showing that this diversity is also observed in the social behavior of chimpanzees, hint at the possibility of an evolutionary origin behind this heterogeneity.

Email addresses: charles.ayoubi@epfl.ch (Charles Ayoubi), boris.thurm@epfl.ch (Boris Thurm)
Preliminary version. Please do not cite without authors' permission

Our goal in this paper is to assess the evolutionary foundation of the coexistence of more than one type of preference in a population, and to evaluate what types of preferences prevail then.

Scholars have long challenged the choice of selfish utility in economics. Ever since Smith (1759) suggested moral motives in his *Theory of moral sentiments*, economists have considered several alternative preferences such as altruism (Becker, 1974b), warm glow (Andreoni, 1990), fairness (Rabin, 1993), empathy (Stark and Falk, 1998), reciprocity (Fehr and Gächter, 1998), reciprocal altruism (Levine, 1998), inequity aversion (Fehr and Schmidt, 1999) or morality in the Kantian sense[1] (Laffont, 1975; Brekke et al., 2003). Recently, Alger and Weibull (2013, 2016) have provided a theoretical justification for the latter. In a model of preference evolution under incomplete information and assortative matching, they show that a new type of preference, called *homo moralis*, arises endogenously as the most favored by evolution. A *homo moralis* individual maximizes a weighted sum of her selfish *homo oeconomicus* payoff and of her moral payoff, defined as the payoff that she would get if everybody acted like her.[2]

The *homo moralis* preferences elegantly tackle the shortcomings of selfish preferences. However, building on the classical definition of evolutionary stability by Maynard Smith and Price (1973), Alger and Weibull (2013, 2016) investigate the survival of only one type of preference in the society. Maynard Smith and Price (1973) and Maynard Smith (1974) first aimed to identify the strategy providing an evolutionary advantage in animal conflicts between members of a given species. Therefore they defined the concept of *evolutionarily stable strategy*, a strategy adopted by most of the members of a population (called the "resident" strategy) giving a higher reproductive fitness than any other "mutant" strategy. Alger and Weibull (2013) generalize this definition of evolutionary stability, applying it to preference evolution, in order to identify an *evolutionarily stable preference*. A *homo moralis* type of preference emerge in this framework as evolutionarily stable under assortative matching. However, assuming the presence of only one homogeneous resident preference, their approach overlooks the empirically observed heterogeneity of preferences among individuals. Our aim is to fill this gap.

After discussing the conditions for cohabitation of two resident types in a population, we prove the existence of a heterogeneous *evolutionarily stable population* in symmetric 2×2 fitness games and characterize the conditions for this existence. We show that the evolutionarily stable preferences in a heterogeneous population are context-dependent. As an illustration, we display the conditions under which a population made of two kinds of *homo moralis*, the selfish *homo oeconomicus*, and the fully-moral *homo kantiensis*, can coexist and be evolutionarily stable in a prisoner's dilemma.

The organization of the rest of the paper is as follows: in Section 2 we present the model and the main definitions, and we extend the assortment function to a population of several types introducing

[1] Kant (1870) first formulation of his categorical imperative is: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.".

[2] Bergstrom (1995) also showed the evolutionary stability of a "semi-Kantian" utility function (a *homo moralis* with morality coefficient one half) in the special case of symmetric interactions between siblings.

the assortment matrix. In Section 3 we discuss the conditions under which two different types can coexist. In Section 4 we evaluate the conditions for the evolutionary stability of a heterogeneous population. In Section 5 we discuss our results, their main implications and limitations, and we conclude in Section 6.

2. Model

In this section, we present the model and the main definitions. We consider a large population of individuals of different types, i.e. holding different preferences (Section 2.1). While individuals' behavior is driven by their preferences, their evolutionary success is determined by the payoffs they get (Section 2.2). The payoffs the agents obtain partly depend on the counterparts they interact with. Individuals interact in pairs and the matching process is assortative (Section 2.3). We then question the evolutionary success of heterogeneous agents by introducing the concept of *evolutionarily stable population* (Section 2.4). Finally, after defining the *homo moralis* type of preference in Section 2.5, we examine the case of a population of two types of *homo moralis* , namely *homo oeconomicus* and *homo kantiensis*, involved in a prisoners' dilemma (Section 2.6). The rest of the paper, analyzes the evolutionary stability of this population.

2.1 Heterogeneous Population

We consider a large population of individuals whose behaviors depend on their type $\theta_i \in \Theta$, i.e. their preferences. In the classical setting, a population is composed of two types $(\theta_1, \theta_\tau) \in \Theta^2$ (Alger and Weibull, 2013). The two types and their respective shares define a population state $s = (\theta_1, \theta_\tau, \lambda_\tau)$, where $\lambda_\tau \in (0, 1)$ is the population share of θ_τ. If λ_τ is small, θ_1 is called the resident type and θ_τ the mutant type.

We expand the classical model by allowing for the presence of three types[3] $(\theta_1, \theta_2, \theta_\tau) \in \Theta^3$. Let $I = \{1, 2, \tau\}$, then for all $i \in I$, we denote by $\lambda_i \in (0, 1)$ the share of type θ_i in the population. The three types and their respective shares define a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \lambda_\tau)$. By normalizing the population size to unity, we have: $\sum_{i \in I} \lambda_i = 1$. Therefore, the population state s can be described with only two population shares instead of three. For convenience, we will often use $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with λ the relative share of θ_2 with respect to θ_1, i.e. $\lambda = \lambda_2/(\lambda_1 + \lambda_2)$. Note that we have: $\lambda_1 = (1 - \lambda)(1 - \lambda_\tau)$ and $\lambda_2 = \lambda(1 - \lambda_\tau)$.

When λ_τ is small, i.e. when $\lambda_\tau << min(\lambda_1, \lambda_2)$, θ_1 and θ_2 are called the resident types and θ_τ the mutant type.[4] A population with at least two resident types is called *heterogeneous*, while a population with one resident type is called homogeneous.

[3] The model and definitions of this section can be extended to a heterogeneous population of n types. However, given the needs of the rest of the analysis in the sections below, we limit the model here to the case of three types.

[4] By extension, we will sometimes use residents (mutants) to refer to individuals of the resident (mutant) type.

2.2 Fitness game

Individuals are randomly matched into pairs and they engage in a symmetric interaction.[5] Each individual is as likely to be in one or the other side of the interaction. For all $(i,j) \in I$, the conditional probability that an individual of type θ_j is matched with an individual of type θ_i is called $p_{i|j}$.[6] We assume that the common strategy set X is a nonempty, compact and convex set in a topological vector space.[7]

Following Güth and Yaari (1992), we adopt an indirect evolutionary framework. The behavior of individuals, i.e. the strategy they play, is driven by the maximization of personal preferences, which are described by a continuous utility function $u_{\theta_i} : X^2 \to \mathbb{R}$. On the other hand, the individuals' evolutionary success is given by some exogenous payoff (fitness) function π, where we assume $\pi : X^2 \to \mathbb{R}$ to be continuous. The pair $< X, \pi >$ is called the *fitness game*.

To prevent individuals from deviating from their utility-maximization, we consider the individuals' preferences as their private information.[8] A Bayesian Nash Equilibrium (BNE) is then a set of strategies, one for each type, where each strategy is a best reply to the others in the given population state:

Definition 1 (Bayesian Nash Equilibrium)**.** In a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, $(x_1, x_2, x_\tau) \in X^3$ is a type-homogeneous Bayesian Nash equilibrium if:

$$\forall i \in I : \quad x_i \in \underset{x \in X}{\operatorname{argmax}} \quad \sum_{j \in I} p_{j|i} \cdot u_{\theta_i}(x, x_j) \tag{1}$$

The set of Bayesian Nash Equilibria in population state s, i.e. all solutions (x_1, x_2, x_τ) of (Eq. 1), is called $B^{NE}(s) \subseteq X^3$.

Remark 1. The definition of Bayesian Nash equilibrium remains valid when there is no mutant in the population, i.e. when the population is made of two types. In this case, (x_1, x_2) is a Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \lambda)$ if for all $i \in \{1,2\}$, $x_i \in \underset{x \in X}{\operatorname{argmax}} \sum_{j \in \{1,2\}} p_{j|i} \cdot u_{\theta_i}(x, x_j)$.

Property 1. Since in the state $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ the residents are matched between them, as if there were no mutants in the population (cf. Lemma 3 below), if $(x_1^\circ, x_2^\circ) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$, then for

[5] The framework can also be extended to asymmetric interactions with ex-ante symmetry.

[6] Note that all the probabilities are a function of the population state s but we drop this precision for readability purposes.

[7] More precisely, we assume that X is a locally convex Hausdorff space. However, most of our analysis will focus on the simpler case of a finite two-player extensive-form game where X is the set of mixed strategies.

[8] A large body of research has studied preferences evolution under complete and incomplete information, showing that individuals adjust their behavior under complete information (e.g. Robson, 1990; Ellingsen, 1997; Bester and Güth, 1998; Possajennikov, 2000; Ok and Vega-Redondo, 2001; Sethi and Somanathan, 2001; Heifetz et al., 2007; Dekel et al., 2007). For example, suppose that two individuals are playing a prisoner's dilemma, where the first player prefers to defect and the second prefers to cooperate. Under incomplete information, each individual will stick to their original preference. But if the cooperator knows the preference of the defector, then she will deviate and also defect (See also Ockenfels, 1993, for a discussion of cooperation in prisoners' dilemma).

any strategy $x_\tau^\circ \in X$ such that $x_\tau^\circ \in \underset{x \in X}{\text{argmax}} \sum_{j \in I} p_{j|\tau} \cdot u_{\theta_\tau}(x, x_j^\circ)$, we have $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$. Reciprocally, if $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, then $(x_1^\circ, x_2^\circ) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$.

We now define the equilibrium correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1)^2 \rightrightarrows X^3$. This correspondence maps the population share of each type to the associated equilibria. Using the definition of assortativity (see Definition 5, in Section 2.3), it can be extended by continuity to $(0,1) \times [0,1)$ to cover the limit when the mutant share λ_τ goes to zero. The following lemma will be useful for the evolutionary stability analysis:

Lemma 1. *$B^{NE}(s)$ is compact for each $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau) \in \Theta^3 \times (0,1) \times [0,1)$.*
If for all $i \in I$ u_{θ_i} are concave in their first arguments, then $B^{NE}(s) \neq \emptyset$.
The correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is upper hemi-continuous.

Proof. In Appendix AppendixB.1. □

An individual of type θ_i who plays strategy $x_i \in X$ when her opponent of type θ_j plays strategy $x_j \in X$ gets material payoff $\pi(x_i, x_j)$. For simplicity, we will often note $\pi(x_i, x_j) \equiv \pi_{ij}$. The evolutionary success of individuals of a given type depends on the average payoff they perceive. We call this average payoff the type fitness. Formally, we have:

Definition 2 (Type fitness)**.** In a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, let $(x_1, x_2, x_\tau) \in B^{NE}(s)$. For all $i \in I$, the fitness of a type θ_i is given by:

$$\Pi_{\theta_i}(x_1, x_2, x_\tau, s) = \sum_{j \in I} p_{j|i} \cdot \pi(x_i, x_j) \tag{2}$$

2.3 Matching

A key feature of the model lies in the matching process. The matching process determines the probability for two individuals in a population to interact. It therefore affects both the decision process of agents (Nash equilibria) and their evolutionary success (Type fitness). Building on Bergstrom (2003), we consider that the meeting probability between two individuals follows an exogenous[9] assortative matching process rather than the more classical uniform random matching. This assortative matching makes it more likely for a given individual to meet an individual of her same type.

Assortative Matching

In a situation of assortative matching, the probability to meet an individual of type θ_i is not necessarily the same for an individual θ_i and for an individual θ_j, i.e. we can have $p_{i|i} \neq p_{i|j}$. This

[9] Allowing individuals to select their partners (Becker, 1973, 1974a; Gunnthorsdottir et al., 2010; Jackson and Watts, 2010) would require to include informational and strategic features beyond the scope of this study.

contrasts with the case of uniform-random matching in which the probability to meet an individual of type θ_i is always equal to the share λ_i of θ_i in the population, i.e. for all $(i,j) \in I$, $p_{i|j} = p_{i|i} = \lambda_i$.

In the setting with two types in the population, Bergstrom (2003) introduced an assortment function in order to model assortative encounters. Building on his approach, we introduce a type-by-type assortment matrix function allowing for assortative matching in interactions between individuals of three (or more) distinct types.

Definition 3 (Assortment matrix)**.** In a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, for all $(i,j) \in I^2$, let $\phi_{ij}(\lambda, \lambda_\tau)$ be the difference between the conditional probability to be matched with type θ_i, given that the individual herself is of type θ_i, and the probability to be matched with type θ_i, given that the individual is of type θ_j: $\phi_{ij}(\lambda, \lambda_\tau) = p_{i|i} - p_{i|j}$.
For all $(i,j) \in I^2$, $\phi_{ij} : (0,1)^2 \to [-1,1]$. These assortment functions ϕ_{ij} define an exogenous assortment matrix: $\Phi = ((\phi_{ij}(\lambda, \lambda_\tau)))_{(i,j)\in I^2}$.

Extending the concept of assortment function, the assortment matrix embeds *homophily* effects, i.e. the tendency of individuals to interact more with counterparts holding similar characteristics such as family, ethnicity, age, gender, language, religion, geographic proximity, education, work, association activity or income (Ibarra, 1993; McPherson et al., 2001). The assortment matrix allows accounting for the higher probability of interacting with similar others (Byrne, 1971; Lakin and Chartrand, 2003), relating to the notion of distance in network economics (Currarini et al., 2009; Iijima and Kamada, 2017). Some alternative approaches to model *homophily* in an evolutionary framework include evolutionary graph theory and evolutionary set theory (Nowak et al., 2010). In the former, individuals occupy the vertices of a graph and their interactions are governed by edges (Lieberman et al., 2005; Ohtsuki and Nowak, 2008; Shakarian et al., 2012). In the latter, individuals belong to several sets (e.g. school, company, living location, associations, etc.) and the more sets they have in common, the more interactions between them (Tarnita et al., 2009). The assortment matrix defined above is exogenous and hence allows for large flexibility in the setting of the assortment as a function of the state s. It can therefore be used in a variety of contexts such as economics, sociology, biology or management, with the possibility to calibrate its values empirically.

We now introduce a particular type of assortment matrix extending the case of constant assortment often used in single-resident populations (Alger and Weibull, 2012; Salmon and Wilson, 2013) derived from the Wright's coefficient of relatedness in biology (Wright, 1922). This definition will be useful in the evolutionary stability analysis in Section 3 and Section 4.

Definition 4 (Uniformly constant assortment matrix)**.** An assortment matrix Φ is called *uniformly constant* when all of its non-diagonal components are independent of the population shares and equal to the same value.[10] In other words, we will say that Φ is *uniformly constant*[11] when, for all

[10] By definition of the assortment functions, the matrix Φ has a diagonal of zeros.

[11] By extension, we will say that the assortment is *uniformly constant* when the assortment matrix is *uniformly constant*.

$(i,j,k,l) \in I^4$ such that $i \neq j$ and $k \neq l$:

$$\begin{cases} \phi_{ij} : (0,1)^2 \to [-1,1] \quad \text{is} \quad \text{constant}, \\ \phi_{ij}(\cdot) = \phi_{kl}(\cdot) \end{cases}$$

Note that the case of uniform random matching is a special case of uniformly-constant assortment where each assortment function is constant and equal to zero: $\Phi = ((0))_{(i,j)\in I^2}$.

We assume that for all $(i,j) \in I^2$, $\phi_{ij}(\cdot)$ is continuous in λ_τ (the mutant share in the population) and converges as λ_τ goes to zero.We then define the assortativity σ as follows:

Definition 5 (Assortativity)**.** The assortativity $\sigma \in [0,1]$ is the limit for all $i \in \{1,2\}$ of $\phi_{\tau i}$ when λ_τ goes to zero:

$$\forall\, i \in \{1,2\} : \quad \lim_{\lambda_\tau \to 0} \phi_{\tau i}(\lambda, \lambda_\tau) = \sigma$$

Using the definition of assortativity, the assortment functions $\phi_{ij} : (0,1)^2 \to [-1,1]$ can be extended by continuity to $(0,1) \times [0,1)$ to cover the limit when the mutant share θ_τ goes to zero. We will also note $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ the population state when the mutant share goes to zero.

Remark 2. The continuity of the assortment functions and the definition of assortativity $\sigma \in [0,1]$ imply that any uniformly-constant assortment matrix can be written as a function of the unit-matrix[12] J and the identity matrix I as follows: $\Phi = \sigma(J - I)$.

Remark 3. At the limit when λ_τ goes to zero, we have for all $i \in \{1,2\}$, $\phi_{\tau i}(\lambda, 0) = \sigma = p_{\tau\tau}$. Indeed, according to the balancing conditions (see Property 3 below), the probability for a resident to be matched with a mutant $p_{\tau i}$ is zero. Thus, the assortativity is independent of the resident types, and we also have $\sigma \in [0,1]$.

Matching probabilities

The matching process must satisfy some properties in order to be well defined. We detail these properties in this section and show how the matching probabilities can be written as a function of the population shares and the assortment matrix only. In the following, for the sake of readability, we use the notation ϕ_{ij} to designate $\phi_{ij}(\lambda, \lambda_\tau)$.

Property 2 (Matching conditions)**.** The conditional probabilities satisfy the matching conditions if each individual is matched with another individual with probability one, i.e. nobody is left behind without a match:

$$\forall\, i \in I : \quad \sum_{j \in I} p_{j|i} = 1$$

[12] The unit-matrix J is the matrix having each of its components equal to one.

Property 3 (Balancing conditions)**.** The conditional probabilities satisfy the balancing conditions if the probability of the event "being of type θ_i and being matched with an individual of type θ_j" is the same as the probability of the event "being of type θ_j and being matched with an individual of type θ_i":

$$\forall (i,j) \in I^2 : \quad \lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i}$$

The balancing conditions ensure the coherence of the matching process. Similarly, in order to be well defined, the assortment matrix must satisfy some conditions that we call the assortment balancing conditions:

Property 4 (Assortment balancing condition)**.** The assortment matrix satisfies the *assortment balancing conditions* when:

$$\forall (i,j) \in I^2 : \quad \lambda_j \cdot \left[\left(\sum_{k \in I} \lambda_k \phi_{ik} \right) - \phi_{ij} \right] = \lambda_i \cdot \left[\left(\sum_{k \in I} \lambda_k \phi_{jk} \right) - \phi_{ji} \right]$$

If the matching process satisfies the matching and balancing conditions, then the assortment matrix must satisfy the assortment balancing conditions.

Proof. In Appendix AppendixA.1. □

The assortment balancing conditions impose a particular relationship between the assortment functions. As noted by Bergstrom (2003) in the case of assortative encounters between two types, the assortment $\phi_{12} = p_{1|1} - p_{1|2}$ defined between a type θ_1 and a type θ_2 is equal to the assortment $\phi_{21} = p_{2|2} - p_{2|1}$ defined between θ_2 and θ_1. When a third type θ_τ is part of the population, this result does not hold anymore, i.e. we do not necessarily have $\phi_{12}(\lambda, \lambda_\tau) = \phi_{21}(\lambda, \lambda_\tau)$. However, at the limit when the mutant share goes to zero, the residents are matched between them, as if there was no mutants, and thus we get the same relation $\phi_{12} = \phi_{21}$. Formally:

Lemma 2 (Assortment between residents)**.** *When $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, if the matching process satisfies the matching and balancing conditions, then we have $\phi_{12}(\lambda, 0) = \phi_{21}(\lambda, 0)$.*

Proof. In Appendix AppendixA.2. □

Knowing the assortment matrix Φ, we have a system of equations on the conditional probabilities $p_{i|j}$ defined by:

- The matching conditions: for all $i \in I$, $\sum_{j \in I} p_{j|i} = 1$ (Property 2)
- The balancing conditions: for all $(i,j) \in I^2$, $\lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i}$ (Property 3)
- The assortment matrix conditions: for all $(i,j) \in I^2$, $\phi_{ij} = p_{i|i} - p_{i|j}$ (Definition 3)

When the assortment matrix satisfies the assortment balancing conditions, this system has a unique solution, i.e. we can express the conditional probabilities in function of the population shares and assortment functions:

Proposition 1 (Matching probabilities)**.** *When the assortment matrix* Φ *satisfies the assortment balancing conditions (Property 4), the system defined by matching conditions (Property 2), balancing conditions (Property 3) and assortment matrix conditions (Definition 3) has a unique solution:*

$$\forall (i,j) \in I^2 : \quad p_{i|j} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \tag{3}$$

Proof. In Appendix AppendixA.3. □

Remark 4. Since for all $(i,j) \in I^2$, $p_{i|j} \in [0,1]$, the assortment functions should respect another set of conditions to be coherent with the matching process:

$$\forall (i,j) \in I^2 : \quad 0 \leq \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \leq 1$$

Remark 5. Note that under uniform random matching, for all $(i,j) \in I^2$ $\phi_{ij} = 0$ and we obtain $p_{i|j} = \lambda_i$, i.e. each individual is matched with an individual of type θ_i according to the population share λ_i of individuals of type θ_i.

It is also interesting to detail the conditional probabilities $p_{i|i}$:

$$\forall\, i \in I : \quad p_{i|i} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik}$$

The conditional probabilities $p_{i|i}$ are the sum of several terms. The first, λ_i, is the population share of individuals of type θ_i. The others, $\lambda_k \phi_{ik}$, represent the additional matching between individual of type θ_i at the expense of matching with individuals of type θ_k, weighted by λ_k the population share of individuals of type θ_k.

Finally, we will need to know the limits of the conditional probabilities when the mutant share λ_τ goes to zero.

Lemma 3 (Conditional probabilities in a population of two residents and one mutant)**.** *When* $s =$

$(\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, *if Proposition 1 is satisfied, then we have:*

$$\begin{aligned}
p_{1|1} &= (1-\lambda) + \lambda \cdot \phi_{12} \\
p_{1|2} &= (1-\lambda) \cdot (1-\phi_{12}) \\
p_{1|\tau} &= (1-\lambda) \cdot (1-\sigma) - \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{2|1} &= \lambda \cdot (1-\phi_{12}) \\
p_{2|2} &= \lambda + (1-\lambda) \cdot \phi_{12} \\
p_{2|\tau} &= \lambda \cdot (1-\sigma) + \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{\tau|1} &= 0 \\
p_{\tau|2} &= 0 \\
p_{\tau|\tau} &= \sigma
\end{aligned}$$

where $\Gamma = \lim_{\lambda_\tau \to 0} \frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}$.

Proof. In Appendix AppendixA.4 □

Note that when λ_τ goes to zero, we have $p_{\tau|1} = p_{\tau|2} = 0$, and individuals of type θ_1 and θ_2 are matched as if individuals θ_τ were not in the population. The conditional probabilities $p_{1|1}$, $p_{2|1}$, $p_{1|2}$ and $p_{2|2}$ are then consistent with the classical setting (Bergstrom, 2003; Alger and Weibull, 2013).

When the assortment matrix is uniformly constant, we have $\phi_{12} = \sigma$ and $\Gamma = 0$. The limit Γ can be interpreted as the matching probability difference between mutants and residents of the two types: $\Gamma = \lim_{\lambda_\tau \to 0}(p_{\tau 2} - p_{\tau 1})/\lambda_\tau$. In other words, if individuals θ_1 and θ_2 meet the mutants at the same rate when they enter the population, then $\Gamma = 0$, while if residents of one type meet the mutants at a higher rate than the other residents do then $\Gamma \neq 0$. Finally, when the assortment functions $\phi_{\tau 1}$ and $\phi_{\tau 2}$ are right-differentiable in $\lambda_\tau = 0$, we have $\Gamma = \partial\phi_{\tau 1}(\lambda, 0)/\partial\lambda_\tau - \partial\phi_{\tau 2}(\lambda, 0)/\partial\lambda_\tau$.[13] Therefore, Γ is also the marginal assortment difference between mutants and residents of the two types.

2.4 Evolutionarily stable population

In order to analyze the evolutionary stability of a heterogeneous population, we need to extend the concept of *evolutionarily stable preference* (Alger and Weibull, 2013). An *evolutionarily stable population* should respect two conditions. First, the two resident types should earn the same type fitness to coexist. We call this condition the Type-fitness Equality. Second, the population must resist a small-scale invasion of any other type by earning a greater type fitness. Formally:

Definition 6 (Evolutionarily stable population)**.** A population in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ is evolutionarily stable against a mutant type $\theta_\tau \in \Theta$ such that for all $i \in \{1, 2\}$ $\theta_\tau \neq \theta_i$ if:

1. θ_1 and θ_2 earn the same type fitness: $\Pi_{\theta_1}(x_1^\circ, x_2^\circ, s^\circ) = \Pi_{\theta_2}(x_1^\circ, x_2^\circ, s^\circ)$ in all Bayesian Nash equilibria (x_1°, x_2°) in the population state s°;

[13] Because $\phi_{\tau 1}(\lambda, 0) = \phi_{\tau 2}(\lambda, 0) = \sigma$

2. θ_1 and θ_2 earn a greater type fitness than a small share of mutants: there exists an $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.

Moreover, a population is evolutionarily stable if it is evolutionarily stable against all types $\theta_\tau \in \Theta$ such that for all $i \in \{1,2\}$, $\theta_\tau \neq \theta_i$.

The first condition of an *evolutionarily stable population* requires that the two residents earn the same type fitness. In the framework of evolutionary game dynamics, the evolution of strategies (and preferences) is dictated by an evolutionary process called a replicator, which usually depends on the difference between the fitness obtained and the average fitness in the population. If the fitness of a given type is greater than the average fitness, then the population share of this type will increase. Hence, the two resident types should get the same fitness for the population share λ° to be stable.

In the second condition defining an *evolutionarily stable population*, i.e. when the mutants enter the population, we allow the relative share of the two residents λ to change around a small neighborhood of its initial value λ°. However, in this case ($\lambda_\tau > 0$), we only impose that the two residents earn a greater type fitness than the mutant, and not that the two residents earn the same type fitness. Such a condition would be too restrictive. Thus, by entering the population, the mutant could destabilize the residents, i.e. one type could overcome (or invade) the other. To analyze if an *evolutionarily stable population* is robust to mutant entry, one would need to model the evolutionary dynamics. The results would then depend on the evolutionary process selected, which could be challenging in economics since this evolutionary process depend on genetic, cultural and technological transmission (Norton et al., 1998; Van Damme, 1991).

The definition of *evolutionarily stable population* is consistent with the classical setting: an *evolutionarily stable preference* is an *evolutionarily stable population* when there is only one resident type and one mutant type. Moreover, this definition is similar to the concept of *evolutionarily stable configuration* introduced by Dekel et al. (2007). A configuration (a distribution of preference and the associated equilibria) is evolutionarily stable if it is balanced, i.e. if all types earn the same fitness, and if mutants do not outperform residents. Thus, an *evolutionarily stable population* can be understood as an *evolutionarily stable configuration* in which the distribution of preferences consists in the shares of each type. However, there are a few differences between the two definitions. First, the definition of *evolutionarily stable population* applies to preferences, and thus to all Bayesian Nash equilibria of the population. Second, by requiring that the mutant type is different from the residents in the definition of *evolutionarily stable population*, we can impose that resident individuals earn a strictly greater payoff than the mutants. Finally, the introduction of assortative matching limits the analysis to a finite number of types.

We now derive two useful results linking the second condition of evolutionary stability with the situation at the limit when the mutant share goes to zero. Recall that $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ denotes a population state when the mutant share goes to zero. We have:

Lemma 4. *When the population state is* $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$*, if for all* $i \in \{1,2\}$, $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) >$

$\Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ *for all* $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ *then there exists an* $\bar{\varepsilon} > 0$ *such that for all* $i \in \{1,2\}$*:* $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ *in all Bayesian Nash equilibria* (x_1, x_2, x_τ) *in all states* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ *with* $\lambda_\tau \in (0, \bar{\varepsilon})$ *and* $|\lambda - \lambda^\circ| < \bar{\varepsilon}$*.*

Proof. In Appendix AppendixB.2. □

Lemma 5. *When the population state is* $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$*, if there exists* $i \in \{1,2\}$ *such that* $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) < \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ *with* $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ *a singleton, then is no an* $\bar{\varepsilon} > 0$ *such that for all* $i \in \{1,2\}$*:* $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ *in all Bayesian Nash equilibria* (x_1, x_2, x_τ) *in all states* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ *with* $\lambda_\tau \in (0, \bar{\varepsilon})$ *and* $|\lambda - \lambda^\circ| < \bar{\varepsilon}$*.*

Proof. In Appendix AppendixB.3. □

Lemmas 4 and 5 mean that it is generally sufficient to only study what is happening at the limit when the mutant share goes to zero when analyzing the evolutionary stability of a population. If the two residents earn the same type-fitness and a strictly greater playoff than any mutant $\theta_\tau \neq \theta_1, \theta_2$ in all Bayesian Nash equilibria in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, then the population $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ is evolutionarily stable. Else, the population is generally not evolutionarily stable.[14] Note that the proof of Lemma 5 actually develops a stronger argument than "not evolutionarily stable". If the residents earn the same type fitness in $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ and if the assumptions of Lemma 5 are satisfied, the proof shows that there exists an $\bar{\varepsilon} > 0$ such that the mutant earns a greater type fitness in all Bayesian Nash equilibria in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$. Alger and Weibull (2013) call this property *evolutionary unstability*.

2.5 *Homo moralis*

In the setting of a homogeneous population, Alger and Weibull (2013) show that the only *evolutionarily stable preference* is the one of *homo hamiltonensis*, a particular kind of *homo moralis*.

Definition 7 (Homo moralis and homo hamiltonensis)**.** An individual is a *homo moralis* if her utility function is of the form:

$$u_\kappa(x, y) = (1 - \kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x) \tag{4}$$

where $\kappa \in [0, 1]$ is her degree of morality.
A *homo moralis* maximizes a convex combination of her classical selfish payoff, with a weight $(1-\kappa)$, and of her "moral" payoff, defined as the payoff she would get if her opponent plays like her, with a weight κ. If $\kappa = 0$, then the individual is a *homo oeconomicus* (fully selfish). If $\kappa = 1$, then the

[14] There are two undetermined cases in this situation: (a) The two residents earn the same type-fitness but there exists a mutant θ_τ and a Bayesian Nash equilibria $(x_1^\circ, x_2^\circ, x_\tau^\circ)$ of the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ such that the residents and the mutant earn the same type-fitness: $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s) = \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s)$. (b) The two residents earn the same type-fitness but we are in the case of Lemma 5 except that $B^{NE}(s^\circ)$ is not a singleton and there also exists a Bayesian Nash equilibrium such that the residents earn a greater type fitness than the mutant.

individual is a *homo kantiensis* (fully moral). If the degree of morality κ is equal to the assortativity σ, then the individual is called *homo hamiltonensis*[15].

In our analysis, we often encounter *homo hamiltonensis*, and more precisely the strategies played by *homo hamiltonensis* individuals when all residents are of this type, called *Hamiltonian strategies*, play a key role in the analysis of evolutionary stability.

Definition 8 (Hamiltonian strategies)**.** $x_\sigma \in X$ is a *Hamiltonian strategies* if:

$$x_\sigma \in \underset{x \in X}{\operatorname{argmax}} \quad u_\sigma(x, x_\sigma)$$

For all $y \in X$, we call $\beta_\sigma(y) = \operatorname{argmax}_{x \in X} u_\sigma(x, y)$ the best-reply correspondence of *homo hamiltonensis* individuals, and we denote by $X_\sigma = \{x \in X : x \in \beta_\sigma(x)\}$ the set of fixed-points of *homo hamiltonensis*.

Consider a homogeneous population of *homo hamiltonensis* and a small group of mutants that wish to enter the population. If the mutants are not a "behavioral-alike"[16] to *homo hamiltonensis*, the mutants will always get a lower type fitness than *homo hamiltonensis*. Hence, if the mutant is a *homo moralis* with a degree of morality different from the assortativity ($\kappa \neq \sigma$), such that this *homo moralis* and *homo hamiltonensis* are not behaviorally-alike, then to enter the population, the degree of morality of the *homo moralis* should evolve in direction of the assortativity.

But is this homogeneity a required feature of evolutionary stability? What happens when the population is more diverse? We explore these questions in this paper, using as an illustration a population of *homo oeconomicus* and *homo kantiensis* involved in a prisoners' dilemma.

2.6 *Homo oeconomicus* and *homo kantiensis* in a prisoners' dilemma

A prisoners' dilemma is a finite symmetric fitness game with two pure strategies: cooperate (C) or defect (D). We note π^{ij} the payoff obtained when pure strategy i is played against pure strategy j. A prisoners' dilemma is well defined when $\pi^{CD} < \pi^{DD} < \pi^{CC} < \pi^{DC}$. In other words, players benefit if they both cooperate instead of defecting ($\pi^{DD} < \pi^{CC}$), but each of them has an incentive to deviate ($\pi^{CD} < \pi^{DD}$ and $\pi^{CC} < \pi^{DC}$). In our analysis, the sum S_π will play an important role:

$$S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC} \tag{5}$$

[15] Alger and Weibull (2013) named *homo hamiltonensis* in homage to the late biologist William Donald Hamilton. See Grafen (2004) for a biography.

[16] Types θ and τ are called behavioral-alike if they are behaviorally indistinguishable. Precisely, with θ being the resident, the set of of types τ that are behaviorally alike to θ is called Θ_θ:

$$\Theta_\theta = \{\tau \in \Theta : \exists\, x \in X_\theta \; s.t. \; (x, x) \in B^{NE}(\theta, \tau, 0)\}$$

Since $\pi^{CC} - \pi^{CD}$ is the gain minus the cost of cooperation and $\pi^{DC} - \pi^{DD}$ is the gain minus the cost of defection, S_π can be interpreted as the net benefit of cooperation minus the net benefit of defection. When $S_\pi = 0$ the game is sometimes called additive. Throughout this paper, we will use three examples of prisoners' dilemma: (a) $S_\pi < 0$, (b) $S_\pi = 0$ and (c) $S_\pi > 0$.

Table 1: Prisoner's dilemma examples

(a)

	C	D
C	$(4,4)$	$(0,6)$
D	$(6,0)$	$(1,1)$

S_π=-1<0

(b)

	C	D
C	$(4,4)$	$(0,5)$
D	$(5,0)$	$(1,1)$

S_π=0

(c)

	C	D
C	$(4,4)$	$(0,4.5)$
D	$(4.5,0)$	$(1,1)$

S_π=0.5>0

Let A be the matrix of the payoffs in the game, $A = [\pi^{CC}, \pi^{CD}; \pi^{DC}, \pi^{DD}]$. We allow players to use mixed strategies so that the strategy set X is the segment $\Delta = \{z \in \mathbb{R}^2_+ : z_1 + z_2 = 1\}$, where z_1 the probability to cooperate and z_2 the probability to defect. The payoff obtained by an individual playing strategy $x_1 \in X = \Delta$ when matched with an individual playing $x_2 \in X$ is then: $\pi(x_1, x_2) = x_1^\intercal A x_2$, where and $\pi : X^2 \to \mathbb{R}$ is a bilinear function. Since X is a segment, individuals' decision is fully characterized by their probability to cooperate. We will note $\alpha_i \in [0,1]$ the probability of an individual of type θ_i to cooperate. Hence, the payoff obtained by an individual θ_1 playing strategy $x_1 \in X$ when matched with an individual θ_2 playing $x_2 \in X$ is:

$$\pi(x_1, x_2) = \alpha_1\alpha_2\pi^{CC} + \alpha_1(1-\alpha_2)\pi^{CD} + (1-\alpha_1)\alpha_2\pi^{DC} + (1-\alpha_1)(1-\alpha_2)\pi^{DD}$$

Individuals *homo oeconomicus* are fully selfish, their morality coefficient is $\kappa = 0$ so that their utility is $u_0(x,y) = \pi(x,y)$. Hence, they always defect in a prisoner's dilemma because $\pi^{CD} < \pi^{DD}$ and $\pi^{CC} < \pi^{DC}$. Formally, for all $(x,y) \in X^2$ with $x = (\alpha_x; 1-\alpha_x)$, $\alpha_x \neq 0$ (i.e. x is not defection) and $y = (\alpha_y; 1-\alpha_y)$, we have:

$$\begin{aligned} u_0(D,y) - u_0(x,y) &= \left[\alpha_y\pi^{DC} + (1-\alpha_y)\pi^{DD}\right] \\ &\quad - \left[\alpha_x\alpha_y\pi^{CC} + \alpha_x(1-\alpha_y)\pi^{CD} + (1-\alpha_x)\alpha_y\pi^{DC} + (1-\alpha_x)(1-\alpha_y)\pi^{DD}\right] \\ &= \alpha_x\left[\alpha_y\left(\pi^{DC} - \pi^{CC}\right) + (1-\alpha_y)\left(\pi^{DD} - \pi^{CD}\right)\right] \\ &> 0 \end{aligned}$$

On the other hand, individuals *homo kantiensis* are fully moral, their morality coefficient is $\kappa = 1$ so that their utility is $u_1(x,y) = \pi(x,x)$. They always cooperate in a prisoner's dilemma because $\pi^{CC} > \pi^{DD}$. It is worth noting that the utility of a *homo kantiensis* individual does not depend on her opponent strategy but only on her own strategy.

Throughout this paper, we will mainly focus on a population of *homo oeconomicus* (θ_1) and *homo kantiensis* (θ_2) in the state $s = (\theta_1, \theta_2, \lambda)$, with $\lambda \in (0,1)$ the share of *homo kantiensis*. Consequently, the only Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \lambda)$ is $(x_1, x_2) = (D, C)$, or alternatively $(\alpha_1, \alpha_2) = (0, 1)$. Moreover, the share of *homo kantiensis* λ is also equal to the

cooperation share in the population.

3. On the coexistence of *homo oeconomicus* and *homo kantiensis*

The first condition of evolutionary stability requires that the residents earn the same type fitness in all Bayesian Nash equilibria in the state s (Definition 6). In this section, we explore when this condition is satisfied.

Let θ_1 be *homo oeconomicus*, θ_2 *homo kantiensis* and $\lambda^\circ \in (0,1)$ the share of *homo kantiensis*. The only Bayesian Nash equilibrium in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ is $(x_1, x_2) = (D, C)$ (see Section 2.6). Hence, *homo oeconomicus* and *homo kantiensis* earn the same type fitness if and only if:

$$\Pi_{\theta_1}(D, C, s^\circ) = \Pi_{\theta_2}(D, C, s^\circ) \tag{6}$$

Using Lemma 3 and noting $\phi_{12} \equiv \phi_{12}(\lambda^\circ, 0)$, we can write the type fitness of *homo oeconomicus* and *homo kantiensis* as a function of the share λ° and of the assortment between *homo oeconomicus* and *homo kantiensis* when there is no mutant in the population:

$$\begin{aligned} \Pi_{\theta_1}(D, C, s^\circ) &= [(1-\lambda^\circ) + \lambda^\circ \cdot \phi_{12}] \cdot \pi^{DD} + [\lambda^\circ(1-\phi_{12})] \cdot \pi^{DC} \\ \Pi_{\theta_2}(D, C, s^\circ) &= [(1-\lambda^\circ)(1-\phi_{12})] \cdot \pi^{CD} + [\lambda^\circ + (1-\lambda^\circ)\phi_{12}] \cdot \pi^{CC} \end{aligned} \tag{7}$$

Consequently, noting $\Pi_{\theta_{1-2}} \equiv \Pi_{\theta_1}(D, C, s^\circ) - \Pi_{\theta_2}(D, C, s^\circ)$ we have:

$$\begin{aligned} \Pi_{\theta_{1-2}} &= [(1-\lambda^\circ) + \lambda^\circ\phi_{12}] \cdot \pi^{DD} + [\lambda^\circ(1-\phi_{12})] \cdot \pi^{DC} \\ &\quad - [(1-\lambda^\circ)(1-\phi_{12})] \cdot \pi^{CD} - [\lambda^\circ + (1-\lambda^\circ)\phi_{12}] \cdot \pi^{CC} \\ &= \left[\pi^{DD} - \pi^{CD} - \phi_{12}\left(\pi^{CC} - \pi^{CD}\right)\right] - \lambda^\circ(1-\phi_{12})\left[\pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}\right] \end{aligned} \tag{8}$$

Similarly, rearranging the terms differently, we also have:

$$\Pi_{\theta_{1-2}} = (1-\lambda^\circ)(1-\phi_{12})\left[\pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}\right] - \left[\pi^{CC} - \pi^{DC} - \phi_{12}\left(\pi^{DD} - \pi^{DC}\right)\right] \tag{9}$$

We define: $Q_\pi \equiv \pi^{DD} - \pi^{CD} - \phi_{12}(\pi^{CC} - \pi^{CD})$ and $R_\pi \equiv \pi^{CC} - \pi^{DC} - \phi_{12}(\pi^{DD} - \pi^{DC})$. Note that we have: $Q_\pi + R_\pi = (1-\phi_{12})S_\pi$, with $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$. Rewriting the type-fitness equality (Equation 6) with Equations 8 and 9, we obtain two equivalent conditions, one for λ° and the other for $(1-\lambda^\circ)$:

$$\begin{aligned} \lambda^\circ(1-\phi_{12})S_\pi &= Q_\pi \\ (1-\lambda^\circ)(1-\phi_{12})S_\pi &= R_\pi \end{aligned}$$

We have the following proposition:

Proposition 2 (Type-fitness equality)**.** *In the population state* $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ *with* $\lambda^\circ \in (0,1)$*, homo oeconomicus* (θ_1) *and homo kantiensis* (θ_2) *earn the same type fitness if and only if:*

1. $S_\pi = 0$ *and* $Q_\pi = 0$*, i.e.* $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.
2. $S_\pi \neq 0$ *and* $\lambda^\circ = Q_\pi / \left[(1 - \phi_{12})\, S_\pi\right]$.

Moreover, if homo oeconomicus and homo kantiensis earn the same type fitness, then we must have $\phi_{12} \in (0,1)$.

Proof. In Appendix AppendixB.4. □

Proposition 2 characterizes the conditions under which *homo oeconomicus* and *homo kantiensis* can coexist in any prisoners' dilemma. In other words, the proposition provides information on the existence of a population of *homo oeconomicus* and *homo kantiensis* earning the same type fitness. If there exists $\lambda^\circ \in (0,1)$ such that $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$ when $S_\pi = 0$ or $\lambda^\circ = Q_\pi / \left[(1 - \phi_{12})\, S_\pi\right]$ when $S_\pi \neq 0$, then *homo oeconomicus* and *homo kantiensis* earn the same type fitness in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$.

Although our analysis is static, there is a link between Proposition 2 and the evolutionary game dynamics framework. In fact, at the equilibrium in a dynamic game, the two types should earn the same fitness. Thus, Proposition 2 allows to quickly identify the candidate population-state for an equilibrium in a dynamic game. The remaining question in this context is then whether or not this equilibrium can be reached. The answer depends not only on the replicator but also on the shape of the assortment function.

Finally, the last part of the Proposition stipulates that the assortment should be in a given range $\phi_{12} \in (0,1)$ to allow *homo oeconomicus* and *homo kantiensis* to earn the same type fitness. This range is detailed in the proof of the Proposition:

1. When $S_\pi < 0$: $(\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) < \phi_{12} < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$.
2. When $S_\pi = 0$: $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.
3. When $S_\pi > 0$: $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \phi_{12} < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.

We now discuss this result more in details in the case of a uniformly-constant assortment.

3.1 Coexistence under uniformly-constant assortment

We here consider the case of a uniformly-constant assortment (Definition 4), which is an extension of uniform random matching accounting for assortatively-matched interactions. Under uniformly-constant assortment, the assortment functions are constant and equal to the assortativity σ (Definition 5) by continuity: for all $\lambda \in (0,1)$, $\phi_{12}(\lambda, 0) = \sigma \in [0,1]$.

The following Corollary recaps the results of Proposition 2 under uniformly-constant assortment:

Corollary 1 (Type-fitness equality under uniformly-constant assortment)**.** *In the population state* $s = (\theta_1, \theta_2, \lambda^\circ)$ *with* $\lambda^\circ \in (0,1)$*, homo oeconomicus* (θ_1) *and homo kantiensis* (θ_2) *earn the same type*

fitness under uniformly-constant assortment if and only if:

1. *When* $S_\pi < 0$*:* $(\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) < \sigma < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$ *and* $\lambda^\circ = Q_\pi / [(1-\sigma)S_\pi]$.
2. *When* $S_\pi = 0$*:* $\sigma = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.
3. *When* $S_\pi > 0$*:* $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$ *and* $\lambda^\circ = Q_\pi / [(1-\sigma)S_\pi]$.

Proof. In Appendix AppendixB.4. □

There exists a population share $\lambda^\circ \in (0,1)$ such that *homo oeconomicus* and *homo kantiensis* earn the same type fitness if the assortativity is in a given range. This result is quite intuitive. In fact, if the assortment is too low then *homo oeconomicus* individuals earn a greater type fitness than *homo kantiensis* ones. For instance under uniform random matching (for all $\lambda \in (0,1)$, $\phi_{12} = \sigma = 0$), we have:

$$\Pi_{\theta_1}(D, C, s^\circ) = (1-\lambda^\circ)\pi^{DD} + \lambda^\circ \pi^{DC}$$
$$\Pi_{\theta_2}(D, C, s^\circ) = (1-\lambda^\circ)\pi^{CD} + \lambda^\circ \pi^{CC}$$

Since $\pi^{CD} < \pi^{DD}$ and $\pi^{CC} < \pi^{DC}$, $\Pi_{\theta_1}(D, C, s^\circ) > \Pi_{\theta_2}(D, C, s^\circ)$. Conversely, if the assortment is too high then *homo kantiensis* earns a greater type-fitness than *homo oeconomicus*. For instance, let $\sigma = 1$. This means that *homo oeconomicus* and *homo kantiensis* individuals only meet individuals of their own type. Thus, we have $\Pi_{\theta_1}(D, C, s^\circ) = \pi^{DD}$, and $\Pi_{\theta_2}(D, C, s^\circ) = \pi^{CC}$. Since $\pi^{CC} > \pi^{DD}$, $\Pi_{\theta_1}(D, C, s^\circ) < \Pi_{\theta_2}(D, C, s^\circ)$.

Note that when $S_\pi = 0$, i.e. when the the game is additive, there is a unique assortativity σ allowing *homo oeconomicus* and *homo kantiensis* to earn the same type-fitness. When the assortativity is below this threshold, *homo oeconomicus* dominates, while *homo kantiensis* dominates when the assortativity is above this threshold. This result is in line with the literature. For instance, Bergstrom (2003) and Allen and Nowak (2015) have studied the evolution of cooperative strategy in an evolutionary game dynamics framework, finding that assortment allows cooperation in prisoner's dilemma. Since at the equilibrium, strategies must earn the same fitness, their results are consistent with ours. In particular, in a simplified version of the game, where payoffs are additive ($\pi^{CD} = -c$, $\pi^{DD} = 0$, $\pi^{CC} = b - c$ and $\pi^{DC} = b$ with $b > c > 0$, $S_\pi = 0$) and the assortment constant, they highlight that cooperation is favored when a condition similar to the Hamilton's rule is satisfied.[17] We obtain an analogous condition in this simplified game: cooperation will outperform defection when $b\sigma > c$.

We now illustrate Corollary 1 with the examples defined in Section 2.6.

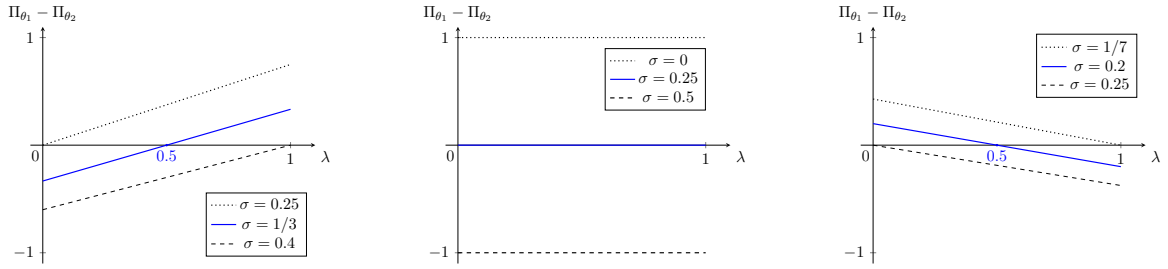
[17] Hamilton's rule stipulates that the frequency of an altruistic gene will increase if $br > c$, with b the reproductive gain for the recipient of the altruistic act, c the reproductive cost for the altruist individual, and r the genetic relatedness of the recipient to the actor (Hamilton, 1964b,a).

(a) First, let $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$ and $\pi^{DC} = 6$. We then have $S_\pi = -1 < 0$, $Q_\pi = 1 - 4\sigma$ and $R_\pi = -2 + 5\sigma$. Thus, there exists a heterogeneous population satisfying type-fitness equality when $0.25 < \sigma < 0.4$ (see Figure 1a). With $\sigma = 1/3$, then $\lambda^\circ = 0.5$ and *homo kantiensis* and *homo oeconomicus* co-exist and get the same type fitness equal to $\Pi_\theta = 8/3$. If the assortment is too low ($\sigma \leq 0.25$), only *homo oeconomicus* survives. In contrast, when the assortment is too high ($\sigma \geq 0.4$), *homo kantiensis* would dominate.

(b) Now let $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$ and $\pi^{DC} = 5$. We have $S_\pi = 0$, $Q_\pi = 1 - 4\sigma$ and $R_\pi = -1 + 4\sigma$. Thus, the only assortativity value consistent with type-fitness equality is $\sigma = 0.25$ (see Figure 1b). But then, for any population share $\lambda^\circ \in (0, 1)$, *homo kantiensis* and *homo oeconomicus* earn the same type-fitness.

(c) Finally, let $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$ and $\pi^{DC} = 4.5$. We have $S_\pi = 0.5 > 0$, $Q_\pi = 1 - 4\sigma$ and $R_\pi = -0.5 + 3.5\sigma$. Thus, there exists a heterogeneous population satisfying type-fitness equality when $1/7 < \sigma < 0.25$ (see Figure 1c). For example, when $\sigma = 0.2$, then $\lambda^\circ = 0.5$ and *homo kantiensis* and *homo oeconomicus* live together and get the same type-fitness equal to $\Pi = 2.4$. As above, the assortment plays a key role: if too low or too high, one type will dominate.

The assortativity allowing a heterogeneous population when $S_\pi = 0$ is $\sigma = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) = 0.25$. It is also the minimum assortativity for a heterogeneous population when $S_\pi < 0$ and the maximum assortativity for a heterogeneous population when $S_\pi > 0$. This comes as no surprise. Indeed, as discussed in Section 2.6, S_π can be interpreted as the net benefit of cooperation minus the net benefit of defection. Hence, when $S_\pi < 0$, defectors (*homo oeconomicus*) have an advantage and only high value of assortativity allows a heterogeneous population. Reciprocally, when $S_\pi > 0$, the game favors more cooperators (*homo kantiensis*) and lower value of assortativity is needed to get a heterogeneous population.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 1: Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under uniformly-constant assortment

3.2 Coexistence under state-dependent assortment

As highlighted in the literature, the phenomenon of homophily is highly dependent on the context. The size and demographic characteristics of the community considered affect the degree of homophily

among its members (McPherson et al., 2001; Currarini et al., 2009).[18] Therefore, going beyond the case of uniformly-constant[19] assortment, we pursue our analysis with the general case of a state-dependent assortment.

For this purpose, we define the function $\Pi_{\theta_{1-2}} : (0,1) \to \mathbb{R}$ as the type-fitness difference in prisoner's dilemma between *homo oeconomicus* and *homo kantiensis*. From Equation 8, we have for all $\lambda \in (0,1)$:

$$\Pi_{\theta_{1-2}}(\lambda) = Q_\pi(\lambda) - \lambda\left(1 - \phi_{12}(\lambda)\right) S_\pi$$

Where $\phi_{12}(\lambda) \equiv \phi_{12}(\lambda, 0)$ and $Q_\pi(\lambda) \equiv \pi^{DD} - \pi^{CD} - \phi_{12}(\lambda)\left(\pi^{CC} - \pi^{CD}\right)$. By assumption, the assortment function is continuous in λ. Moreover, the examples considered in this section converge when λ goes to zero (i.e. *homo kantiensis* is a mutant) and when λ goes to one (i.e. *homo oeconomicus* is a mutant). Thus, the function $\Pi_{\theta_{1-2}}$ can be extended by continuity to $[0,1]$.

Given the great number of cases offered by the relaxation of the uniformly-constant assortment hypothesis, we consider three specific cases to illustrate Proposition 2:

1. In the first case, we suppose that ϕ_{12} is linear: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$ (see Figure 2). Thus, when the share of *homo kantiensis* λ goes to zero, $\phi_{12}(0) = 0.32$. This means that when *homo kantiensis* is a mutant, the probability for a *homo kantiensis* individual to meet another *homo kantiensis* is $p_{22} = 0.32$ (see Lemma 3). Reciprocally, when the share of *homo kantiensis* λ goes to one, $\phi_{12}(1) = 0.08$ so that the probability for a *homo oeconomicus* individual to meet another *homo oeconomicus* is $p_{11} = 0.08$. Hence, the shape of $\phi_{12}(\cdot)$ increases the evolutionary-success opportunities of each type: a *homo oeconomicus* is better off when its probability to meet another *homo oeconomicus* is low, while a *homo kantiensis* is better off meeting another *homo kantiensis* with a high probability.
2. In the second case, we suppose that ϕ_{12} is a U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$ (see Figure 2). With this shape, there is a high assortment when the population is imbalanced (i.e. when one resident accounts for a high share of the population), and the assortment is lower when the population is more balanced. This could represent a population where individuals are living nearby each other when their share in the population is low (or in other words, mutants enter the population in a specific area) while individuals are more mixed when the population is more balanced.
3. In the third case, we suppose that ϕ_{12} is an inverse U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$ (see Figure 2). With this shape, the assortment is higher for a more balanced population. Bergstrom (2003) has shown that in a prisoners' dilemma involving cooperators

[18] Precisely, Currarini et al. (2009) find that the *homophily* in most US ethnic groups is nonlinear and non-monotonous in the group size and McPherson et al. (2001) shows that *homophily* depends on sociodemographic, behavioral, and intrapersonal characteristics.

[19] Recall that under uniformly-constant assortment, the assortative matching is uniform across all types in the population and independent of the shares in the population

and defectors, the assortment could have this shape when players have some choice about their partners. Moreover, such an assortment function is consistent with empirical evidences on the homophiliy in US ethnic groups (Currarini et al., 2009).

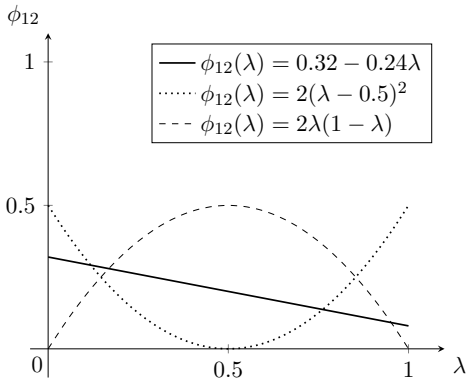


Figure 2: Illustrative state-dependent assortment functions between *homo oeconomicus* and *homo kantiensis*

For each case, we consider the same examples studied above and defined in Section 2.6: $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$, and (a) $\pi^{DC} = 6$, (b) $\pi^{DC} = 5$, (c) $\pi^{DC} = 4.5$. Thus, $Q_\pi(\lambda) = 1 - 4\phi_{12}(\lambda)$ and $S_\pi = 5 - \pi^{DC}$.

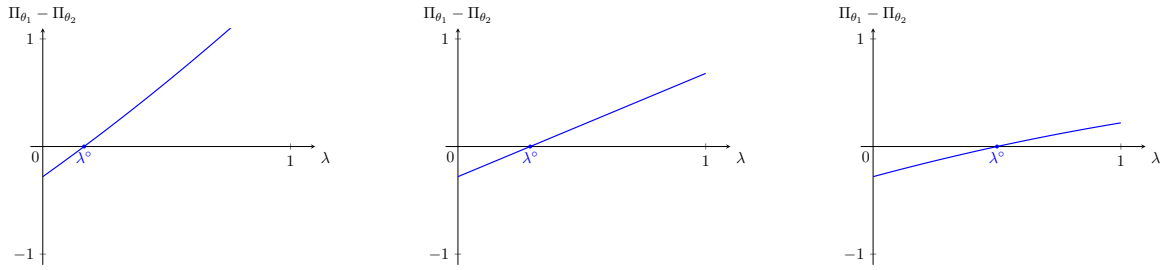
1. When $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$, $\Pi_{\theta_{1-2}}(\lambda) = -0.24S_\pi\lambda^2 + (0.96 - 0.68S_\pi)\lambda - 0.28$.
   (a) $S_\pi = -1 < 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has one root $\lambda^\circ \in (0,1)$: $\lambda^\circ = 1/6$ and then $\phi_{12}(\lambda^\circ) = 0.28$ (See Figure 3a).
   (b) $S_\pi = 0$, $\Pi_{\theta_{1-2}}$ is a line which intersects the x-axis for $\lambda^\circ = 7/24 \in (0,1)$, and then $\phi_{12}(\lambda^\circ) = 0.25$ (See Figure 3b).
   (c) $S_\pi = 0.5 > 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has one root $\lambda^\circ \in (0,1)$: $\lambda^\circ = 0.5$ and then $\phi_{12}(\lambda^\circ) = 0.2$ (See Figure 3c).

2. When $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$, $\Pi_{\theta_{1-2}}(\lambda) = 2S_\pi\lambda^3 - (2S_\pi + 8)\lambda^2 + (8 - 0.5S\pi)\lambda - 1$.
   (a) $S_\pi = -1 < 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.130$ and $\lambda^{\circ+} \approx 0.943$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.274$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.393$ (See Figure 4a).
   (b) $S_\pi = 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has two roots in $(0,1)$: $\lambda^{\circ-} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ+} = 0.5 + 0.25\sqrt{2}$, and then $\phi_{12}(\lambda^{\circ-}) = \phi_{12}(\lambda^{\circ+}) = 0.25$ (See Figure 4b).
   (c) $S_\pi = 0.5 > 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.157$ and $\lambda^{\circ+} \approx 790$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.235$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.168$ (See Figure 4c).

3. When $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$, $\Pi_{\theta_{1-2}}(\lambda) = -2S_\pi\lambda^3 + (2S_\pi + 8)\lambda^2 - (8 + S\pi)\lambda + 1$.
   (a) $S_\pi = -1 < 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.169$ and $\lambda^{\circ+} \approx 0.756$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.280$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.369$ (See Figure 5a).
   (b) $S_\pi = 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has two roots in $(0,1)$: $\lambda^{\circ-} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ+} = 0.5 + 0.25\sqrt{2}$, and then $\phi_{12}(\lambda^{\circ-}) = \phi_{12}(\lambda^{\circ+}) = 0.25$ (See Figure 5b). This case is actually symmetric to 2.(b) so that the equilibrium cooperation shares are the same.

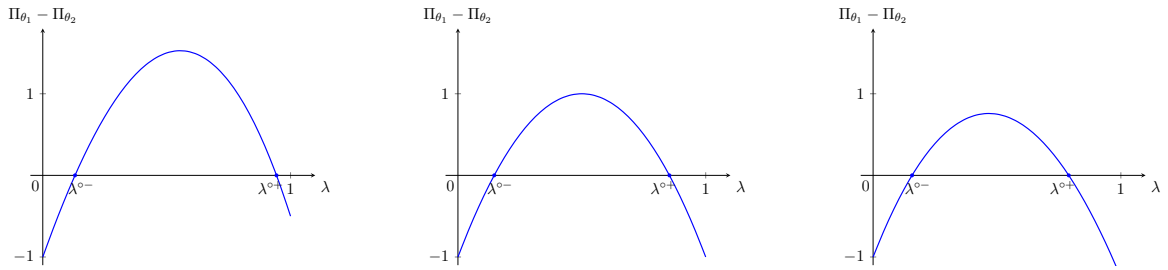
(c) $S_\pi = 0.5 > 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.137$ and $\lambda^{\circ+} \approx 0.917$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.237$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.153$ (See Figure 5c).

In each game, we find one cooperation share λ° allowing for a heterogeneous population with linear assortment (case 1) and two equilibrium cooperation shares with quadratic assortment (cases 2 and 3). However, this is not a general property of linear and quadratic assortment. The number of cooperation shares satisfying type-fitness equality depends on the game payoffs and on the assortment functions. Moreover, under linear assortment, note that the equilibrium cooperation share increases with S_π. Nonetheless, this is also not a general feature. For instance, with $\phi_{12}(\lambda) = 0.2\lambda + 0.2$, the equilibrium cooperation share decreases with S_π.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

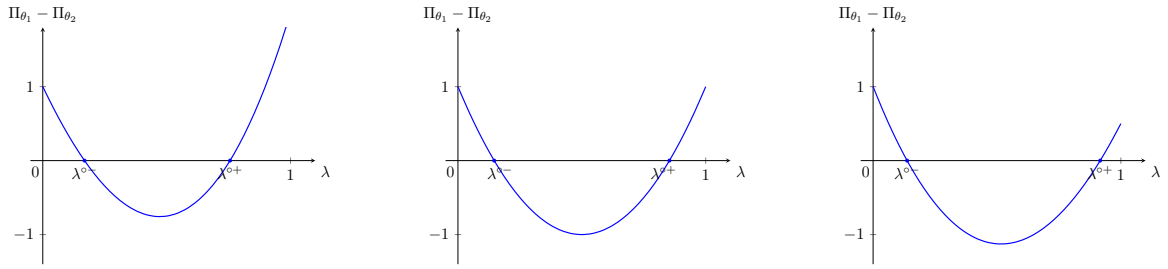
Figure 3: Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under state-dependent assortment $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 4: Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under state-dependent assortment $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$

State-dependent assortment brings more complexity but also more interesting equilibria. It will play a key role in the evolutionary-stability analysis (Section 4). Furthermore, the shape of the assortment function determines if an equilibrium cooperation share can be reached or not.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 5: Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under state-dependent assortment $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$

4. Is a population of *homo oeconomicus* and *homo kantiensis* favored by evolution?

An *evolutionarily stable population* satisfies two conditions: residents earn the same type fitness and they resist a small-scale invasion of any other type (Definition 6). In the previous section, we studied when the first condition is met for a population of *homo oeconomicus* (θ_1) and *homo kantiensis* (θ_2). We turn now our analysis to the second condition, assuming that the residents earn the same type fitness in the Bayesian Nash equilibrium $(x_1, x_2) = (D, C)$ in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$, with $\lambda^\circ \in (0, 1)$.

As shown in Lemmas 4 and 5, it is generally sufficient to only study what is happening at the limit when the mutant share goes to zero when analyzing the evolutionary stability of a population. Let $\theta_\tau \in \Theta$ be a mutant and (x_1, x_2, x_τ) a Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Note that since *homo oeconomicus* individuals always defect no matter their opponent strategy while *homo kantiensis* individuals always cooperate, we have $(x_1, x_2, x_\tau) = (D, C, x_\tau)$. Using Lemma 3 and noting $\pi_{ij} \equiv \pi(x_i, x_j)$ and $\Pi_{\theta_i} \equiv \Pi_{\theta_i}(x_1, x_2, x_\tau, s)$ for all $(i, j) \in I^2$, we can write the type fitness of each type:

$$\begin{cases} \Pi_{\theta_1} = (1 - \lambda^\circ + \lambda^\circ\phi_{12}) \cdot \pi_{11} + \lambda^\circ(1 - \phi_{12}) \cdot \pi_{12} \\ \Pi_{\theta_2} = (1 - \lambda^\circ)(1 - \phi_{12}) \cdot \pi_{21} + [\lambda + (1 - \lambda)\phi_{12}] \cdot \pi_{22} \\ \Pi_{\theta_\tau} = [(1 - \lambda^\circ)(1 - \sigma) - \lambda^\circ(1 - \lambda^\circ)\Gamma] \cdot \pi_{\tau 1} + [\lambda^\circ(1 - \sigma) + \lambda^\circ(1 - \lambda^\circ)\Gamma] \cdot \pi_{\tau 2} + \sigma \cdot \pi_{\tau\tau} \end{cases}$$

Note that $\pi_{1\tau}$ and $\pi_{2\tau}$ do not appear in the expression of the type fitness of *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) because at the limit when the mutant share goes to zero, the residents are matched between them as if there were no mutants in the population. Consequently, since by assumption *homo oeconomicus* and *homo kantiensis* earn the same type fitness in the Bayesian Nash equilibrium (D, C) in the population state s°, they also earn the same type fitness in all Bayesian Nash equilibria in the population state s, i.e. $\Pi_{\theta_1} = \Pi_{\theta_2} \equiv \Pi_\theta$.

Next, since we are in a two-strategies game, we can express the strategy x_τ in function of the strategies x_1 and x_2. For this purpose, recall that for all $i \in I$, $\alpha_i \in [0, 1]$ is the probability that θ_i individuals attach to cooperation, so that $x_i = (\alpha_i; 1 - \alpha_i)$. When $\alpha_1 \neq \alpha_2$, there exists $\gamma \in \mathbb{R}$ such

that $\alpha_\tau = (1-\gamma)\alpha_1 + \gamma\alpha_2$. The following Lemma depicts the difference in type-fitness between the residents and any mutant:

Lemma 6 (Difference in type fitness between residents and mutant)**.** *Let a population $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, with $\lambda^\circ \in (0,1)$, engaged in a prisoners' dilemma such that the residents earn the same type fitness Π_θ for $(x_1, x_2) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$ with $x^1 \neq x^2$. Then, the difference in type-fitness between the residents and the mutant for $(x_1, x_2, x_\tau) \in B^{NE}(s)$ is:*

$$\begin{aligned}\Pi_\theta - \Pi_{\theta_\tau} = &[\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12} - \sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma] \cdot (\alpha_2 - \alpha_1)^2 \cdot S_\pi \\ &+ [(\gamma - \lambda^\circ)(\phi_{12} - \sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma] \cdot (\alpha_2 - \alpha_1) \cdot [\alpha_2(\pi^{CC} - \pi^{CD}) + (1-\alpha_2)(\pi^{DC} - \pi^{DD})]\end{aligned}$$

Proof. In Appendix AppendixB.5. □

When studying the sign of this difference in type fitness, it is useful to understand what is the sign of $\gamma(1-\gamma)$. Without loss of generality and by symmetry we can assume $\alpha_1 < \alpha_2$, i.e. individuals θ_1 play the pure strategy 1 with a lower probability than individuals θ_2. We have $\alpha_\tau = (1-\gamma)\alpha_1 + \gamma\alpha_2$. If $\alpha_\tau \in (\alpha_1, \alpha_2)$, then $\gamma \in (0,1)$ and $\gamma(1-\gamma) > 0$. If $\alpha_\tau = \alpha_1$ or $\alpha_\tau = \alpha_2$, then $\gamma(1-\gamma) = 0$. Else $\gamma(1-\gamma) < 0$ (see Figure 6). In our study, since *homo oeconomicus* individuals defect and *homo kantiensis* individuals cooperate, we have $\alpha_1 = 0$ and $\alpha_2 = 1$. Because the residents play pure strategies, $\gamma = \alpha_\tau$. Thus, for all $\alpha_\tau \in (0,1)$, $\gamma(1-\gamma) > 0$.

Figure 6: Sign of $\gamma(1-\gamma)$ depending on the probabilities attached to the first pure strategy

We now have all the ingredients to examine the evolutionary stability of a population of *homo oeconomicus* and *homo kantiensis*. As in the coexistence analysis, we start with the case of uniformly-constant assortment before looking at the case of state-dependent assortment.

4.1 Evolutionary stability under uniformly-constant assortment

Under uniformly-constant assortment, we have $\phi_{12} = \sigma$ (by definition, see Remark 2) and $\Gamma = 0$. Indeed, $\Gamma = \lim_{\lambda_\tau \to 0}(\phi_{\tau 1} - \phi_{\tau 2})/\lambda_\tau$, and $\phi_{\tau 1} = \phi_{\tau 2} = \sigma$. As discussed in Section 2.3, Γ can be interpreted as the marginal matching-probability difference between mutants and residents of the two types. When individuals θ_1 and θ_2 meet the mutants at the same rate when they enter the population, then $\Gamma = 0$. We can rewrite Lemma 6 for the case of uniformly-constant assortment:

Corollary 2 (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, when θ_1 is homo oeconomicus, θ_2 is homo kantiensis and $\lambda^\circ \in (0,1)$, engaged in a prisoners' dilemma such*

that the residents earn the same type fitness Π_θ *for* $(D,C) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$*. Then, the difference in type-fitness between the residents and the mutant for* $(D,C,x_\tau) \in B^{NE}(s)$ *is:*

$$\Pi_\theta - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. In Appendix AppendixB.5. □

This expression is much simpler than the general case. The difference in type-fitness between the residents and the mutant depends only on the assortativity, on the mutant's strategy and on the net benefit of cooperation minus the net benefit of defection S_π. Moreover, from Corollary 1, we know that $\sigma > 0$ because we assumed that *homo oeconomicus* and *homo kantiensis* were earning the same type fitness. Note also that $\alpha_\tau(1-\alpha_\tau) \geq 0$ because $(\alpha_\tau) \in [0,1]$ and if mutants do not play a pure strategy, the inequality is strict, i.e. $\alpha_\tau(1-\alpha_\tau) > 0$. Hence, the sign of the difference in type-fitness depends only on the sign of S_π.

Interestingly, the same expression remains valid in a more general case, when the mutant share is not equal to zero:

Lemma 7 (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$*, when* θ_1 *is homo oeconomicus,* θ_2 *is homo kantiensis, engaged in a prisoners' dilemma. Then, we have for any* $(D,C,x_\tau) \in B^{NE}(s)$*:*

$$(1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau\Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. In Appendix AppendixB.6. □

Before stating our main results, we need to introduce two additional notions. First, the type set Θ is called *rich* if for each strategy $x \in X$, there exists some type $\theta \in \Theta$ for which this strategy is strictly dominant: $u_\theta(x,y) > u_\theta(x',y)$ for all $x' \neq x$ and y in X. When Θ is *rich*, for any strategy $x \in X$ it is always possible to find a mutant playing x. Second, we call Θ_{12} the set of mutants τ that are behaviorally indistinguishable from residents θ_1 and θ_2:

$$\Theta_{12} = \left\{\theta_\tau \in \Theta : \exists\, x \in X \text{ such that } (x_1,x,x) \text{ or } (x,x_2,x) \in B^{NE}(s)\right\}$$

In our study, the set Θ_{12} includes all the mutants that cooperate or defect when their share goes to zero, i.e. $\Theta_{12} = \left\{\theta_\tau \in \Theta : (D,C,D) \text{ or } (D,C,C) \in B^{NE}(s)\right\}$. We have the following Theorem:

Theorem 1 (Evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis*)**.** *In a prisoners' dilemma under uniformly-constant assortment when* Θ *is rich, there exists a heterogeneous evolutionarily stable population of homo oeconomicus and homo kantiensis against all types* $\theta_\tau \notin \Theta_{12}$ *if and only if* $S_\pi > 0$ *and* $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$*.*

Moreover, if there exists a heterogeneous evolutionarily stable population of homo oeconomicus and homo kantiensis, then it is unique and the cooperation share satisfies $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi)$.

Proof. In Appendix AppendixB.7. □

Theorem 1 fully characterizes the existence and uniqueness of *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* in prisoners' dilemma under uniformly-constant assortment. In particular, there does not exist a heterogeneous *evolutionarily stable population* when $S_\pi \leq 0$. When $S_\pi > 0$, there exists a unique heterogeneous *evolutionarily stable population* when the assortativity belongs to a range such that *homo oeconomicus* and *homo kantiensis* can coexist.

We made a few assumptions to derive the Theorem. First, we assumed that the type set Θ was rich. If it was not, then *homo oeconomicus* and *homo kantiensis* would be the only types in Θ. Then, any heterogeneous population satisfying type-fitness equality would be evolutionarily stable, even when $S_\pi \leq 0$ (because there does not exist any mutant). We could actually relax this assumption by assuming that there exists one type $\theta_\tau \in \Theta$ committed to a mixed strategy. In fact, any mixed strategy enables the mutant to earn a greater type fitness than at least one of the resident. Second, the population is evolutionarily stable against all types $\theta_\tau \notin \Theta_{12}$, i.e. the types which are not behaviorally-alike to the residents. Indeed, if mutants cooperate or defect, then the share of cooperation changes and the mutant cannot earn a strictly smaller type fitness in all Bayesian Nash equilibria in a neighborhood of λ°.

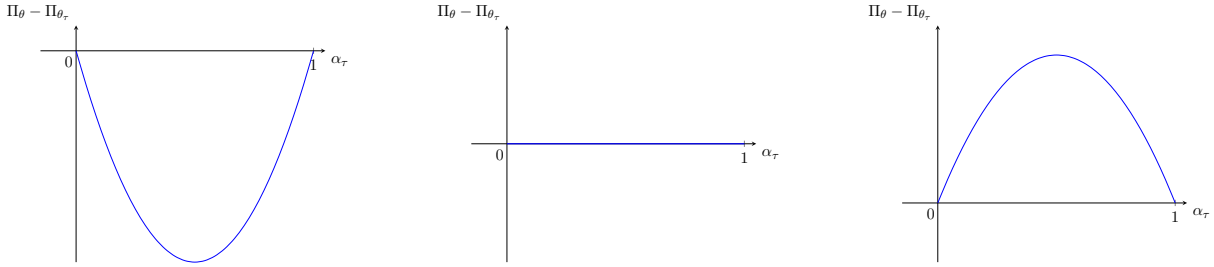
We now illustrate the Theorem going back to the examples defined in Section 2.6: $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$, and (a) $\pi^{DC} = 6$, (b) $\pi^{DC} = 5$, (c) $\pi^{DC} = 4.5$.

(a) First, $S_\pi < 0$. With a uniformly-constant assortment $\sigma = 1/3$, then with $\lambda^\circ = 0.5$ the population satisfies type-fitness equality and $\Pi_\theta = 8/3$ (see Section 3). However, we have $S_\pi = -1$, and since the difference in type fitness between the residents and the mutant at the limit is: $\Pi_\theta - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$ (Corollary 2), any mutant would earn more than the residents at the limit as illustrated in Figure 7a. Hence, we can conclude that the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable.

(b) Second, $S_\pi = 0$ and the game is additive. As discussed in Section 3, the only uniformly-constant assortment allowing type-fitness equality is $\sigma = 0.25$. With this value, for any $\lambda^\circ \in (0,1)$ *homo oeconomicus* and *homo kantiensis* earns the same type fitness. On the other hand, any mutant would also earn the same type-fitness at the limit (see Figure 7b). From Lemma 7, the mutant would also earn a greater type-fitness than at least one of the resident when its share λ_τ increases. Thus the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable

(c) Finally, $S_\pi > 0$. With a uniformly-constant assortment $\sigma = 0.2$, then with $\lambda^\circ = 0.5$ the population satisfies type-fitness equality and $\Pi_\theta = 2.4$ (see Section 3). Moreover, we have $S_\pi = 0.5$, and the difference in type fitness between the residents and the mutant at the limit is: $\Pi_\theta - \Pi_{\theta_\tau} = \sigma\gamma(1-\gamma)S_\pi$ (Corollary 2). Thus, for all $\alpha_\tau \in (0,1)$, $\Pi_\theta - \Pi_\tau > 0$ (see Figure 7c) and we can conclude

that the population of *homo oeconomicus* and *homo kantiensis* is evolutionarily stable against all mutants which do not cooperate or defect.



(a) Non evolutionarily-stable ($S_\pi < 0$, $\sigma = 1/3$, $\lambda^\circ = 0.5$)

(b) Non evolutionarily-stable ($S_\pi = 0$, $\sigma = 0.25$)

(c) Evolutionarily stable ($S_\pi > 0$, $\sigma = 0.2$, $\lambda^\circ = 0.5$)

Figure 7: Type-fitness difference between a heterogeneous population of *homo oeconomicus* and *homo kantiensis* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under uniformly-constant assortment.

There is a link between evolutionary stability in heterogeneous population and homogeneous population. The only *evolutionarily stable preference* in a homogeneous population is *homo hamiltonensis*, a *homo moralis* with degree of morality equal to the assortativity (Alger and Weibull, 2013). When *homo hamiltonensis* is the only resident, individuals of this type play *Hamiltonian strategies* $x_\sigma \in X_\sigma$ (see Definition 8). It turns out that under uniformly-constant assortment, the residents of a heterogeneous *evolutionarily stable population* should also play *Hamiltonian strategies*:

Proposition 3 (Non evolutionarily-stable population)**.** *In a symmetric* 2×2 *fitness game where the assortment matrix is uniformly constant and strictly positive, let* $s = (\theta_1, \theta_2, \lambda)$ *be a heterogeneous population.*
If there exists $(x^1, x^2) \in B^{NE}(s)$ *such that* $(x^1, x^2) \notin X_\sigma^2$ *and if* Θ *is rich, then the population is not evolutionarily stable.*

Proof. In AppendixB.8. □

Theorem 2 (Evolutionarily stable population)**.** *In a symmetric* 2×2 *fitness game where the assortment matrix is uniformly constant and strictly positive, let* $s = (\theta_1, \theta_2, \lambda)$ *be a heterogeneous population.*
If for all $(x^1, x^2) \in B^{NE}(s)$, $(x^1, x^2) \in X_\sigma^2$, *if* $\lambda = Q_\pi/((1-\sigma)S_\pi)$, *and if* $\beta_\sigma(x)$ *is a singleton for all* $x \in X_\sigma$, *then the population* $(\theta_1, \theta_2, \lambda)$ *is evolutionarily stable against all types* $\theta_\tau \notin \Theta_{12}$.

Proof. In AppendixB.9 □

.

4.2 Evolutionary stability under state-dependent assortment

Under state-dependent assortment when *homo oeconomicus* and *homo kantiensis* earn the same type fitness in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$, the difference in type fitness between the residents and any

mutant at the limit when the share of the mutant goes to zero is (Lemma 6):

$$\begin{aligned}\Pi_\theta - \Pi_{\theta_\tau} = [\alpha_\tau(1-\alpha_\tau)\sigma + (1-\alpha_\tau)\lambda^\circ(\phi_{12}-\sigma) + (1-\alpha_\tau)\lambda^\circ(1-\lambda^\circ)\Gamma] \cdot \cdot S_\pi \\ + [(\alpha_\tau - \lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma] \cdot (\pi^{CC} - \pi^{CD})\end{aligned}$$

We can first observe that if $\sigma = 1$ and if the mutants cooperate, they will always earn a greater type fitness than the residents. Indeed, in this setting, the mutants are matched between themselves earning $\Pi_{\theta_\tau} = \pi^{CC}$. On the other hand, the residents earn $\Pi_\theta < \pi^{CC}$. We can observe this by looking at the type fitness of *homo kantiensis*: *homo kantiensis* individuals earn π^{CC} when matched with another *homo kantiensis* but they earn $\pi^{CD} < \pi^{CC}$ when matched with a *homo oeconomicus*. Consequently, there is a maximum value of assortativity allowing for a heterogeneous *evolutionary stable population*:

Proposition 4 (Evolutionary stability under state-dependent assortment)**.** *In a prisoners' dilemma, if Θ is rich then there exists $\bar{\sigma} < 1$ such that there does not exist a heterogeneous evolutionary stable population of homo oeconomicus and homo kantiensis for all $\sigma > \bar{\sigma}$.*

Proof. In Appendix AppendixB.10. □

To illustrate Proposition 4, we focus on the same cases studied in Section 3:

1. We suppose that ϕ_{12} is linear: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$.
2. We suppose that ϕ_{12} is a U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$.
3. We suppose that ϕ_{12} is an inverse U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$.

Moreover, we assume that $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$. This shape allows $p_{1\tau}$ and $p_{2\tau}$ to belong to $[0,1]$. Moreover, it means that $p_{1\tau} = 1-\sigma$ and $p_{2\tau} = 0$, i.e. a mutant either meet a *homo oeconomicus* or another mutant, which increases the likelihood that the population of *homo oeconomicus* and *homo kantiensis* is evolutionarily stable.

For each case, we consider the same examples studied above and defined in Section 2.6: $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$, and (a) $\pi^{DC} = 6$, (b) $\pi^{DC} = 5$, (c) $\pi^{DC} = 4.5$.

1. When $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$:
   (a) $S_\pi =< 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^\circ = 1/6$. Moreover, for $\sigma < 0.4$, the residents earn a strictly greater type fitness than any mutant at the limit, and thus following the same arguments as in Theorem 1, the population is evolutionarily stable (see Figure 8a).
   (b) $S_\pi = 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^\circ = 7/24$. Moreover, for $\sigma < 0.46875$, the residents earn a strictly greater type fitness than any mutant (see Figure 8b). Hence, there exists a heterogeneous *evolutionarily stable population* when $\sigma < 0.46875$.

(c) $S_\pi > 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^\circ = 0.5$. They also earn a strictly greater type fitness than any mutant when $\sigma < 0.6$ and thus the population is evolutionarily stable (see Figure 8c).

2. When $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$:

   (a) $S_\pi < 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ-} \approx 0.130$ and $\lambda^{\circ+} \approx 0.943$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.368$ for $\lambda^{\circ-}$ and $\bar{\sigma} \approx 0.966$ for $\lambda^{\circ+}$ (see Figure 9a).

   (b) $S_\pi = 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness when $\lambda^{\circ-} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ+} = 0.5 + 0.25\sqrt{2}$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.360$ for $\lambda^{\circ-}$ and $\bar{\sigma} \approx 0.890$ for $\lambda^{\circ+}$ (see Figure 9b).

   (c) $S_\pi > 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ-} \approx 0.157$ and $\lambda^{\circ+} \approx 790$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.355$ for $\lambda^{\circ-}$ and $\bar{\sigma} \approx 0.825$ for $\lambda^{\circ+}$ (see Figure 9c).

3. When $\phi_{12}(\lambda) = 2\lambda(1 - \lambda)$:

   (a) $S_\pi < 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ-} \approx 0.169$ and $\lambda^{\circ+} \approx 0.756$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.402$ for $\lambda^{\circ-}$ and $\bar{\sigma} \approx 0.846$ for $\lambda^{\circ+}$ (see Figure 10a).

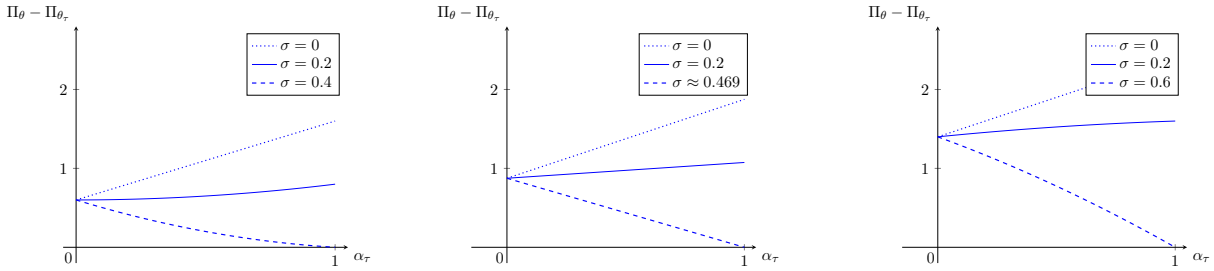
   (b) $S_\pi = 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ-} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ+} = 0.5 + 0.25\sqrt{2}$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.360$ for $\lambda^{\circ-}$ and $\bar{\sigma} \approx 0.890$ for $\lambda^{\circ+}$ (see Figure 10b).

   (c) $S_\pi > 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ-} \approx 0.137$ and $\lambda^{\circ+} \approx 0.917$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.342$ for $\lambda^{\circ-}$ and $\bar{\sigma} \approx 0.929$ for $\lambda^{\circ+}$ (see Figure 10c).

With our assumptions, we find *evolutionarily stable populations* of *homo oeconomicus* and *homo kantiensis* in all games. This contrasts with the case of a uniformly-constant assortment, in which there is no *evolutionarily stable population* when S_π is negative. Note also that under state-dependent assortment, the heterogeneous *evolutionarily stable population* can also resist to the invasion of behaviorally-alike mutants, i.e. mutants that cooperate or defect. This was not the case under uniformly-constant assortment.

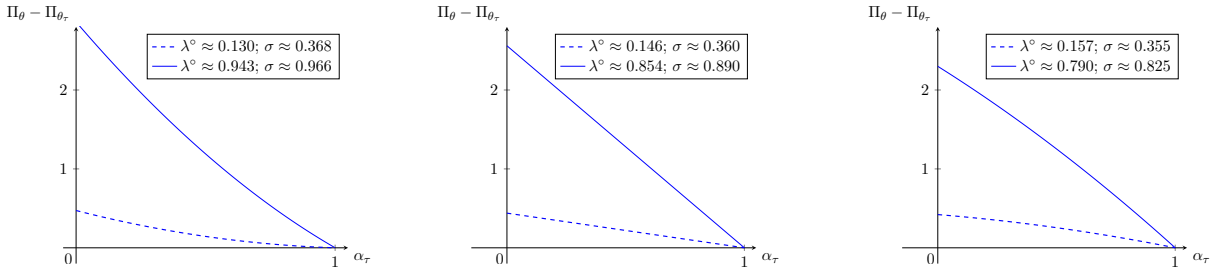
Both *homo oeconomicus* and *homo kantiensis* are important for the evolutionary success of the population, but in a different way. On the one hand, individuals *homo kantiensis* drive up the average fitness of the population since Π_θ increases with the share of *homo kantiensis*. As a result, there exists heterogeneous *evolutionarily stable population* for higher values of assortativity (see Figures 9 and 10). On the other hand, individuals *homo oeconomicus* help to resist the invasion of mutants. In fact, suppose that mutants are matched with *homo kantiensis* with greater probability than with *homo*

oeconomicus. For instance, let $\Gamma = (1-\sigma)/\lambda^\circ$. Then, in the case of $\sigma = 0$, the mutants would at least earn $\Pi_{\theta_\tau} = 4$ on average, so that they earn a greater type fitness than the residents. Hence, the population of *homo oeconomicus* and *homo kantiensis* would not be evolutionarily stable. The matching speed Γ governs which residents the mutants are more likely to meet. Thus, Γ plays a central role in the analysis of evolutionary stability.



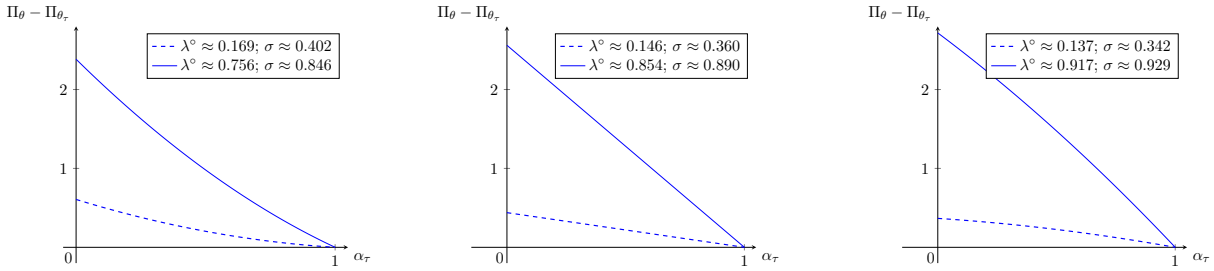
(a) $S_\pi < 0$, $\lambda^\circ = 1/6$ **(b)** $S_\pi = 0$, $\lambda^\circ = 7/24$ **(c)** $S_\pi > 0$, $\lambda^\circ = 0.5$

Figure 8: Difference in type fitness between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$ and $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$.



(a) $S_\pi < 0$, **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 9: Difference in type fitness between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$ and $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 10: Difference in type fitness between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$ and $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$.

5. Discussion

In this section, we first discuss the differences between a heterogeneous *evolutionarily stable population* and a population constituted by a single type of resident, *homo hamiltonensis*. Then,we discuss what determines the types of preferences favored by evolution in our framework.

5.1 Homogenous vs heterogeneous *evolutionarily stable population*

Homo hamiltonensis in a heterogeneous population

Expanding the framework of evolutionary stability formally established by Maynard Smith and Price (1973), Alger and Weibull (2013) proved the evolutionary stability of a particular type of preference, *homo hamiltonensis* in a single-type homogeneous population. As first expectation, we could have hypothesized that a heterogeneous *evolutionarily stable population* would "on average" have a *homo hamiltonensis* preference. In other words, an intuitively good candidate for a heterogeneous *evolutionarily stable population* would be a population composed by fully-selfish and fully-moral individuals with a share σ of fully moral individuals in order to "mimic" a *homo hamiltonensis* utility. However, such a population is not evolutionarily stable in most cases.[20] Instead, our results show that a heterogeneous *evolutionarily stable population* under uniformly-constant assortment depends on *Hamiltonian strategies*.

Since *homo hamiltonensis* individuals play *Hamiltonian strategies* when they are the only residents, one could ask if *homo hamiltonensis* can always be part of a heterogeneous *evolutionarily stable population*. The answer is no. In fact, consider an *evolutionarily stable population* of two types θ_1 and θ_2 committed to two different *Hamiltonian strategies* x_σ^1 and x_σ^2. In finite 2×2 fitness game under uniformly-constant and strictly positive assortment, this means that individuals of each type play the two pure strategies. Now suppose *homo hamiltonensis* replaces one of the residents, what happens then? Without loss of generality, let *homo hamiltonensis* replaces θ_1. In this setting, θ_2 individuals always play x_σ^2 while *homo hamiltonensis* individuals solve the following maximization problem:

$$x_h \in \underset{x \in X}{\operatorname{argmax}} \{p_{11}\left((1-\sigma)\pi(x, x_h) + \sigma\pi(x, x)\right) + p_{21}\left((1-\sigma)\pi(x, x_\sigma^2) + \sigma\pi(x, x)\right)\} \tag{10}$$

Since x_σ^2 is a *Hamiltonian strategy*, we have for all $x \in X : \pi(x_\sigma^2, x_\sigma^2) \geq (1-\sigma)\pi(x, x_\sigma^2) + \sigma\pi(x, x)$, and x_σ^2 is also a solution of the maximization problem (10). Consequently, (x_σ^2, x_σ^2) is a Bayesian Nash Equilibrium for the population of *homo hamiltonensis* and θ_2. But it is not the only one. Indeed, x_σ^1 is solution of (10) when:

$$p_{11}\left[\pi(x_\sigma^1, x_\sigma^1) - (1-\sigma)\pi(x_\sigma^2, x_\sigma^1) - \sigma\pi(x_\sigma^2, x_\sigma^2)\right] \geq p_{21}\left[\pi(x_\sigma^2, x_\sigma^2) - (1-\sigma)\pi(x_\sigma^1, x_\sigma^2) - \sigma\pi(x_\sigma^1, x_\sigma^1)\right]$$

[20] The only case in which this population is evolutionarily stable is when $\sigma = \lambda$ and σ is a solution of $\sigma = (\pi^{11} - \pi^{21} - \sigma(\pi^{22} - \pi^{21}))/((1-\sigma)(\pi^{11} + \pi^{22} - \pi^{12} - \pi^{21}))$.

Rewriting, with $p_{11} = 1 - \lambda + \lambda\sigma$ and $p_{21} = \lambda(1-\sigma)$:

$$(1 - \lambda + \lambda\sigma)\, Q_\pi \geq \lambda(1-\sigma)\, R_\pi$$

This inequality boils down to $\sigma \geq 0$ and is thus always respected.[21] Therefore, (x^1_σ, x^2_σ) is also a Bayesian Nash equilibrium for the population of *homo hamiltonensis* and θ_2. Hence, *homo hamiltonensis* individuals can play the two pure strategies x^1_σ and x^2_σ. Can they also play a mixed strategy? Let $x_h = (\alpha_h, 1-\alpha_h) = \alpha_h x^1_\sigma + (1-\alpha_h)x^2_\sigma$ a mixed strategy ($\alpha_h \in (0,1)$), x_h is solution of (10) when:

$$\begin{aligned} &p_{11}\left[(1-\sigma)(\alpha_h \pi(x^1_\sigma, x^1_\sigma) + (1-\alpha_h)\pi(x^1_\sigma, x^2_\sigma)) + \sigma\pi(x^1_\sigma, x^1_\sigma)\right] + p_{21}\left[(1-\sigma)\pi(x^1_\sigma, x^2_\sigma) + \sigma\pi(x^1_\sigma, x^1_\sigma)\right] \\ =&p_{11}\left[(1-\sigma)(\alpha_h \pi(x^2_\sigma, x^1_\sigma) + (1-\alpha_h)\pi(x^2_\sigma, x^2_\sigma)) + \sigma\pi(x^2_\sigma, x^2_\sigma)\right] + p_{21}\left[\pi(x^2_\sigma, x^2_\sigma)\right] \end{aligned}$$

Using $p_{11} = 1 - \lambda + \lambda\sigma$ and $R_\pi/((1-\sigma)S_\pi) = 1 - \lambda$, this equation can be rewritten as:

$$\alpha_h = \frac{1-\lambda}{1-\lambda+\lambda\sigma} \quad \in (0,1)$$

Consequently, when $x_h = (\frac{1-\lambda}{1-\lambda+\lambda\sigma}, \frac{\lambda\sigma}{1-\lambda+\lambda\sigma})$, (x_h, x^2_σ) is also a Bayesian Nash equilibrium for the population of *homo hamiltonensis* and θ_2. Since the definition of evolutionary stability encompasses all Bayesian Nash equilibria, this means that the population of *homo hamiltonensis* and θ_2 is not evolutionarily stable.[22] In other words, *homo hamiltonensis* individuals cannot be part of a heterogeneous *evolutionarily stable population* playing diverse strategies.

Equilibrium implications

In the classical setting of a homogeneous, single-type resident population, all resident individuals in the population play the same strategy. We show that this characteristic is not necessary for evolutionary stability by proving the existence of a heterogeneous population exhibiting diverse strategies played by resident individuals without infringing the evolutionary stability (Theorem 1). For example, in the prisoner's dilemma, the classical setting suggests that, when no mixed *Hamiltonian strategy* exists (i.e. when $\pi_{CC} + \pi_{DD} - \pi_{DC} - \pi_{CD} > 0$), all *homo hamiltonensis* individuals either cooperate or defect, i.e. they all behave as a *homo oeconomicus* and defect, or they all behave as a *homo kantiensis* and cooperate. Yet, Theorem 1 establishes the existence of a heterogeneous *evolutionarily stable population* with a share of defectors *homo oeconomicus* and of cooperators *homo kantiensis*.

This last result is more in line with empirical observations. In single trial public goods experiments for instance, results display between 40% and 60% contribution to the public good (Marwell and

[21] Because $Q_\pi = \lambda(1-\sigma)S_\pi$ and $R_\pi = (1-\lambda)(1-\sigma)S_\pi$.

[22] Proposition 3 insures that only *Hamiltonian strategies* can be candidates for evolutionary stability, i.e. only the two pure strategies in this context.

Ames, 1981; Dawes and Thaler, 1988). A population of *homo hamiltonensis* all playing a mixed strategy in prisoner's dilemma could support this empirical observation when $S_\pi < 0$ but not when $S_\pi > 0$ (Lemma AppendixB.5). In the latter case, only a heterogeneous population would justify the observations.

Finally, the introduction of assortative matching between preferences has a key implication when studying and interpreting equilibria in games. In his thesis, John Nash discussed two interpretations of a mixed Nash equilibrium (Nash, 1950, 1951). In the first interpretation, an individual randomizes his play before acting, for instance by throwing a dice or a coin. In the second, called "mass-action", individuals of a large population play one of the pure strategies composing the mixed equilibrium with the share of people playing each strategy being equal to the weight of the strategy in the equilibrium.[23] Similarly, in the original and static evolutionary game theory framework (Maynard Smith, 1974), a mixed *evolutionarily stable strategy* can either describe a "monomorphic" population of identical individuals randomizing their behavior, or a heterogeneous population (called "polymorphic" in biology) of several types of individuals, each type playing a pure strategy. Under uniform random matching, the two interpretations are equivalent. Thus, the static framework could not distinguish between a monomorphic and a polymorphic population, which led to the emergence of the evolutionary game dynamics framework (Bergstrom and Godfrey-Smith, 1998). However, when the matching is assortative, a monomorphic and a polymorphic population would not yield the same equilibrium, as already observed by Grafen (1979) and Hines and Maynard Smith (1979). In other words, the first and second interpretation of a mixed equilibrium are no longer equivalent when a distinct preference is associated to each strategy.

5.2 Context-based preferences

Game-dependent diversity

A key property in the case of a homogeneous single-type resident population is the evolutionary stability of the *homo hamiltonensis* type of preference regardless of the game being played. In other words, as long as the assortativity is set and constant, in any game between assortatively matched individuals, only those behaviorally alike to *homo hamiltonensis* will resist mutant invasion. In this paper, proving the evolutionary stability of other types of preferences when allowing for the presence of more than one resident type in the population, we show that this evolutionary stability depends on the game being played. Specifically, we find that both the assortment properties and the game payoffs determine whether a heterogeneous population is evolutionarily stable. In a prisoner's dilemma for instance, under uniformly-constant assortment, the evolutionary stability of a population of *homo oeconomicus* and *homo kantiensis* individuals depends on the sign of S_π and the value of assortativity σ (Theorem 1). Hence, the prevailing preferences in a population depend on the context. This finding is in line with earlier research stating that the economic environment determines the prevalence of self-interested or altruistic behaviors (Bester and Güth, 1998) and of self-interested or fair

[23] See also Leonard (1994) and Weibull (1994) for a discussion of the mass-action interpretation of Nash equilibria.

behaviors (Fehr and Schmidt, 1999). Empirical evidence also suggests that choices and preferences can change according to the context (Tversky and Simonson, 1993; Rieskamp et al., 2006; Masatlioglu et al., 2012; Bordalo et al., 2013). As examples, economic crises modify the attitude toward risks (Schildberg-Hörisch, 2018) and the social, economic and institutional settings affect cooperative behaviors (Shogren and Taylor, 2008). In our framework, a socio-economic shock would translate into a change in the payoffs and in the homophily (i.e. the assortment), which would, in turn, affect the prevailing preferences in the population.

This dependence on the context has significant implications for empirical testing. Since the game and the context affect the behavior of agents, experiments should give particular attention to the conditions under which experiments are performed (statement of payoffs, cost of actions, available options, ties between subjects, etc.). While empirical behavioral research often aims at finding the parameters of the preferences of individuals, it would be an interesting challenge to try to estimate how diverse a population is. Considering a distribution of *homo moralis* with different morality coefficients, what is the shape of this distribution? The framework developed in this paper could be tested in lab experiments. For instance, in the case of the prisoner's dilemma, does our simplified model explain the share of individuals cooperating? Is there assortment between individuals with similar preferences, and if so, what is the shape of assortment functions in different contexts and cultures? In all these experiments, the choice of payoffs in the game is central, since different payoffs lead to different evolutionary stability profiles.

Unobserved diversity of preferences

In Theorem 1, we have detailed the conditions under which a population of selfish *homo oeconomicus* and fully-moral *homo kantiensis* can be evolutionarily stable in a prisoner's dilemma under uniformly-constant assortment. This result can be extended to the behaviorally-alike of *homo oeconomicus* and *homo kantiensis*. In particular, individuals caring only for the payoff of others such as fully-altruistic or fully-empathetic individuals would always cooperate in a prisoner's dilemma.[24] Thus, they can be part of a heterogeneous *evolutionarily stable population* with *homo oeconomicus* individuals.

Is it more likely to find moral or altruist individuals in a population? Our framework provides a theoretical-justification to the observed diversity of behaviors and preferences but cannot answer this question. Thus, it would be interesting to empirically test which social preferences explain individuals' choices better. For instance, Miettinen et al. (2017) have recently shown that *homo moralis* has a higher explanatory power than altruistic preferences in a sequential prisoners' dilemma. However, scientists can only observe the strategies chosen by individuals and not their true preferences. As discussed above, these strategies are context-dependent. Hence, further investigation varying the games and the context of the experiment would help identify individual preferences with greater

[24]The utility of fully-altruistic and fully-empathetic individuals is $u(x, y) = \pi(y, x)$. See also Alger and Weibull (2017) for a discussion of the strategic behaviors of moralists and altruists.

precision and better understand the individual motives behind the observed decisions.

6. Conclusion

In many countries and contexts we observe individuals exhibiting a wide heterogeneity of preferences in labor market choice, saving decisions and prosocial behavior (Falk et al., 2018). Following this empirical observation, we extend the classical framework of evolutionary stability of preferences by allowing heterogeneity in individual preferences in the context of assortative interaction with imperfect information. We generalize the concept of assortment function to define an assortment matrix modeling homophily between the different types of preferences in a population. In the case of uniformly-constant assortment, we prove that a heterogeneous *evolutionarily stable population* composed of two types always exists: individuals of this population earn the same payoff and resist a small-scale invasion of mutants. Moreover, we find that the two types should play *Hamiltonian strategies*, the strategies played by a certain *homo moralis* when this type is the only one in the population. Finally, we show how and when a heterogeneous population made of fully-selfish individuals, *homo oeconomicus*, and fully-moral ones, *homo kantiensis*, is evolutionarily stable in prisoner's dilemmas.

In a heterogeneous environment, individuals do not necessarily play the same strategy. Thus, our work helps in understanding the driving forces behind strategic behavior such as cooperation and defection in social dilemma or the diverse contribution to public goods. We believe that an in-depth investigation of the observed variability of behaviors among agents when voting, performing environmentally friendly actions or donating money is necessary. Hence, further work on the implications of accounting for the diversity of preferences in a population would bring valuable insights for policy makers and allow a better crafting of public policies.

More generally, this paper intends to give a theoretical framework pushing the development of analyses accounting for a diversity of preferences under assortative matching. Many extensions and improvements can be undertaken to deepen the understanding of heterogeneous populations. First, further exploring the case of non uniformly-constant assortment, of which we analyzed three different cases in Section 3.2 and Section 4.2, is key to better comprehend the role assortment plays in allowing for the diversity of preferences. Then, it would be interesting to study how to define assortment in the case of a distribution of preferences in order to reconcile our framework with the one of Dekel et al. (2007). The assortment could also be rendered endogenous by including informational and strategic features into the game. Second, the analysis of a heterogeneous *evolutionarily stable population* could be extended to finite games with more than two pure strategies and more than two resident types, and to infinite games. Would *Hamiltonian strategies* still be favored under uniformly-constant assortment? Finally, in our analysis, we favored a static framework because we investigated under which conditions a heterogeneous population is evolutionarily stable to the invasion of a mutant preference. It would be helpful to analyze how the preferences in a heterogeneous population evolve under assortative matching using an evolutionary game dynamics framework. We expect that some equilibria we found in the static case could not be reached in a dynamic setting, depending on the evolutionary process,

i.e. the replicator.

This paper aims at opening the way towards better consideration of the diversity of preferences and of assortative matching, moving away from the more classical use of representative agents and homogeneous populations in future theoretical and empirical studies.

References

Ingela Alger and Jörgen W Weibull. A generalization of Hamilton's rule — Love others how much? *Journal of Theoretical Biology*, 299:42–54, 2012.

Ingela Alger and Jörgen W Weibull. Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302, 2013.

Ingela Alger and Jörgen W Weibull. Evolution and Kantian morality. *Games and Economic Behavior*, 98:56–67, 2016.

Ingela Alger and Jörgen W Weibull. Strategic behavior of moralists and altruists. *Games*, 8(3):38, 2017.

Benjamin Allen and Martin A Nowak. Games among relatives revisited. *Journal of Theoretical Biology*, 378:103–116, 2015.

James Andreoni. Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477, 1990.

James Andreoni and John Miller. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753, 2002.

Gary S Becker. A theory of marriage: Part I. *Journal of Political Economy*, 81(4):813–846, 1973.

Gary S Becker. A theory of marriage: Part II. *Journal of Political Economy*, 82(2, Part 2):S11–S26, 1974a.

Gary S Becker. A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093, 1974b.

Carl T Bergstrom and Peter Godfrey-Smith. On the evolution of behavioral heterogeneity in individuals and populations. *Biology and Philosophy*, 13(2):205–231, 1998.

Theodore C Bergstrom. On the evolution of altruistic ethical rules for siblings. *The American Economic Review*, 85(1):58–81, 1995.

Theodore C Bergstrom. The algebra of assortative encounters and the evolution of cooperation. *International Game Theory Review*, 5(03):211–228, 2003.

Helmut Bester and Werner Güth. Is altruism evolutionarily stable? *Journal of Economic Behavior & Organization*, 34(2):193–209, 1998.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013.

Kjell Arne Brekke, Snorre Kverndokk, and Karine Nyborg. An economic model of moral motivation. *Journal of Public Economics*, 87(9-10):1967–1983, 2003.

Stephen V Burks, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini. Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences*, 106(19):7745–7750, 2009.

Donn Erwin Byrne. *The attraction paradigm*, volume 11. Academic Press, 1971.

Sergio Currarini, Matthew O Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009.

Robyn M Dawes and Richard H Thaler. Anomalies: cooperation. *Journal of Economic Perspectives*, 2(3):187–197, 1988.

Eddie Dekel, Jeffrey C Ely, and Okan Yilankaya. Evolution of preferences. *The Review of Economic Studies*, 74(3):685–704, 2007.

Tore Ellingsen. The evolution of bargaining behavior. *The Quarterly Journal of Economics*, 112(2): 581–602, 1997.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692, 2018.

Ernst Fehr and Simon Gächter. Reciprocity and economics: The economic implications of *Homo Reciprocans*. *European Economic Review*, 42(3-5):845–859, 1998.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

Alan Grafen. The hawk-dove game played between relatives. *Animal Behaviour*, 27:905–907, 1979.

Alan Grafen. William Donald Hamilton. 1 august 1936—7 march 2000, 2004.

Anna Gunnthorsdottir, Roumen Vragov, Stefan Seifert, and Kevin McCabe. Near-efficient equilibria in contribution-based competitive grouping. *Journal of Public Economics*, 94(11-12):987–994, 2010.

Werner Güth and Menahem Yaari. An evolutionary approach to explain reciprocal behavior in a simple strategic game. *U. Witt. Explaining Process and Change–Approaches to Evolutionary Economics. Ann Arbor*, pages 23–34, 1992.

William D Hamilton. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1):17–52, 1964a.

William D Hamilton. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16, 1964b.

Aviad Heifetz, Chris Shannon, and Yossi Spiegel. The dynamic evolution of preferences. *Economic Theory*, 32(2):251–286, 2007.

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, 91(2):73–78, 2001.

William Gord S Hines and John Maynard Smith. Games between relatives. *Journal of Theoretical Biology*, 79(1):19–30, 1979.

Herminia Ibarra. Personal networks of women and minorities in management: A conceptual framework. *Academy of management Review*, 18(1):56–87, 1993.

Ryota Iijima and Yuichiro Kamada. Social distance and network structures. *Theoretical Economics*, 12(2):655–689, 2017.

Matthew O Jackson and Alison Watts. Social games: Matching and the play of finitely repeated games. *Games and Economic Behavior*, 70(1):170–191, 2010.

Immanuel Kant. *Grundlegung zur metaphysik der sitten*, volume 28. L. Heimann, 1870.

Jean-Jacques Laffont. Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. *Economica*, 42(168):430–437, 1975.

Jessica L Lakin and Tanya L Chartrand. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14(4):334–339, 2003.

Robert J Leonard. Reading Cournot, reading Nash: The creation and stabilisation of the Nash equilibrium. *The Economic Journal*, pages 492–511, 1994.

David K Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622, 1998.

Erez Lieberman, Christoph Hauert, and Martin A Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312, 2005.

Gerald Marwell and Ruth E Ames. Economists free ride, does anyone else?: Experiments on the provision of public goods, IV. *Journal of Public Economics*, 15(3):295–310, 1981.

Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y Ozbay. Revealed attention. *American Economic Review*, 102(5):2183–2205, 2012.

John Maynard Smith. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, 47(1):209–221, 1974.

John Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

Topi Miettinen, Michael Kosfeld, Ernst Fehr, and Jorgen W Weibull. Revealed preferences in a sequential prisoners' dilemma: A horse-race between five utility functions. 2017.

John Nash. *Non-cooperative games*. PhD thesis, Princeton, 1950.

John Nash. Non-cooperative games. *Annals of Mathematics*, pages 286–295, 1951.

Bryan Norton, Robert Costanza, and Richard C Bishop. The evolution of preferences: why sovereign' preferences may not lead to sustainable policies and what to do about it. *Ecological Economics*, 24 (2-3):193–211, 1998.

Martin A Nowak, Corina E Tarnita, and Tibor Antal. Evolutionary dynamics in structured populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):19–30, 2010.

Peter Ockenfels. Cooperation in prisoners' dilemma: An evolutionary approach. *European Journal of Political Economy*, 9(4):567–579, 1993.

Hisashi Ohtsuki and Martin A Nowak. Evolutionary stability on graphs. *Journal of Theoretical Biology*, 251(4):698–707, 2008.

Efe A Ok and Fernando Vega-Redondo. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, 97(2):231–254, 2001.

Thomas Piketty. Social mobility and redistributive politics. *The Quarterly Journal of Economics*, 110(3):551–584, 1995.

Alex Possajennikov. On the evolutionary stability of altruistic and spiteful preferences. *Journal of Economic Behavior & Organization*, 42(1):125–129, 2000.

Matthew Rabin. Incorporating fairness into game theory and economics. *The American Economic Review*, pages 1281–1302, 1993.

Jörg Rieskamp, Jerome R Busemeyer, and Barbara A Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3):631–661, 2006.

Arthur J Robson. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, 144(3):379–396, 1990.

Catherine Salmon and Margo Wilson. Kinship: The conceptual hole in psychological studies of social cognition and close relationships. *Evolutionary Social Psychology*, page 265, 2013.

Hannah Schildberg-Hörisch. Are Risk Preferences Stable? *Journal of Economic Perspectives*, 32(2): 135–54, 2018.

Bodo B Schlegelmilch, Greg M Bohlen, and Adamantios Diamantopoulos. The link between green purchasing decisions and measures of environmental consciousness. *European Journal of Marketing*, 30(5):35–55, 1996.

Rajiv Sethi and Eswaran Somanathan. Preference evolution and reciprocity. *Journal of Economic Theory*, 97(2):273–297, 2001.

Paulo Shakarian, Patrick Roos, and Anthony Johnson. A review of evolutionary graph theory with applications to game theory. *Biosystems*, 107(2):66–80, 2012.

Jason F Shogren and Laura O Taylor. On behavioral-environmental economics. *Review of Environmental Economics and Policy*, 2(1):26–44, 2008.

Adam Smith. *The Theory of Moral Sentiments: By Adam Smith.* A. Millar; and A. Kincaid and J. Bell, in Edinburgh, 1759.

Oded Stark and Ita Falk. Transfers, empathy formation, and reverse transfers. *The American Economic Review*, 88(2):271–276, 1998.

Corina E Tarnita, Tibor Antal, Hisashi Ohtsuki, and Martin A Nowak. Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences*, 106(21):8601–8604, 2009.

Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management science*, 39(10): 1179–1189, 1993.

Eric Van Damme. Evolutionary game theory. In *Stability and Perfection of Nash Equilibria*, pages 214–258. Springer, 1991.

Edwin JC Van Leeuwen, Katherine A Cronin, Daniel BM Haun, Roger Mundry, and Mark D Bodamer. Neighbouring chimpanzee communities show different preferences in social grooming behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20121543, 2012.

Jörgen W Weibull. The mass-action interpretation of nash equilibrium. Technical report, IUI Working Paper, 1994.

Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645):330–338, 1922.

AppendixA. The algebra of assortative matching: Proofs

In this section, we provide the proofs of properties, lemmas and proposition of Section 2.3 on assortative encounters.

We are in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ (equivalently $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \lambda_\tau)$). Let $I = \{1, 2, \tau\}$, the assortment matrix is $\Phi = ((\phi_{ij}(\lambda, \lambda_\tau)))_{(i,j)\in I^2}$ such that for all $(i, j) \in I^2$, $\phi_{ij}(\lambda, \lambda_\tau) = p_{i|i} - p_{i|j}$ (Definition 3). To be well defined, the matching process must satisfy two sets of conditions:

- The matching conditions: for all $i \in I$, $\sum_{j\in I} p_{j|i} = 1$ (Property 2)
- The balancing conditions: for all $(i, j) \in I^2$, $\lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i}$ (Property 3)

AppendixA.1 Proof of Property 4

Property (Assortment balancing condition)**.** The assortment matrix satisfies the *assortment balancing conditions* when:

$$\forall\,(i,j) \in I^2 : \quad \lambda_j \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{ik}\right) - \phi_{ij}\right] = \lambda_i \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{jk}\right) - \phi_{ji}\right]$$

If the matching process satisfies the matching and balancing conditions, then the assortment matrix must satisfy the assortment balancing conditions.

Proof.

$$\begin{aligned}
& \lambda_j \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{ik}\right) - \phi_{ij}\right] - \lambda_i \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{jk}\right) - \phi_{ji}\right] \\
\underset{\text{(Def.3)}}{=} & \sum_{k\in I} \lambda_j \lambda_k p_{i|i} - \sum_{k\in I} \lambda_j \lambda_k p_{i|k} - \lambda_j p_{i|i} + \lambda_j p_{i|j} - \sum_{k\in I} \lambda_i \lambda_k p_{j|j} + \sum_{k\in I} \lambda_i \lambda_k p_{j|k} + \lambda_i p_{j|j} - \lambda_i p_{j|i} \\
\underset{\text{(Prop.3)}}{=} & \lambda_j p_{i|i} - \sum_{k\in I} \lambda_j \lambda_i p_{k|i} - \lambda_j p_{i|i} + \lambda_i p_{j|i} - \lambda_i p_{j|j} + \sum_{k\in I} \lambda_i \lambda_j p_{k|j} + \lambda_i p_{j|j} - \lambda_i p_{j|i} \\
= & \lambda_i \lambda_j \left[\sum_{k\in I} p_{k|j} - \sum_{k\in I} p_{k|i}\right] \\
\underset{\text{(Prop.2)}}{=} & 0
\end{aligned}$$

□

AppendixA.2 Proof of Lemma 2

Lemma (Assortment between residents)**.** *When $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, if the matching process satisfies the matching and balancing conditions, then we have $\phi_{12}(\lambda, 0) = \phi_{21}(\lambda, 0)$.*

Proof. If the matching process satisfies the matching and balancing conditions, then the assortment matrix must satisfy the assortment balancing conditions (Property 4). The assortment balancing conditions are:

$$\begin{aligned}
\lambda_2\left(\lambda_2\phi_{12}+\lambda_\tau\phi_{1\tau}-\phi_{12}\right) &= \lambda_1\left(\lambda_1\phi_{21}+\lambda_\tau\phi_{2\tau}-\phi_{21}\right)\\
\lambda_\tau\left(\lambda_2\phi_{12}+\lambda_\tau\phi_{1\tau}-\phi_{1\tau}\right) &= \lambda_1\left(\lambda_1\phi_{\tau1}+\lambda_2\phi_{\tau2}-\phi_{\tau1}\right)\\
\lambda_\tau\left(\lambda_1\phi_{21}+\lambda_\tau\phi_{2\tau}-\phi_{2\tau}\right) &= \lambda_2\left(\lambda_1\phi_{\tau1}+\lambda_2\phi_{\tau2}-\phi_{\tau2}\right)
\end{aligned}$$

Rewriting the first equation, we get:

$$\phi_{21}=\frac{\lambda_2(1-\lambda_2)\phi_{12}+\lambda_\tau(\lambda_1\phi_{2\tau}-\lambda_2\phi_{1\tau})}{\lambda_1(1-\lambda_1)}$$

Note that for all $(i,j)\in I^2$, $\phi_{ij}=p_{i|i}-p_{i|j}$ is bounded and belongs to $[-1,1]$, and $\lambda_1,\lambda_2\in(0,1)$. Thus, $\lim_{\lambda_\tau\to0}\lambda_\tau(\lambda_1\phi_{2\tau}-\lambda_2\phi_{1\tau})=0$. Moreover, let $\lambda(\lambda_\tau)\in(0,1)$ be the share of θ_2 with respect to θ_1. We thus have $\lambda_1=(1-\lambda(\lambda_\tau))(1-\lambda_\tau)$, and $\lambda_2=\lambda(\lambda_\tau)(1-\lambda_\tau)$. Then noting $\lambda\in(0,1)$ the share of θ_2 with respect to θ_1 when λ_τ goes to zero, we have: $\lim_{\lambda_\tau\to0}\lambda_2(1-\lambda_2)=\lim_{\lambda_\tau\to0}\lambda_1(1-\lambda_1)=\lambda(1-\lambda)$. Consequently, $\lim_{\lambda_\tau\to0}\phi_{12}(\lambda,\lambda_\tau)=\lim_{\lambda_\tau\to0}\phi_{21}(\lambda,\lambda_\tau)$. □

AppendixA.3 Proof of Proposition 1

Proposition (Matching probabilities)**.** *When the assortment matrix Φ satisfies the assortment balancing conditions (Property 4), the system defined by matching conditions (Property 2), balancing conditions (Property 3) and assortment matrix conditions (Definition 3) has a unique solution:*

$$\forall(i,j)\in I^2:\quad p_{i|j}=\lambda_i+\sum_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}$$

Proof. Let (S) be the system of equations defined by matching conditions, balancing conditions and assortment matrix conditions:

$$(S)\begin{cases}\forall\, i\in I,\sum_{j\in I}p_{j|i}=1\\ \forall\,(i,j)\in I^2,\lambda_j\cdot p_{i|j}=\lambda_i\cdot p_{j|i}\\ \forall\,(i,j)\in I^2,\phi_{ij}=p_{i|i}-p_{i|j}\end{cases}$$

Suppose there exists matching probabilities $p_{i|j}$ solutions of the system (S). Since $\sum_{k\in I}p_{k|i}=1$, we have $\sum_{k\in I}\lambda_i p_{k|i}=\lambda_i$ for all $i\in I$. Using the balancing conditions, we get $\lambda_i-\sum_{k\in I}\lambda_k p_{i|k}=0$. Moreover, since $\sum_{k\in I}\lambda_k=1$, we have $p_{i|i}=\sum_{k\in I}\lambda_k p_{i|i}$. Adding these two equations, we obtain $p_{i|i}=\lambda_i+\sum_{k\in I}\lambda_k(p_{i|i}-p_{i|k})$ for all $i\in I$, i.e. $p_{i|i}=\lambda_i+\sum_{k\in I}\lambda_k\phi_{ik}$. Since for all $(i,j)\in I^2$, $p_{i|j}=p_{i|i}-\phi_{ij}$, we get $p_{i|j}=\lambda_i+\sum_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}$. We have proven that if a solution of (S) exists, then it must be $p_{i|j}=\lambda_i+\sum_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}$.

We now show that $q_{i|j} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij}$ is solution of (S) using the assortment balancing conditions. First, $q_{i|j}$ satisfies the matching conditions:

$$\forall\, j \in I,\ \sum_{i \in I} q_{i|j} = \sum_{i \in I} \left[\lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \right] = 1 + \sum_{i \in I} \frac{\lambda_i}{\lambda_j} \left[\sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right]$$
$$= 1 + \frac{1}{\lambda_j} \left[\sum_{k \in I} \lambda_k \phi_{jk} - \sum_{i \in I} \lambda_i \phi_{ji} \right] = 1$$

Second, $q_{i|j}$ satisfies the balancing conditions:

$$\forall\, (i,j) \in I^2,\ \lambda_j q_{i|j} - \lambda_i q_{j|i} = \lambda_j \lambda_i + \lambda_j \left[\sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \right] - \lambda_i \lambda_j - \lambda_i \left[\sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right] = 0$$

Finally, $q_{i|j}$ satisfies the assortment matrix conditions:

$$\forall\, (i,j) \in I^2,\ q_{i|i} - q_{i|j} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \lambda_i - \sum_{k \in I} \lambda_k \phi_{ik} + \phi_{ij} = \phi_{ij}$$

□

AppendixA.4 Proof of Lemma 3

Lemma (Conditional probabilities in a population of two residents and one mutant)**.** *When* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$*, if Proposition 1 is satisfied, then we have:*

$$\begin{aligned}
p_{1|1} &= (1-\lambda) + \lambda \cdot \phi_{12} \\
p_{1|2} &= (1-\lambda) \cdot (1-\phi_{12}) \\
p_{1|\tau} &= (1-\lambda) \cdot (1-\sigma) - \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{2|1} &= \lambda \cdot (1-\phi_{12}) \\
p_{2|2} &= \lambda + (1-\lambda) \cdot \phi_{12} \\
p_{2|\tau} &= \lambda \cdot (1-\sigma) + \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{\tau|1} &= 0 \\
p_{\tau|2} &= 0 \\
p_{\tau|\tau} &= \sigma
\end{aligned}$$

where $\Gamma = \lim_{\lambda_\tau \to 0} \frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}$.

Proof. If Proposition 1 is satisfied, the conditional probabilities are:

$$\begin{aligned}
p_{1|1} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \lambda_\tau \cdot \phi_{1\tau} \\
p_{1|2} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \lambda_\tau \cdot \phi_{1\tau} - \phi_{12} \\
p_{1|\tau} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \lambda_\tau \cdot \phi_{1\tau} - \phi_{1\tau} \\
p_{2|1} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \lambda_\tau \cdot \phi_{2\tau} - \phi_{21} \\
p_{2|2} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \lambda_\tau \cdot \phi_{2\tau} \\
p_{2|\tau} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \lambda_\tau \cdot \phi_{2\tau} - \phi_{2\tau} \\
p_{\tau|1} &= \lambda_\tau + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2} - \phi_{\tau 1} \\
p_{\tau|2} &= \lambda_\tau + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2} - \phi_{\tau 2} \\
p_{\tau|\tau} &= \lambda_\tau + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2}
\end{aligned}$$

We can then calculate the limits of the conditional probabilities when the mutant share λ_τ goes to zero. First note that for all $(i,j) \in I^2$, ϕ_{ij} is bounded, and thus $\lim_{\lambda_\tau \to 0} \lambda_\tau \phi_{ij} = 0$. Also, the definition of assortativity implies that: for all $i \in \{1,2\}$, $\lim_{\lambda_\tau \to 0} \phi_{\tau i} = \sigma$.

Let $\lambda(\lambda_\tau) \in (0,1)$ be the share of θ_2 with respect to θ_1. We thus have $\lambda_1 = (1 - \lambda(\lambda_\tau))(1 - \lambda_\tau)$, and $\lambda_2 = \lambda(\lambda_\tau)(1 - \lambda_\tau)$. Then noting $\lambda \in (0,1)$ the share of θ_2 with respect to θ_1 when λ_τ goes to zero, we have: $\lim_{\lambda_\tau \to 0} \lambda_2 = \lambda$ and $\lim_{\lambda_\tau \to 0} \lambda_1 = (1 - \lambda)$.

From Lemma 2, we also have: $\phi_{12}(\lambda, 0) = \phi_{21}(\lambda, 0) \equiv \phi_{12}$.

Finally, we need to compute the limits of $\phi_{1\tau}$ and $\phi_{2\tau}$. We will use the assortment balancing conditions:

$$\begin{aligned}
\lambda_2 \left(\lambda_2 \phi_{12} + \lambda_\tau \phi_{1\tau} - \phi_{12}\right) &= \lambda_1 \left(\lambda_1 \phi_{21} + \lambda_\tau \phi_{2\tau} - \phi_{21}\right) \\
\lambda_\tau \left(\lambda_2 \phi_{12} + \lambda_\tau \phi_{1\tau} - \phi_{1\tau}\right) &= \lambda_1 \left(\lambda_1 \phi_{\tau 1} + \lambda_2 \phi_{\tau 2} - \phi_{\tau 1}\right) \\
\lambda_\tau \left(\lambda_1 \phi_{21} + \lambda_\tau \phi_{2\tau} - \phi_{2\tau}\right) &= \lambda_2 \left(\lambda_1 \phi_{\tau 1} + \lambda_2 \phi_{\tau 2} - \phi_{\tau 2}\right)
\end{aligned}$$

Rewriting the second and third assortment balancing conditions, we get:

$$\begin{aligned}
\phi_{1\tau} &= \frac{\lambda_2}{1 - \lambda_\tau} \phi_{12} + \frac{\lambda_1}{1 - \lambda_\tau} \frac{(1 - \lambda_1)\phi_{\tau 1} - \lambda_2 \phi_{\tau 2}}{\lambda_\tau} \\
\phi_{2\tau} &= \frac{\lambda_1}{1 - \lambda_\tau} \phi_{21} + \frac{\lambda_2}{1 - \lambda_\tau} \frac{(1 - \lambda_2)\phi_{\tau 2} - \lambda_1 \phi_{\tau 1}}{\lambda_\tau}
\end{aligned}$$

Taking the limit when λ_τ goes to zero:

$$\begin{aligned}
\lim_{\lambda_\tau \to 0} \phi_{1\tau} &= \lambda\phi_{12} + (1-\lambda) \lim_{\lambda_\tau \to 0} \frac{\left[\lambda(\lambda_\tau) + \lambda_\tau - \lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 1} - \left[\lambda(\lambda_\tau) - \lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 2}}{\lambda_\tau} \\
&= \lambda\phi_{12} + (1-\lambda) \lim_{\lambda_\tau \to 0} \left[(1-\lambda(\lambda_\tau))\phi_{\tau 1} + \lambda(\lambda_\tau)\phi_{\tau 2} + \lambda(\lambda_\tau)\frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}\right] \\
&= \lambda\phi_{12} + (1-\lambda)\sigma + \lambda(1-\lambda)\Gamma \\
\lim_{\lambda_\tau \to 0} \phi_{2\tau} &= (1-\lambda)\phi_{12} + \lambda \lim_{\lambda_\tau \to 0} \frac{\left[1-\lambda(\lambda_\tau) + \lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 2} - \left[1-\lambda(\lambda_\tau) - \lambda_\tau + \lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 1}}{\lambda_\tau} \\
&= (1-\lambda)\phi_{12} + \lambda \lim_{\lambda_\tau \to 0} \left[(1-\lambda(\lambda_\tau))\phi_{\tau 1} + \lambda(\lambda_\tau)\phi_{\tau 2} - (1-\lambda(\lambda_\tau))\frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}\right] \\
&= (1-\lambda)\phi_{12} + \lambda\sigma - \lambda(1-\lambda)\Gamma
\end{aligned}$$

where $\Gamma = \lim_{\lambda_\tau \to 0} \frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}$.

Putting it all together, the limits of the conditional probabilities are:

$$\begin{aligned}
p_{1|1} &= (1-\lambda) + \lambda \cdot \phi_{12} \\
p_{1|2} &= (1-\lambda) \cdot (1-\phi_{12}) \\
p_{1|\tau} &= (1-\lambda) \cdot (1-\sigma) - \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{2|1} &= \lambda \cdot (1-\phi_{12}) \\
p_{2|2} &= \lambda + (1-\lambda) \cdot \phi_{12} \\
p_{2|\tau} &= \lambda \cdot (1-\sigma) + \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{\tau|1} &= 0 \\
p_{\tau|2} &= 0 \\
p_{\tau|\tau} &= \sigma
\end{aligned}$$

□

AppendixB. Analysis of evolutionary stability: Proofs

In this section, we provide the proofs related to the analysis of evolutionary stability. We are in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ (equivalently $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \lambda_\tau)$).

AppendixB.1 Proof of Lemma 1

Lemma. *$B^{NE}(s)$ is compact for each $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau) \in \Theta^3 \times (0,1) \times [0,1)$.*
If for all $i \in I$ u_{θ_i} are concave in their first arguments, then $B^{NE}(s) \neq \emptyset$.
The correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is upper hemi-continuous.

Proof. This proof extends the proof provided by Alger and Weibull (2013) for a population of two types to a population of three types. It follows similar arguments and reasoning.

First, from the definition of a Bayesian Nash equilibrium (Definition 1), we have that, in a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, $(x_1, x_2, x_\tau) \in X^3$ is a type-homogeneous Bayesian Nash equilibrium if:

$$\forall i \in I : \quad x_i \in \underset{x \in X}{\operatorname{argmax}} \quad \sum_{j \in I} p_{j|i} \cdot u_{\theta_i}(x, x_j)$$

With $\lambda_1 = (1-\lambda)(1-\lambda_\tau)$ and $\lambda_2 = \lambda(1-\lambda_\tau)$, we can rewrite the matching probabilities in function of the assortment functions and population shares (Proposition 1). Thus, we get:

$$\forall i \in I : \quad x_i \in \underset{x \in X}{\operatorname{argmax}} \quad \sum_{j \in I} \left(\left[\lambda_j + \sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right] \cdot u_{\theta_i}(x, x_j) \right)$$

Fixing the population state s, i.e. fixing $(\theta_i)_{i \in I}$ and $(\lambda, \lambda_\tau) \in (0,1) \times [0,1)$, we note for all $i \in I$ $U_{s,i} : X^4 \to \mathbb{R}$ the functions defined by:

$$U_{s,i}(x, x_1, x_2, x_\tau) = \sum_{j \in I} \left(\left[\lambda_j + \sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right] \cdot u_{\theta_i}(x, x_j) \right)$$

For all $i \in I, u_{\theta_i}$ is continuous and thus $U_{s,i}$ is also continuous. Since X is compact, then the solution correspondence $\beta_{s,i} : X^3 \rightrightarrows X$ defined by $\beta_{s,i}(x_1, x_2, x_\tau) = \underset{x \in X}{\operatorname{argmax}} U_{s,i}(x, x_1, x_2, x_\tau)$ are non-empty and compact-valued by the Weierstrass's maximum theorem. Hence, the combined correspondence $B_s : X^3 \rightrightarrows X^3$, defined by $B_s(x_1, x_2, x_\tau) = \times_{i \in I} \beta_{s,i}(x_1, x_2, x_\tau)$ is compact valued and, by Berge's maximum theorem, upper hemi-continuous. Hence, B_s has a closed graph and the set of fixed points of B_s, i.e. $B^{NE}(s) = \{(x_i)_{i \in I} : (x_i)_{i \in I} \in B_s((x_i)_{i \in I})\}$, is closed, so that $B^{NE}(s)$ is compact for each $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau) \in \Theta^3 \times (0,1) \times [0,1)$.

Second, since for all $i \in I, u_{\theta_i}$ is concave in their first arguments then so are $U_{s,i}$.Thus, B_s is convex-valued and has a fixed point by Kakutani's fixed point theorem, i.e. $B^{NE}(s)$ is non-empty.

Third, fixing $(\theta_i)_{i \in I}$, we write for all $i \in I$ $V_{\theta,i} : X^4 \times (0,1) \times [0,1) \to \mathbb{R}$ the functions defined by:

$$V_{\theta,i}(x, x_1, x_2, x_\tau, \lambda, \lambda_\tau) = \sum_{j \in I} \left(\left[\lambda_j + \sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right] \cdot u_{\theta_i}(x, x_j) \right)$$

Since for all $(i,j) \in I^2, u_{\theta_i}$ and ϕ_{ij} are continuous, so are $V_{\theta,i}$. Let $V^*_{\theta,i} : X^3 \times (0,1) \times [0,1) \to \mathbb{R}$ the functions defined by $V^*_{\theta,i}(x_1, x_2, x_\tau, \lambda, \lambda_\tau) = \max_{x \in X} V_{\theta,i}(x, x_1, x_2, x_\tau, \lambda, \lambda_\tau)$. By Berge's maximum theorem, $V^*_{\theta,i}$ are continuous. Moreover, by definition of $B^{NE}(s)$, we have, $(x_1, x_2, x_\tau) \in B^{NE}(s)$ if and only if for all $i \in I$:

$$V^*_{\theta,i}(x_1, x_2, x_\tau, \lambda, \lambda_\tau) - V_{\theta,i}(x, x_1, x_2, x_\tau, \lambda, \lambda_\tau) \geq 0 \quad \forall x \in X$$

Let $< \lambda_t >_{t \in \mathbb{N}} \to \lambda^0$ and $< \lambda_{\tau,t} >_{t \in \mathbb{N}} \to \lambda^0_\tau$, and suppose that $(x_{1,t}, x_{2,t}, x_{\tau,t}) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \lambda_{\tau,t})$

and for all $i \in I$, $x_{i,t} \to x_i^0$. By continuity of $V_{\theta,i}$ and $V^*_{\theta,i}$, we have for all $i \in I$:

$$V^*_{\theta,i}(x_1^0, x_2^0, x_\tau^0, \lambda^0, \lambda_\tau^0) - V_{\theta,i}(x, x_1^0, x_2^0, x_\tau^0, \lambda^0, \lambda_\tau^0) \geq 0 \quad \forall x \in X$$

This last results proves that $(x_1^0, x_2^0, x_\tau^0) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^0, \lambda_\tau^0)$ and therefore that the correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is upper hemi-continuous. □

AppendixB.2 Proof of Lemma 4

Lemma. *When the population state is $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, if for all $i \in \{1,2\}$, $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) > \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ for all $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ then there exists an $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.*

Proof. Suppose that in the population state $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, we have for all $i \in \{1,2\}$, $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) > \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ for all $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$.

For all $i \in I$, the type-fitness Π_{θ_i} are continuous by continuity of the game payoffs and of the assortment functions. Thus, the strict inequalities hold for all $(\hat{x}_1, \hat{x}_2, \hat{x}_\tau)$ in a neighborhood $U \subset X^3 \times (0,1) \times [0,1)$ of $(x_1, x_2, x_\tau, \lambda^\circ, 0)$. Using Lemma 1, we know that $B^{NE}(\theta_1, \theta_2, \tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. If $(x_{1,t}, x_{2,t}, x_{\tau,t}) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \lambda_{\tau,t})$ for all $t \in \mathbb{N}$, $(\lambda_t, \lambda_{\tau,t}) \to (\lambda^\circ, 0)$ and $\langle (x_{1,t}, x_{2,t}, x_{\tau,t}) \rangle_{t \in \mathbb{N}}$ converges, then the limit point (x_1^*, x_2^*, x_τ^*) necessarily belongs to $B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \lambda_{\tau,t} < \bar{\varepsilon}$, $|\lambda_t - \lambda^\circ| < \bar{\varepsilon}$ and $(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t}) \in U$, so that for all $i \in I$, $\Pi_{\theta_i}(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t}) > \Pi_{\theta_\tau}(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t})$. □

AppendixB.3 Proof of Lemma 5

Lemma. *When the population state is $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, if there exists $i \in \{1,2\}$ such that $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) < \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ with $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ a singleton, then there does not exist an $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.*

Proof. Suppose that in the population state is $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, there exists $i \in \{1,2\}$ such that $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) < \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ with $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ a singleton.

For all $i \in I$, the type-fitness Π_{θ_i} are continuous by continuity of the game payoffs and of the assortment functions. Thus, the strict inequalities hold for all $(\hat{x}_1, \hat{x}_2, \hat{x}_\tau)$ in a neighborhood $U \subset X^3 \times (0,1) \times [0,1)$ of $(x_1, x_2, x_\tau, \lambda^\circ, 0)$. Using Lemma 1, we know that $B^{NE}(\theta_1, \theta_2, \tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. If $(x_{1,t}, x_{2,t}, x_{\tau,t}) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \lambda_{\tau,t})$ for all $t \in \mathbb{N}$, $(\lambda_t, \lambda_{\tau,t}) \to (\lambda^\circ, 0)$ and $\langle (x_{1,t}, x_{2,t}, x_{\tau,t}) \rangle_{t \in \mathbb{N}}$ converges, then the limit point (x_1^*, x_2^*, x_τ^*) necessarily belongs to $B^{NE}(s^\circ)$. Since by assumption $B^{NE}(s^\circ)$ is a singleton, we have $(x_1^*, x_2^*, x_\tau^*) = (x_1^\circ, x_2^\circ, x_\tau^\circ)$.

Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \lambda_{\tau,t} < \bar{\varepsilon}$, $|\lambda_t - \lambda^\circ| < \bar{\varepsilon}$ and $(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t}) \in U$, so that $\Pi_{\theta_i}(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t}) < \Pi_{\theta_\tau}(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t})$. □

AppendixB.4 Proof of Proposition 2 and Corollary 1

Proposition (Type-fitness equality)**.** *In the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ \in (0,1)$, homo oeconomicus (θ_1) and homo kantiensis (θ_2) earn the same type fitness if and only if:*

1. *When $S_\pi = 0$: $Q_\pi = 0$, i.e. $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.*
2. *When $S_\pi \neq 0$: $\lambda^\circ = Q_\pi / [(1 - \phi_{12}) S_\pi]$.*

Moreover, if homo oeconomicus and homo kantiensis earn the same type fitness, then $\phi_{12} \in (0,1)$.

Corollary (Type-fitness equality under uniformly-constant assortment)**.** *In the population state $s = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ \in (0,1)$, homo oeconomicus (θ_1) and homo kantiensis (θ_2) earn the same type fitness under uniformly-constant assortment if and only if:*

1. *When $S_\pi < 0$: $(\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) < \sigma < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$ and $\lambda^\circ = Q_\pi / [(1-\sigma) S_\pi]$.*
2. *When $S_\pi = 0$: $\sigma = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.*
3. *When $S_\pi > 0$: $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$ and $\lambda^\circ = Q_\pi / [(1-\sigma) S_\pi]$.*

Proof. Suppose that *homo oeconomicus* and *homo kantiensis* earn the same type fitness, i.e. $\Pi_{\theta_1}(D, C, s^\circ) = \Pi_{\theta_2}(D, C, s^\circ)$ with:

$$\begin{aligned} \Pi_{\theta_1}(D, C, s^\circ) &= [(1 - \lambda^\circ) + \lambda^\circ \cdot \phi_{12}] \cdot \pi^{DD} + [\lambda^\circ (1 - \phi_{12})] \cdot \pi^{DC} \\ \Pi_{\theta_2}(D, C, s^\circ) &= [(1 - \lambda^\circ)(1 - \phi_{21})] \cdot \pi^{CD} + [\lambda^\circ + (1 - \lambda^\circ)\phi_{21}] \cdot \pi^{CC} \end{aligned}$$

Then we have:

$$\lambda^\circ (1 - \phi_{12}) S_\pi = Q_\pi \tag{B.1}$$

$$(1 - \lambda^\circ)(1 - \phi_{12}) S_\pi = R_\pi \tag{B.2}$$

Where $Q_\pi \equiv \pi^{DD} - \pi^{CD} - \phi_{12}(\pi^{CC} - \pi^{CD})$, $R_\pi \equiv \pi^{CC} - \pi^{DC} - \phi_{12}(\pi^{DD} - \pi^{DC})$ and $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$.

We first show that $\phi_{12} < 1$. Recall that $\phi_{12} \in [-1, 1]$ by definition of the assortment (Definition 3). Suppose that $\phi_{12} = 1$. This means that *homo oeconomicus* and *homo kantiensis* individuals only meet individuals of their own type. Thus, the type-fitness of *homo oeconomicus* is $\Pi_{\theta_1}(D, C, s^\circ) = \pi^{DD}$, and the type-fitness of *homo kantiensis* is $\Pi_{\theta_2}(D, C, s^\circ) = \pi^{CC}$. Since $\pi^{CC} > \pi^{DD}$ by definition of a prisoner's dilemma, *homo kantiensis* earns a strictly greater type-fitness than *homo oeconomicus*, which contradicts our assumption that the two types earn the same fitness. Hence, $\phi_{12} < 1$.

We now distinguish two cases: $S_\pi = 0$ and $S_\pi \neq 0$.

When $S_\pi = 0$, then $Q_\pi = 0$ (Equation B.1). Thus, $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) > 0$ because $0 < \pi^{DD} - \pi^{CD} < \pi^{CC} - \pi^{CD}$ by definition of a prisoner's dilemma ($\pi^{CD} < \pi^{DD} < \pi^{CC} < \pi^{DC}$), and we are in case 1. of the proposition.[25] Under uniformly-constant assortment $\phi_{12} = \sigma$ and we are in case 2. of the corollary.

When $S_\pi \neq 0$, we have $\lambda^\circ > 0$ and $(1 - \lambda^\circ) > 0$ since $\lambda^\circ \in (0, 1)$ by assumption. Moreover, $(1 - \phi_{12}) > 0$ since $\phi_{12} < 1$. Thus, $Q_\pi \neq 0$, $R_\pi \neq 0$ and Q_π and R_π are of the same sign than S_π (Equations B.1 and B.2). Hence, $Q_\pi \cdot R_\pi > 0$ and $\lambda^\circ = Q_\pi / [(1 - \phi_{12}) S_\pi]$, and we are in case 2. of the proposition. When $S_\pi < 0$, then $Q_\pi < 0$ and $R_\pi < 0$. Thus, $0 < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) < \phi_{12}$ and $\phi_{12} < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$ by definition of a prisoner's dilemma, which proves $\phi_{12} > 0$. Under uniformly-constant assortment $\phi_{12} = \sigma$ and we are in case 1. of the corollary. Similarly, when $S_\pi > 0$, then $Q_\pi > 0$ and $R_\pi > 0$. Thus, $\phi_{12} < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$ and $0 < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \phi_{12}$ by definition of a prisoner's dilemma, which proves $\phi_{12} > 0$. Under uniformly-constant assortment $\phi_{12} = \sigma$ and we are in case 3. of the corollary.

For the converse, if one of the two cases of the Proposition (or one of the three cases of the Corollary) is true, then Equation (B.1) is satisfied and *homo oeconomicus* and *homo kantiensis* earn the same type fitness. □

AppendixB.5 Proof of Lemma 6 and Corollary 2

Lemma (Difference in type fitness between residents and mutant)**.** *Let a population $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, with $\lambda^\circ \in (0, 1)$, engaged in a prisoners' dilemma such that the residents earn the same type fitness Π_θ for $(x_1, x_2) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$ with $x^1 \neq x^2$. Then, the difference in type-fitness between the residents and the mutant for $(x_1, x_2, x_\tau) \in B^{NE}(s)$ is:*

$$\begin{aligned}\Pi_\theta - \Pi_{\theta_\tau} = {} & [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12} - \sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma] \cdot (\alpha_2 - \alpha_1)^2 \cdot S_\pi \\ & + [(\gamma - \lambda^\circ)(\phi_{12} - \sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma] \cdot (\alpha_2 - \alpha_1) \cdot [\alpha_2(\pi^{CC} - \pi^{CD}) + (1-\alpha_2)(\pi^{DC} - \pi^{DD})]\end{aligned}$$

Corollary (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, when θ_1 is homo oeconomicus, θ_2 is homo kantiensis and $\lambda^\circ \in (0, 1)$, engaged in a prisoners' dilemma such that the residents earn the same type fitness Π_θ for $(D, C) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$. Then, the difference in type-fitness between the residents and the mutant for $(D, C, x_\tau) \in B^{NE}(s)$ is:*

$$\Pi_\theta - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1 - \alpha_\tau)S_\pi$$

Proof. Let $(x_1, x_2, x_\tau) \in X^3$ be a Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Using Lemma 3 and noting $\pi_{ij} \equiv \pi(x_i, x_j)$ and $\Pi_{\theta_i} \equiv \Pi_{\theta_i}(x_1, x_2, x_\tau, s)$ for all $(i, j) \in I^2$, we can write

[25] Note that we also have $R_\pi = 0$ (Equation B.2) so that $\phi_{12} = (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$. Indeed, since $S_\pi = \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC} = 0$, $\pi^{DD} - \pi^{CD} = \pi^{DC} - \pi^{CC}$ and $\pi^{CC} - \pi^{CD} = \pi^{DC} - \pi^{DD}$.

the type fitness of each type:

$$\begin{cases} \Pi_{\theta_1} = (1-\lambda^\circ+\lambda^\circ\phi_{12})\cdot\pi_{11} + \lambda^\circ(1-\phi_{12})\cdot\pi_{12} \\ \Pi_{\theta_2} = (1-\lambda^\circ)(1-\phi_{12})\cdot\pi_{21} + [\lambda+(1-\lambda)\phi_{12}]\cdot\pi_{22} \\ \Pi_{\theta_\tau} = [(1-\lambda^\circ)(1-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot\pi_{\tau 1} + [\lambda^\circ(1-\sigma)+\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot\pi_{\tau 2} + \sigma\cdot\pi_{\tau\tau} \end{cases}$$

We know from Property 1 that $(x_1, x_2) \in B^{NE}(s^\circ)$ with $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$. By assumption, θ_1 and θ_2 earns the same type fitness Π_θ in s°. Consequently, they also earn the same type fitness in all Bayesian Nash equilibria in the population state s, i.e. $\Pi_{\theta_1} = \Pi_{\theta_2} \equiv \Pi_\theta$ because in the state s the residents are matched between them, i.e. $\pi_{1\tau}$ and $\pi_{2\tau}$ do not appear in the expression of their type fitness.

In a finite symmetric 2×2 fitness games, let A be the matrix of the payoffs in this game, with π^{ij} the payoff when pure strategy i is played against pure strategy j. The payoff obtained by an individual playing strategy x_i when matched with an individual playing x_j is then: $\pi(x_i, x_j) = \pi_{ij} = x_i^\intercal A x_j$. We can rewrite the payoffs in function of the matrix payoff A:

$$\begin{cases} \Pi_{\theta_1} = x_1^\intercal\left[(1-\lambda^\circ)(1-\phi_{12})Ax_1 + \lambda^\circ(1-\phi_{12})Ax_2\right] + \phi_{12}x_1^\intercal Ax_1 \\ \Pi_{\theta_2} = x_2^\intercal\left[(1-\lambda^\circ)(1-\phi_{12})Ax_1 + \lambda^\circ(1-\phi_{12})Ax_2\right] + \phi_{12}x_2^\intercal Ax_2 \\ \Pi_{\theta_\tau} = x_\tau^\intercal\left[((1-\lambda^\circ)(1-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma)Ax_1 + (\lambda^\circ(1-\sigma)+\lambda^\circ(1-\lambda^\circ)\Gamma)Ax_2\right] + \sigma x_\tau^\intercal Ax_\tau \end{cases}$$

Let $\alpha_1, \alpha_2, \alpha_\tau \in [0,1]$ be the probabilities that $\theta_1, \theta_2, \theta_\tau$ individuals attach to the first pure strategy: $x_1 = (\alpha_1, 1-\alpha_1)$, $x_2 = (\alpha_2, 1-\alpha_2)$ and $x_\tau = (\alpha_\tau, 1-\alpha_\tau)$. Since $x_1 \neq x_2$, there exists $\gamma \in \mathbb{R}$ such that $x_\tau = (1-\gamma)x_1 + \gamma x_2$ ($\alpha_\tau = (1-\gamma)\alpha_1 + \gamma\alpha_2$).

From type-fitness equality, we know that $\Pi_{\theta_1} = \Pi_{\theta_2} = \Pi_\theta$. Thus, $(1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} = \Pi_\theta$. We can then write the difference between the payoff of the residents and the payoff of the mutants as follows:

$$\begin{aligned} \Pi_\theta - \Pi_\tau &= (1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} - \Pi_\tau \\ &= [(1-\gamma)\phi_{12} - (1-\gamma)^2\sigma - (1-\gamma)(1-\lambda^\circ)(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_1^\intercal Ax_1 \\ &+ [-\gamma(1-\gamma)\sigma - (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) - (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_1^\intercal Ax_2 \\ &+ [-\gamma(1-\gamma)\sigma - \gamma(1-\lambda^\circ)(\phi_{12}-\sigma) + \gamma\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_2^\intercal Ax_1 \\ &+ [\gamma\phi_{12} - \gamma^2\sigma - \gamma\lambda^\circ(\phi_{12}-\sigma) - \gamma\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_2^\intercal Ax_2 \end{aligned}$$

Rearranging, we get:

$$\begin{aligned} \Pi_\theta - \Pi_\tau = {} & [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot[x_1^\intercal Ax_1 - x_1^\intercal Ax_2 - x_2^\intercal Ax_1 + x_2^\intercal Ax_2] \\ & + [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot[x_2^\intercal A(x_2 - x_1)] \end{aligned}$$

We can further develop this expression, using the pure-strategies payoffs:

$$\begin{cases} x_1^\intercal A x_1 = \alpha_1^2\pi^{11} + \alpha_1(1-\alpha_1)(\pi^{21}+\pi^{12}) + (1-\alpha_1)^2\pi^{22} \\ x_1^\intercal A x_2 = \alpha_1\alpha_2\pi^{11} + \alpha_1(1-\alpha_2)\pi^{12} + (1-\alpha_1)\alpha_2\pi^{21} + (1-\alpha_1)(1-\alpha_2)\pi^{22} \\ x_2^\intercal A x_1 = \alpha_1\alpha_2\pi^{11} + \alpha_2(1-\alpha_1)\pi^{12} + (1-\alpha_2)\alpha_1\pi^{21} + (1-\alpha_1)(1-\alpha_2)\pi^{22} \\ x_2^\intercal A x_2 = \alpha_2^2\pi^{11} + \alpha_2(1-\alpha_2)(\pi^{21}+\pi^{12}) + (1-\alpha_2)^2\pi^{22} \end{cases} \tag{B.3}$$

Therefore:

$$\begin{aligned} &x_1^\intercal A x_1 - x_1^\intercal A x_2 - x_2^\intercal A x_1 + x_2^\intercal A x_2 = (\alpha_1-\alpha_2)^2\left(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21}\right) \\ &x_2^\intercal A(x_2-x_1) = (\alpha_2-\alpha_1)[\alpha_2(\pi^{11}-\pi^{12}) + (1-\alpha_2)(\pi^{21}-\pi^{22})] \end{aligned} \tag{B.4}$$

Consequently, the difference in type fitness when the share of the mutant goes to zero is:

$$\begin{aligned} \Pi_\theta - \Pi_\tau = &[\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)^2\cdot\left(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21}\right) \\ &+ [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)\cdot[\alpha_2(\pi^{11}-\pi^{12}) + (1-\alpha_2)(\pi^{21}-\pi^{22})] \end{aligned} \tag{B.5}$$

In a prisoners' dilemma, the first pure strategy is cooperate (C) and the second pure strategy is defect (D). Hence, with $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$, we have:

$$\begin{aligned} \Pi_\theta - \Pi_\tau = &[\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)^2\cdot S_\pi \\ &+ [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)\cdot[\alpha_2(\pi^{CC}-\pi^{CD}) + (1-\alpha_2)(\pi^{DC}-\pi^{DD})] \end{aligned}$$

When the assortment is uniformly constant, $\phi_{12} = \sigma$ and $\Gamma = 0$. Thus, we obtain:

$$\Pi_\theta - \Pi_\tau = \gamma(1-\gamma)\sigma(\alpha_2-\alpha_1)^2 S_\pi$$

Since *homo oeconomicus* always defect $\alpha_1 = 0$, and since *homo kantiensis* always cooperate $\alpha_2 = 1$. Hence, $\gamma = \alpha_\tau$ and:

$$\Pi_\theta - \Pi_\tau = \alpha_\tau(1-\alpha_\tau)\sigma S_\pi$$

□

AppendixB.6 Proof of Lemma 7

Lemma (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$*, when* θ_1 *is homo oeconomicus,* θ_2 *is homo kantiensis, engaged in a prisoners' dilemma. Then, we have for*

any $(D, C, x_\tau) \in B^{NE}(s)$*:*

$$(1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau \Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. Let $(D, C, x_\tau) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, using Proposition 1 and noting $\lambda_1 = (1-\lambda)(1-\lambda_\tau)$ and $\lambda_2 = \lambda(1-\lambda_\tau)$, we can write the type fitness of each type:

$$\begin{cases} \Pi_{\theta_1} = (\lambda_1(1-\sigma)+\sigma)\cdot\pi^{DD} + \lambda_2(1-\sigma)\cdot\pi^{DC} + \lambda_\tau(1-\sigma)\cdot\pi_{D\tau} \\ \Pi_{\theta_2} = \lambda_1(1-\sigma)\cdot\pi^{CD} + (\lambda_2(1-\sigma)+\sigma)\cdot\pi^{CC} + \lambda_\tau(1-\sigma)\cdot\pi_{C\tau} \\ \Pi_{\theta_\tau} = \lambda_1(1-\sigma)\cdot\pi_{\tau D} + \lambda_2(1-\sigma)\cdot\pi_{\tau C} + (\lambda_\tau(1-\sigma)+\sigma)\cdot\pi_{\tau\tau} \end{cases}$$

We have:

$$\begin{aligned} \pi_{D\tau} &= \alpha_\tau\pi^{DC} + (1-\alpha_\tau)\pi^{CC} \\ \pi_{C\tau} &= \alpha_\tau\pi^{CC} + (1-\alpha_\tau)\pi^{CD} \\ \pi_{\tau D} &= \alpha_\tau\pi^{CD} + (1-\alpha_\tau)\pi^{DD} \\ \pi_{\tau C} &= \alpha_\tau\pi^{CC} + (1-\alpha_\tau)\pi^{DC} \\ \pi_{\tau\tau} &= \alpha_\tau^2\pi^{CC} + \alpha_\tau(1-\alpha_\tau)(\pi^{CD}+\pi^{DC}) + (1-\alpha_\tau)^2\pi^{DD} \end{aligned}$$

Therefore:

$$\begin{aligned} (1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau\Pi_{\theta_2} = &\left[(1-\alpha_\tau)\lambda_1(1-\sigma) + (1-\alpha_\tau)\sigma + (1-\alpha_\tau)^2\lambda_\tau(1-\sigma)\right]\cdot\pi^{DD} \\ &+ \left[(1-\alpha_\tau)\lambda_2(1-\sigma) + \alpha_\tau(1-\alpha_\tau)\lambda_\tau(1-\sigma)\right]\cdot\pi^{DC} \\ &+ \left[\alpha_\tau\lambda_1(1-\sigma) + \alpha_\tau(1-\alpha_\tau)\lambda_\tau(1-\sigma)\right]\cdot\pi^{CD} \\ &+ \left[\alpha_\tau\lambda_2(1-\sigma) + \alpha_\tau\sigma + \alpha_\tau^2\lambda_\tau(1-\sigma)\right]\cdot\pi^{CC} \end{aligned}$$

And:

$$\begin{aligned} \Pi_{\theta_\tau} = &\left[(1-\alpha_\tau)\lambda_1(1-\sigma) + (1-\alpha_\tau)^2\lambda_\tau(1-\sigma) + (1-\alpha_\tau)^2\sigma\right]\cdot\pi^{DD} \\ &+ \left[(1-\alpha_\tau)\lambda_2(1-\sigma) + \alpha_\tau(1-\alpha_\tau)\lambda_\tau(1-\sigma) + \alpha_\tau(1-\alpha_\tau)\sigma\right]\cdot\pi^{DC} \\ &+ \left[\alpha_\tau\lambda_1(1-\sigma) + \alpha_\tau(1-\alpha_\tau)\lambda_\tau(1-\sigma) + \alpha_\tau(1-\alpha_\tau)\right]\cdot\pi^{CD} \\ &+ \left[\alpha_\tau\lambda_2(1-\sigma) + \alpha_\tau^2\lambda_\tau(1-\sigma) + \alpha_\tau^2\sigma\right]\cdot\pi^{CC} \end{aligned}$$

Consequently:

$$(1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau\Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

□

AppendixB.7 Proof of Theorem 1

Theorem (Evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis*)**.** *In a prisoners' dilemma under uniformly-constant assortment when* Θ *is rich, there exists a heterogeneous evolutionarily-stable population of homo oeconomicus and homo kantiensis against all types* $\theta_\tau \notin \Theta_{12}$ *if and only if* $S_\pi > 0$ *and* $(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD}) < \sigma < (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$*. Moreover, if* $S_\pi > 0$ *and* $(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD}) < \sigma < (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$*, the cooperation share in the evolutionarily stable population satisfies* $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi)$*.*

Proof. Suppose that there exists an *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* against all types $\theta_\tau \notin \Theta_{12}$. Then, by definition of evolutionary stability (Definition 6), there exists a state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ such that *homo oeconomicus* and *homo kantiensis* earn the same type fitness Π_θ. From Corollary 1, we know that there are only three possible cases:

1. When $S_\pi < 0$: $(\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) < \sigma < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$ and $\lambda^\circ = Q_\pi/\left[(1-\sigma)S_\pi\right]$.
2. When $S_\pi = 0$: $\sigma = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.
3. When $S_\pi > 0$: $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$ and $\lambda^\circ = Q_\pi/\left[(1-\sigma)S_\pi\right]$.

Let θ_τ a mutant committed to the strategy $\hat{x_\tau} = (1/2; 1/2)$. Such a mutant exists since the type set is rich by assumption. Note also that $\theta_\tau \notin \Theta_{12}$. Then, $(D, C, \hat{x_\tau})$ is a Bayesian Nash equilibrium in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, \lambda_\tau)$ with $\lambda_\tau \in (0,1)$. Using Lemma 7, we have:

$$\frac{\Pi_{\theta_1} + \Pi_{\theta_2}}{2} - \Pi_{\theta_\tau} = \frac{\sigma S_\pi}{4} \tag{B.6}$$

In the three cases satisfying the type-fitness equality, we have $\sigma > 0$ (else *homo oeconomicus* would dominate). Hence, the sign of the left-hand side of Equation B.6 is the same as the sign of S_π. When $S_\pi \leq 0$, we have:

$$\frac{\Pi_{\theta_1} + \Pi_{\theta_2}}{2} \leq \Pi_{\theta_\tau}$$

Hence, θ_τ earns a greater type fitness than the average type-fitness of the residents in all Bayesian Nash equilibria in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, \lambda_\tau)$ with $\lambda_\tau \in (0,1)$. This means that θ_τ earns a greater type fitness than either θ_1 or θ_2 (or both). Thus, the population of *homo oeconomicus* and *homo kantiensis* does not satisfy the second condition for evolutionary stability, which contradicts our initial assumption. Consequently, the only remaining case is $S_\pi > 0$ and then $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.

Conversely, suppose that $S_\pi > 0$ and $(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD}) < \sigma < (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$. Then, from Corollary 1, we know that *homo oeconomicus* and *homo kantiensis* earn the same type fitness Π_θ in their only Bayesian Nash equilibrium (D, C) in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi) \in (0,1)$. Let $\theta_\tau \notin \Theta_{12}$ a mutant and $(D, C, x_\tau) \in B^{NE}(s)$ with $s =$

$(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Using Corollary 2, we can express the difference in type-fitness between the residents and the mutant:

$$\Pi_\theta - \Pi_\tau = \alpha_\tau(1-\alpha_\tau)\sigma S_\pi$$

We have $\sigma > 0$. Moreover, since $\theta_\tau \notin \Theta_{12}$, the mutant does not cooperate or defect, i.e. $\alpha_\tau \in (0,1)$. Thus, $\alpha_\tau(1-\alpha_\tau) > 0$. Finally, $S_\pi > 0$ by assumption. Hence, $\Pi_\theta - \Pi_\tau > 0$. In other words, we have shown that $\Pi_{\theta_1} > \Pi_\tau$ and $\Pi_{\theta_2} > \Pi_\tau$ for any mutant $\theta_\tau \notin \Theta_{12}$ and for any Bayesian Nash equilibrium $(D, C, x_\tau) \in B^{NE}(s)$, with $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Using Lemma 4, we can conclude that the population of *homo oeconomicus* and *homo kantiensis* in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi)$ is evolutionarily stable against all types $\theta_\tau \notin \Theta_{12}$. □

AppendixB.8 Proof of Proposition 3

Proposition (Non evolutionarily-stable population)**.** *In a symmetric 2×2 fitness game where the assortment matrix is uniformly constant and strictly positive, let $s = (\theta_1, \theta_2, \lambda)$ be a heterogeneous population.*
If there exists $(x^1, x^2) \in B^{NE}(s)$ such that $(x^1, x^2) \notin X_\sigma^2$ and if Θ is rich, then the population is not evolutionarily stable.

Proof. The proof follows two steps. First, we show that there always exists a mutant type that earns strictly more than the residents at the limit. Then, we extend this result to a small neighborhood by continuity.
Note that if the population does not respect the Type-fitness equality condition, it is not evolutionarily stable. Thus, we consider next a population that respects the Type-fitness equality condition. (Definition 6.1).

If $x^1 = x^2 = x_\theta \notin X_\sigma$, then there exists $\hat{x} \in X$ such that $u_\sigma(x_\theta, x_\theta) < u_\sigma(\hat{x}, x_\theta)$, i.e. $\pi(x_\theta, x_\theta) < (1-\sigma)\pi(\hat{x}, x_\theta) + \sigma\pi(\hat{x}, \hat{x})$. At the limit when the population share of the mutant goes to zero, this inequality is equivalent to $\Pi_\theta < \Pi_\tau$, for a mutant playing $\hat{x}$. Moreover, since Θ is rich, there exists a type $\theta_\tau \in \Theta$ for which $\hat{x}$ is strictly dominant, i.e. θ_τ always play $\hat{x}$.

If $x^1 \neq x^2$, we know that the difference in payoffs between the residents and mutants τ at the limit satisfies:

$$\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)(\alpha_1 - \alpha_2)^2 (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21})$$

From (B.4), we know that $S_\pi = (\alpha_1 - \alpha_2)^2 (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21})$.[26] Hence, rewriting the expression

[26] Recall that $S_\pi = \pi^{11} + \pi^{22} - \pi^{12} - \pi^{21}$ where π^{ij} denotes the payoff obtained by individual θ_i against individual θ_j; while π_{ij} denotes the payoff of playing pure strategy i against pure strategy j.

above, we have:

$$\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)S_\pi$$

We consider the three different cases of Lemma 8:

1. If $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} > 0$, then $X_\sigma \subseteq \{0, 1\}$,
   Since Θ is rich, if θ_1 or θ_2 individuals do not play pure strategies, it is always possible to find a mutant playing a strategy $\hat{x}$ such that $\gamma(1-\gamma) < 0$ (discussion of Fig.6). In this case, the difference between the two payoffs above is negative and the mutant earns more than the residents at the limit since $\sigma > 0$.
   Else, if θ_1 and θ_2 individuals both play pure strategies, then since $(x^1, x^2) \notin X_\sigma^2$, we have $X_\sigma = \{0\}$ or $X_\sigma = \{1\}$. Thus, one type is playing the Hamiltonian strategy. Without loss of generality and by symmetry, suppose θ_1 individuals are playing the Hamiltonian strategy, and that $X_\sigma = \{1\}$ i.e. θ_1 individuals play the first pure strategy while θ_2 individuals play the second pure strategy. We then have $S_\pi = \pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} > 0$ and we are in case 2. of Proposition 2. So we also have $Q_\pi, R_\pi \geq 0$. Let $x \in X$, such that $x \neq x^2$, i.e. $x = (\eta, 1-\eta)$ with $\eta \in (0, 1]$. Then:

$$\begin{aligned}(1-\sigma)\pi(x, x^2) + \sigma\pi(x, x) &= \pi_{22} - \eta R_\pi - \sigma\eta(1-\eta)S_\pi \\ &< \pi_{22}\end{aligned}$$

   Thus, for all x in X such that $x \neq x^2$, $u_\sigma(x, x^2) < u_\sigma(x^2, x^2)$. This means that the strategy played by individuals θ_2, i.e. the second pure strategy, is also a Hamiltonian strategy. Consequently, $X_\sigma = \{0, 1\}$ which contradicts the assumption $(x^1, x^2) \notin X_\sigma^2$. Hence, this case is impossible.
2. If $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} = 0$, then we have $S_\pi = 0$ ($\alpha_1 \neq \alpha_2$ else the residents play the same strategy). Thus, from Proposition 2, we also have $Q_\pi = R_\pi = 0$. Subtracting, the expression $Q_\pi - S_\pi$, using (B.3), we find:

$$Q_\pi - S_\pi = (\alpha_1 - \alpha_2)[\alpha_2(1+\sigma)(\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21}) + (\pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22})]$$

   Hence, we have $\pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22} = 0$. Therefore, case 2. of Lemma 8 implies that $X_\sigma = [0, 1]$ which contradicts the assumption $(x^1, x^2) \notin X_\sigma^2$, and this case is impossible.
3. If $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} < 0$, then since Θ is rich, it is always possible to find a mutant playing a strategy $\hat{x}$ such that $\gamma(1-\gamma) > 0$ (discussion of Fig.6) so that the mutants earn more than the residents at the limit since $\sigma > 0$.

Consequently, in the different cases when $(x^1, x^2) \notin X_\sigma^2$ and Θ is rich, we have shown that there exists a mutant type θ_τ that earns strictly more than the residents at the limit by being committed

to a strategy $\hat{x}$:

$$\begin{aligned} & \Pi_1(x^1, x^2, \hat{x}, \lambda, 0) < \Pi_\tau(x^1, x^2, \hat{x}, \lambda, 0) \\ \text{and} \quad & \Pi_2(x^1, x^2, \hat{x}, \lambda, 0) < \Pi_\tau(x^1, x^2, \hat{x}, \lambda, 0) \end{aligned}$$

By continuity of the payoffs, these strict inequalities hold for all $(x, y, \hat{x})$ in a neighborhood $U \subset X^3 \times (0,1) \times [0,1)$ of $(x^1, x^2, \hat{x}, \lambda, 0)$. Using Lemma 1, we know that $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. If $(x_t^1, x_t^2, \hat{x}_t) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \varepsilon_t)$ for all $t \in \mathbb{N}$, $(\lambda_t, \varepsilon_t) \to (\lambda, 0)$ and the sequence $\left\langle (x_t^1, x_t^2, \hat{x}_t) \right\rangle_{t \in \mathbb{N}}$ converges, then the limit point $(x^{1*}, x^{2*}, \hat{x}^*)$ necessarily belongs to $B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ which is a singleton by assumption, i.e. $(x^{1*}, x^{2*}, \hat{x}^*) = (x^1, x^2, \hat{x})$. Moreover, since θ_τ is committed to strategy $\hat{x}$, for all $t \in \mathbb{N}$ $\hat{x}_t = \hat{x}$. Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \varepsilon_t < \bar{\varepsilon}$ and $(x_t^1, x_t^2, \hat{x}, \lambda_t, \varepsilon_t) \in U$, so that $\Pi_1(x_t^1, x_t^2, \hat{x}, \lambda_t, \varepsilon_t) < \Pi_\tau(x_t^1, x_t^2, \hat{x}, \lambda_t, \varepsilon_t)$ and $\Pi_2(x_t^1, x_t^2, \hat{x}, \lambda_t, \varepsilon_t) < \Pi_\tau(x_t^1, x_t^2, \hat{x}, \lambda_t, \varepsilon_t)$. □

AppendixB.9 Proof of Theorem 2

Theorem (Evolutionarily stable population)**.** *In a symmetric* 2×2 *fitness game where the assortment matrix is uniformly constant and strictly positive, let* $s = (\theta_1, \theta_2, \lambda)$ *be a heterogeneous population. If for all* $(x^1, x^2) \in B^{NE}(s)$, $(x^1, x^2) \in X_\sigma^2$, *if* $\lambda = Q_\pi/((1-\sigma)S_\pi)$, *and if* $\beta_\sigma(x)$ *is a singleton for all* $x \in X_\sigma$, *then the population* $(\theta_1, \theta_2, \lambda)$ *is evolutionarily stable against all types* $\theta_\tau \notin \Theta_{12}$.

Proof. In the proof, we will need the following lemma showed by Alger and Weibull (2013):

Lemma 8 (Proposition 2 of Alger and Weibull (2013))**.** *Let*

$$\hat{x}(\sigma) = \min\left\{1, \frac{\pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22}}{(1+\sigma)(\pi_{12} + \pi_{21} - \pi_{11} - \pi_{22})}\right\}$$

When $\sigma > 0$,

1. *If* $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} > 0$, *then* $X_\sigma \subseteq \{0, 1\}$.
2. *If* $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} = 0$, *then*

$$X_\sigma = \begin{cases} \{0\}, & \text{if} \quad \pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22} < 0 \\ [0,1], & \text{if} \quad \pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22} = 0 \\ \{1\}, & \text{if} \quad \pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22} > 0 \end{cases}$$

3. *If* $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} < 0$, *then*

$$X_\sigma = \begin{cases} \{0\}, & \text{if} \quad \pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22} \leq 0 \\ \{\hat{x}(\sigma)\}, & \text{if} \quad \pi_{12} + \sigma\pi_{21} - (1+\sigma)\pi_{22} > 0 \end{cases}$$

Now to prove our result, we first show that the residents earn a strictly greater payoff than the mutants at the limit, and then extend the result to a small neighborhood.

First note that if $(x^1, x^2) \in B^{NE}(\theta_1, \theta_2, \lambda)$, then the strategies x^1 and x^2 will also belong to the set of Bayesian Nash equilibria for a population of three types when the mutant share is zero, i.e. $(x^1, x^2, x^\tau) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, where x^τ is the strategy played by mutants τ.

If $(x^1, x^2) \in X_\sigma^2$ such that $x^1 = x^2 = x_\sigma$, then the population $(\theta_1, \theta_2, \lambda)$ satisfies the Payoff Equality condition when the mutant is absent. Moreover, since $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, we have $u_\sigma(x_\sigma, x_\sigma) > u_\sigma(x, x_\sigma)$ for all $x \in X$ such that $x \neq x_\sigma$, i.e. $\pi(x_\sigma, x_\sigma) > (1-\sigma)\pi(x, x_\sigma)+\sigma\pi(x, x)$. In particular when $x = x^\tau$ $(\theta_\tau \notin \Theta_{12})$, we have $\pi(x_\sigma, x_\sigma) > (1-\sigma)\pi(x^\tau, x_\sigma) + \sigma\pi(x^\tau, x^\tau)$. At the limit when the mutant share goes to zero, we have: $\Pi_1 = \Pi_2 = \Pi_\theta = \pi(x_\sigma, x_\sigma)$ and $\Pi_\tau = (1-\sigma)\pi(x^\tau, x_\sigma) + \sigma\pi(x^\tau, x^\tau)$ so that $\Pi_\theta > \Pi_\tau$.

If $(x^1, x^2) \in X_\sigma^2$ such that $x^1 \neq x^2$, then the population $(\theta_1, \theta_2, \lambda)$ satisfies the Payoff Equality condition when the mutant is absent because $\lambda = Q_\pi/((1-\sigma)S_\pi)$ by assumption. Then, we know that the difference in payoffs between the residents and mutants θ_τ at the limit is:

$$\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)(\alpha_1 - \alpha_2)^2 (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21})$$

Moreover, from the proof of Theorem 1, we have $Q_\pi > 0$ and $R_\pi > 0$, and thus from Proposition 2, we also have $S_\pi > 0$. Since $S_\pi = (\alpha_1 - \alpha_2)^2 (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21})$ (B.4), we have $\pi_{11}+\pi_{22}-\pi_{12}-\pi_{21} > 0$. Thus, we are in case 1 of Lemma 8, and since $(x^1, x^2) \in X_\sigma^2$ such that $x^1 \neq x^2$, we know that $X_\sigma = \{0, 1\}$. It means that individuals θ_1 and θ_2 play the two pure strategies. Without loss of generality and by symmetry, we can assume that individuals θ_1 play the pure strategy 1 $(\alpha_1 = 1)$, and that individuals θ_2 play the pure strategy 2 $(\alpha_2 = 0)$. Thus, γ is in fact the probability that θ_τ attaches to the pure strategy 1. Moreover, since $\theta_\tau \notin \Theta_{12}$, mutants cannot play a pure strategy and $\gamma \in (0, 1)$ i.e. $\gamma(1-\gamma) > 0$. We also have $S_\pi = \pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} > 0$, and $\sigma > 0$. Consequently, the difference in payoffs at the limit is strictly positive:

$$\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)S_\pi > 0$$

For both cases ($x^1 = x^2$ and $x^1 \neq x^2$) We have shown:

$$\begin{aligned} &\Pi_1(x^1, x^2, x^\tau, \lambda, 0) > \Pi_\tau(x^1, x^2, x^\tau, \lambda, 0) \\ \text{and} \quad &\Pi_2(x^1, x^2, x^\tau, \lambda, 0) > \Pi_\tau(x^1, x^2, x^\tau, \lambda, 0) \end{aligned}$$

for all $(x^1, x^2, x^\tau) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ and for any $\theta_\tau \notin \Theta_{12}$. Moreover, Π_1, Π_2 and Π_τ are continuous by continuity of the game payoffs and of the assortment functions. Thus, the strict inequalities hold for all $(\hat{x}^1, \hat{x}^2, \hat{x}^\tau)$ in a neighborhood $U \subset X^3 \times (0, 1) \times [0, 1)$ of $(x^1, x^2, x^\tau, \lambda, 0)$. Using Lemma 1, we know that $B^{NE}(\theta_1, \theta_2, \tau, \cdot) : (0, 1) \times [0, 1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. If $(x_t^1, x_t^2, x_t^\tau) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \varepsilon_t)$ for all $t \in \mathbb{N}$, $(\lambda_t, \varepsilon_t) \to (\lambda, 0)$ and $\left\langle (x_t^1, x_t^2, x_t^\tau) \right\rangle_{t\in\mathbb{N}}$ converges, then the limit point $(x^{1*}, x^{2*}, x^{\tau*})$ necessarily belongs to $B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, 0)$. Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \varepsilon_t < \bar{\varepsilon}$ and $(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t) \in U$, so that $\Pi_1(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t) > \Pi_\tau(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t)$ and $\Pi_2(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t) > \Pi_\tau(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t)$. □

AppendixB.10 Proof of Proposition 4

Proposition (Evolutionary stability under state-dependent assortment)**.** *In a prisoners' dilemma, if Θ is rich then there exists $\bar{\sigma} < 1$ such that there does not exist a heterogeneous evolutionary stable population of homo oeconomicus and homo kantiensis for all $\sigma > \bar{\sigma}$.*

Proof. Suppose that *homo oeconomicus* and *homo kantiensis* earn the same type fitness Π_θ in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$. Then, we have $\Pi_\theta = \Pi_{\theta_2} < \pi^{CC}$ because $\Pi_{\theta_2} = p_{1|2} \cdot \pi^{CD} + p_{2|2} \cdot \pi^{CC}$, $\pi^{CD} < \pi^{CC}$ by definition of a prisoners' dilemma and $p_{1|2} > 0$ (since from Proposition 2, $\phi_{12} < 1$). Let $\sigma = 1$ and θ_τ a mutant committed to cooperation. Such a mutant exists since the type set is rich by assumption. Then, the mutants are matched between themselves ($p_{\tau\tau} = 1$) so that $\Pi_{\theta_\tau} = \pi^{CC}$. Hence, we have $\Pi_\theta < \Pi_{\theta_\tau}$ at the limit when the mutant share goes to zero. Since the difference in type fitness between the residents and the mutant is continuous in σ (see Lemma 6), there exists $\bar{\sigma} < 1$ such that the strict inequality holds for all $\sigma > \bar{\sigma}$. Therefore, we have $\Pi_\theta < \Pi_{\theta_\tau}$ for the Bayesian Nash equilibrium $(D, C, C) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, with (D, C, C) a singleton (because each type is committed to its strategy). From Lemma 5, we know that the strict inequality remains valid in a small neighborhood. Consequently, the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable for all $\sigma > \bar{\sigma}$. □