Exploring the diversity of social preferences: Is a heterogeneous population evolutionarily stable under assortative matching?

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 make 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 in a population coexist. 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, Evolutionary stability, 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 of human behaviors (Henrich et al., 2001). Empirical evidence shows that individuals make different decisions when confronted with similar choices. This has been observed in various contexts such as voting behavior (Piketty, 1995), altruism (Andreoni and Miller, 2002), environmental consciousness (Schlegelmilch et al., 1996), risk aversion and saving choices (Burks et al., 2009), health expenditure (Hitiris and Posnett, 1992) and the selection of a life partner (Eastwick and Finkel, 2008), suggesting the existence of distinct preferences among individuals. The diversity in the social behavior of chimpanzees (Van Leeuwen et al., 2012) hints at the possibility of an evolutionary origin behind this heterogeneity. Our goal in this paper is to assess

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 Please do not cite without authors' permission

the evolutionary foundation of the coexistence of more than one type of preference in a population, and to evaluate what types of preferences prevail when they exist.

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. When Maynard Smith and Price (1973) and Maynard Smith (1974) laid the foundation of evolutionary game theory, they intended to identify the strategy providing an evolutionary advantage in animal conflicts between members of a given species. To do so, 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*. This approach led to the emergence of a *homo moralis* type of preference. 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.

We conduct our analysis in the context of strategic interactions between pairwise-matched individuals of a large population. We consider a fitness game which applies to symmetric interactions and asymmetric interactions with ex-ante symmetry. In other words, each individual is as likely to be in one or the other side of the interaction. As 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 (subjective) personal preferences, while their evolutionary success is given by some exogenous payoff (fitness) function. To prevent individuals from deviating from their utility-maximization, we consider the individuals' preferences as their private information.[3] We assume that there are several types, i.e. several preferences, in the population. Each preference is described by a utility function. As in Alger and Weibull (2013), the only restriction we impose on the set of preferences studied is the continuity

[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.

[3] 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).

of the associated utility function.

A key feature of the model lies in the matching process. Building on Bergstrom (2003), we consider that the meeting probability between two individuals follows an exogenous 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. This tendency of individuals to interact more between similar others has been a subject of interest for economics, sociology and biology scholars often called *homophily* (McPherson et al., 2001; Currarini et al., 2009; Golub and Jackson, 2012). To precisely model this *homophily*, we generalize the assortment function suggested by Bergstrom (2003) introducing a novel, type-by-type assortment functions matrix allowing for calibration of the bilateral degree of *homophily* among each potential couple of matched individual.

In this theoretical setting, we investigate the conditions under which two resident types could co-exist and resist the invasion of any mutant. To do so, we generalize the definition of *evolutionarily stable preference* (Alger and Weibull, 2013) by proposing the concept of *evolutionarily stable population*. When a population is made of a unique resident and a mutant, our definition is consistent with the classical setting.

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 in the special case of a uniformly-constant assortment. We also characterize the conditions for this existence. However, not all games welcome a heterogeneous *evolutionarily stable population* playing diverse strategies: in some games all individuals should play the same strategy (heterogeneity in preferences but not in strategies). Somewhat surprisingly, we find a link between the strategies played by the two types of an *evolutionarily stable population*, and the strategies played by an evolutionarily stable *homo moralis* in the framework of Alger and Weibull (2013). Finally, 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.

Our work contributes to two major branches of the literature. First, the evolution of strategies in a heterogeneous population has been extensively studied (mostly in biology) in the context of evolutionary game dynamics.[4] For example, Bergstrom (2003) and Allen and Nowak (2015) study the evolution of cooperation under assortative matching in social dilemmas. Their results are in line with ours when the cooperating individuals are represented by the moral *homo kantiensis* preference and the defectors by the selfish *homo oeconomicus* preference. Second, as argued by Norton et al. (1998), preferences evolve by selection acting on traits that are genetically and culturally transmitted. Thus, economists have adapted the standard evolutionary game theory framework to study the evolution of preferences (Güth and Yaari, 1992). In particular, Dekel et al. (2007) evaluate the evolutionary stability of a distribution of preferences in a population and the associated equilibrium (called a configuration) depending on the observability of preferences. Their definition of stable configuration is close to our concept of *evolutionarily stable population*. Other authors have also analyzed the dynamic evolution of preferences, with application to individualistic preferences (Ok and Vega-Redondo, 2001), to social rewards (Fershtman and Weiss, 1998), to biases (Sandholm, 2001), to negatively interdependent preferences (Koçkesen et al., 2000) or to overconfidence and interdependent preferences (Heifetz et al., 2007). However, most of the literature on preference evolution assumes a uniform random matching. Sethi and Somanathan (2001) provide an interesting discussion of the

[4] See for instance Hofbauer and Sigmund (2003), Sandholm (2010) and Nowak et al. (2010) for a description and review of the field.

evolutionary stability of selfish and reciprocal preferences (as defined by Levine, 1998) under both assortative and non-assortative matching, but without explicitly modeling the assortative matching process. Alger and Weibull (2013) introduce assortative matching but assume only one resident type in the population. Hence, our model helps to bridge this gap in the literature.

The organization of the rest of the paper is as follows: in Section 2 we present the model and extend the assortment function to a population of several types introducing the assortment matrix, in Section 3 we discuss the conditions under which two types can coexist, in Section 4 we define an *evolutionarily stable population* and study the case of uniformly-constant assortment, in Section 5 we discuss our results and delve into the case of state-dependent assortment, and we conclude in Section 6.

2. Model and definitions

As in Alger and Weibull (2013), we consider a large population where individuals are randomly matched into pairs to engage in a symmetric interaction with the common strategy set X. We assume that X is a nonempty, compact and convex set. Individuals are utility maximizers, and their behavior depends on their type $\theta \in \Theta$, i.e. their preferences which are described by a continuous utility function $u_\theta : X^2 \to \mathbb{R}$. Individuals' success in the game is determined by the resulting payoffs: an individual who plays strategy $x \in X$ when the opponent plays strategy $y \in X$ gets material payoff $\pi(x, y)$, where we assume $\pi : X^2 \to \mathbb{R}$ to be continuous.

Before explaining in more details our model framework, we present in the following section the classical setting of a population consisting of two types: one resident type and one mutant type. We introduce some definitions and results useful for the rest of the paper.

2.1 Classical setting

We consider a population of two types $\theta, \tau \in \Theta$. The two types and their respective shares define a population state $s = (\theta, \tau, \varepsilon)$, where $\varepsilon \in (0, 1)$ is the population share of τ. If ε is small, we call θ the resident type and τ the mutant type.[5]

The matching process is random and exogenous[6], and it may be assortative. Let $p_{\tau|\theta}(\varepsilon)$ be the conditional probability that an individual being of type θ is matched with an individual of type τ in the population state $s = (\theta, \tau, \varepsilon)$.[7] Similarly, $p_{\theta|\theta}(\varepsilon)$ is the probability for an individual to be matched with an individual of type θ, conditional on being of type θ. We can then define the assortment function and assortativity:

Definition 1 (Assortment function and assortativity). In a population state $s = (\theta, \tau, \varepsilon)$ with $\epsilon \in (0, 1)$, let $\phi(\varepsilon)$ be the difference between the probability for an individual to be matched with an individual of type θ, conditional on being of type θ herself, and the probability for an individual to be matched with an individual of type θ, conditional on being of type τ.
In other words, we have: $\phi(\varepsilon) = p_{\theta|\theta}(\varepsilon) - p_{\theta|\tau}(\varepsilon)$, defining an assortment function $\phi : (0, 1) \to [-1, 1]$.

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

[6] 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.

[7] This probability is noted $\Pr[\theta|\tau, \varepsilon]$ in Alger and Weibull (2013).

Assuming ϕ is continuous and converges as ε tends to zero, the assortativity $\sigma \in [0,1]$ is the limit of ϕ in zero: $\lim_{\varepsilon \to 0} \phi(\varepsilon) = \sigma$

The assortment function models the *homophily* between two types. Homophily is the tendency of individuals to interact more with others with similar characteristics such as family, ethnicity, age, gender, language, religion, geographic proximity, education, work, association activity or income (Ibarra, 1993; McPherson et al., 2001).

Individuals choose their strategy in order to maximize their utility. A Bayesian Nash Equilibrium (BNE) is a pair of strategies, one for each type, where each strategy is a best reply to the other in the given population state:

Definition 2 (Bayesian Nash Equilibrium). In any state $s = (\theta, \tau, \varepsilon)$, a strategy pair $(x^*, y^*) \in X^2$ is a type-homogeneous Bayesian Nash Equilibrium if:

$$\begin{cases} x^* \in \underset{x \in X}{\operatorname{argmax}} \quad p_{\theta|\theta}(\varepsilon) \cdot u_\theta(x, x^*) + p_{\tau|\theta}(\varepsilon) \cdot u_\theta(x, y^*) \\ y^* \in \underset{y \in X}{\operatorname{argmax}} \quad p_{\theta|\tau}(\varepsilon) \cdot u_\tau(y, x^*) + p_{\tau|\tau}(\varepsilon) \cdot u_\tau(y, y^*) \end{cases}$$

The set of Bayesian Nash Equilibria in population state $s = (\theta, \tau, \varepsilon)$, i.e. all solutions (x^*, y^*), is called $B^{NE}(s) \subseteq X^2$.

Let $\Pi_\theta(x, y, \varepsilon)$ be the average payoff obtained by individuals of type θ when they play $x \in X$ and when individuals of type τ play $y \in X$. Similarly, $\Pi_\tau(x, y, \varepsilon)$ is the average payoff obtained by individuals of type τ. We can express the average payoffs obtained by individuals of type θ and τ in function of the game payoffs:

$$\begin{cases} \Pi_\theta(x, y, \varepsilon) & = p_{\theta|\theta}(\varepsilon) \cdot \pi(x, x) + p_{\tau|\theta}(\varepsilon) \cdot \pi(x, y) \\ \Pi_\tau(x, y, \varepsilon) & = p_{\theta|\tau}(\varepsilon) \cdot \pi(y, x) + p_{\tau|\tau}(\varepsilon) \cdot \pi(y, y) \end{cases} \tag{1}$$

Now consider a population of residents with type θ. What happens when a small group of mutants of type τ "invade" the population? If the residents earn a greater payoff than the mutants, then the resident type θ can withstand a small-scale invasion of the type τ, and type θ is called evolutionarily stable against type τ.

Definition 3 (Evolutionarily stable preference). A type $\theta \in \Theta$ is evolutionarily stable against a type $\tau \in \Theta$ if there exists an $\bar{\varepsilon} > 0$ such that $\Pi_\theta(x^*, y^*, \varepsilon) > \Pi_\tau(x^*, y^*, \varepsilon)$ in all Nash equilibria (x^*, y^*) in all states $s = (\theta, \tau, \varepsilon)$ with $\varepsilon \in (0, \bar{\varepsilon})$. A type $\theta \in \Theta$ is evolutionarily stable if it is evolutionarily stable against all types $\tau \neq \theta \in \Theta$.

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

Definition 4 (Homo moralis and homo hamiltonensis). An individual is a *homo moralis* if her utility function is of the form:

$$u_\kappa(x, y) = (1 - \kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x) \tag{2}$$

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 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*[8].

Definition 5 (Hamiltonian strategies). *Hamiltonian strategies $x_\sigma \in X$ are the strategies played by homo hamiltonensis individuals when all residents are of this type. Formally:*

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

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 population of *homo hamiltonensis* and a small group of mutants that wish to enter the population. If the mutant type is not a "behavioral-alike"[9] to *homo hamiltonensis*, the individuals with the mutant type will always get a lower payoff than the *homo hamiltonensis* individuals. For example, 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.

In Alger and Weibull (2013), the residents all have the same type. But is this a required feature of the population? What happens when the population is more diverse? Is it possible to have a stable population comprised of residents with several types? We explore these questions in this paper.

2.2 A Population with n resident types

We expand the previous model by allowing for the presence of more than one resident type. In this adjusted setting, the population comprises individuals of n resident types (θ_1, θ_2,..., θ_n) and one mutant type θ_τ. If $n > 1$, i.e. if there are at least two resident types in the population, the population is called heterogeneous. The $(n+1)$ types and their respective shares define a population state $s = (\theta_1, \theta_2, ..., \theta_n, \theta_\tau, \lambda_1, \lambda_2, ..., \lambda_n, \varepsilon)$, where $\varepsilon \in (0,1)$ is the population share of θ_τ and for all $i \in [\![1..n]\!]$, $\lambda_i \in (0,1)$ is the share of the residents of type θ_i. Thus, the population share of θ_1 is λ_1, the population share of θ_2 is λ_2 and so on. We have:

$$\sum_{i=1}^{n} \lambda_i = 1 - \varepsilon \tag{3}$$

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

[9] 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:

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

Therefore, s could be described with only n population shares instead of all of them. For example in the classical setting, $s = (\theta, \tau, \varepsilon)$. In the case of two residents and one mutant, we will often use $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \varepsilon)$ with λ the relative share of θ_2 with respect to θ_1, i.e. $\lambda_2 = \lambda(1-\varepsilon)$ and $\lambda_1 = (1-\lambda)(1-\varepsilon)$.

For the sake of tractability, we introduce a new notation. For all $(i,j) \in [\![1..n]\!]^2$, the conditional probability that an individual being of type θ_i is matched with an individual of type θ_j is called p_{ji}.[10] In particular, for all $i \in [\![1..n]\!]$ we note p_{ii} the conditional probability that an individual being of type θ_i is matched with an individual of the same type θ_i. For mutants of type θ_τ, for all $i \in [\![1..n]\!]$ we note $p_{\tau i}$ the conditional probability that an individual being of type θ_i is matched with a mutant and $p_{i\tau}$ the conditional probability that a mutant is matched with an individual of type θ_i. Moreover, for all $i \in [\![1..n]\!]$, we note u_i the utility of residents of type θ_i and u_τ the utility of mutants.

Extending the concept of Bayesian Nash Equilibrium seen above to the case of $(n+1)$ types of individuals $(\theta_1, \theta_2, ..., \theta_n, \theta_\tau)$, we have:

Definition 6 (Bayesian Nash Equilibrium). In a population state $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \varepsilon)$, $(x^1, x^2, ..., x^n, x^\tau) \in X^{n+1}$ is a type-homogeneous Bayesian Nash equilibrium if:

$$\begin{cases} \forall i \in [\![1..n]\!]: \quad x^i \in \underset{x \in X}{\operatorname{argmax}} \quad \sum_{j=1}^{n} (p_{ji} \cdot u_i(x, x^j)) + p_{\tau i} \cdot u_i(x, x^\tau) \\ \qquad\qquad\qquad x^\tau \in \underset{x \in X}{\operatorname{argmax}} \quad \sum_{j=1}^{n} (p_{j\tau} \cdot u_\tau(x^j)) + p_{\tau\tau} \cdot u_\tau(x, x^\tau) \end{cases} \tag{4}$$

Assortative matching

The matching process is still random and exogenous. In the following, we first present some conditions that the matching process should satisfy to be well defined. Then, building on Bergstrom (2003), we introduce a novel, type-by-type assortment matrix function allowing for assortative matching in interactions between individuals of $(n+1)$ distinct types: the n resident type and the mutant type.

Let $I = ([\![1..n]\!] \cup \{\tau\})$. The conditional probabilities p_{ij} should satisfy the following matching conditions:

$$\forall\, i \in I: \quad \sum_{j \in I} p_{ji} = 1 \tag{5}$$

Matching conditions ensure that each individual is matched with another individual with probability one, i.e. nobody is left behind without a match.

The conditional probabilities should also satisfy the balancing conditions:

$$\forall\, (i,j) \in I^2: \quad \lambda_j \cdot p_{ij} = \lambda_i \cdot p_{ji} \tag{6}$$

Balancing conditions ensure the coherence of the matching process. They stipulate that 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

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

θ_i".

Now we are in position to generalize the notion of assortment function and assortativity for a population made of more than two types.

Definition 7 (Assortment matrix). In a population state $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \varepsilon)$, for all $(i,j) \in I^2$, let $\phi_{ij}(\lambda_1, ..., \lambda_n, \varepsilon)$ 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: $\phi_{ij}(\lambda_1, ..., \lambda_n, \varepsilon) = p_{ii} - p_{ij}$.
For all $(i,j) \in I^2$, $\phi_{ij} : (0,1)^{n+1} \to [-1,1]$.[11] This defines an exogenous assortment functions matrix:

$$\Phi = ((\phi_{ij}(\lambda_1, ..., \lambda_n, \varepsilon)))_{(i,j)\in I^2}$$

Extending the concept of assortment function, the assortment matrix embeds *homophily* effects, relating to the notion of distance in network economics (Currarini et al., 2009; Iijima and Kamada, 2017). Along the concept of *homophily*, this matrix allows accounting for the higher probability of interacting with similar others (Byrne, 1971; Lakin and Chartrand, 2003). 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 like economics, sociology, biology or management, with the possibility to calibrate its values empirically.

We now introduce a particular type of assortment matrix extending the classical 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.

Definition 8 (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.[12] In other words, we will say that Φ is *uniformly constant*[13] when, for all $(i,j,k,l) \in I^4$ such that $i \neq j$ and $k \neq l$:

$$\begin{cases} \phi_{ij} : (0,1)^{n+1} \to [-1,1] \quad \text{is} \quad \textit{constant}, \\ \phi_{ij}(\cdot) = \phi_{kl}(\cdot) \end{cases} \tag{7}$$

Note that the classical case of uniform random matching, where the matching process is not assortative, is a special case of uniformly-constant assortment where each assortment function is constant and equal to zero: $\Phi = ((0))_{(i,j)\in I^2}$.

[11] The assortment functions actually depend on n variables instead of $n+1$ because the sum of the population shares is equal to one. For instance, in the classical setting, the assortment function can be defined only in function of the population share of the mutant.

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

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

Following Alger and Weibull (2013), we assume that for all $(i,j) \in I^2$, $\phi_{ij}(\cdot)$ is continuous in ε (the share of mutants in the population) and converges as ε tends to zero.[14] We call assortativity $\sigma \in [0,1]$ the limit for all $i \in [\![1..n]\!]$ of $\phi_{\tau i}$ when ε tends to zero:[15]

$$\forall\, i \in [\![1..n]\!] : \quad \lim_{\varepsilon \to 0} \phi_{\tau i}(\lambda_1, ..., \lambda_n, \varepsilon) = \sigma \tag{8}$$

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 J[16] and the identity matrix I as follows:

$$\Phi = \sigma(J - I) \tag{9}$$

Matching probabilities

In the following, we will use the notation ϕ_{ij} to designate $\phi_{ij}(\lambda_1, ..., \lambda_n, \varepsilon)$, abstracting from the arguments of the assortment functions for simplicity. The above definition of assortment gives assortment conditions on the conditional probabilities p_{ij}:

$$\forall\, (i,j) \in I^2 : \quad \phi_{ij} = p_{ii} - p_{ij} \tag{10}$$

Knowing the assortment matrix Φ, we have a system of equations defined by matching conditions (5), balancing conditions (6) and assortment conditions (10). The system has a (unique) solution when the exogenous assortment matrix verifies some conditions that we call the assortment balancing conditions:[17]

$$\forall\, (i,j) \in I^2 : \quad \lambda_j \cdot \left[\left(\sum_{k \in I} \lambda_k \phi_{ik} \right) - \phi_{ij} \right] = \lambda_i \cdot \left[\left(\sum_{k \in I} \lambda_k \phi_{jk} \right) - \phi_{ji} \right] \tag{11}$$

The assortment balancing conditions ensure the coherence of the assortative matching. In the case of two types in the population, the conditions boil down to one equivalence condition: $\phi_{12}(\lambda_1, \lambda_2) = \phi_{21}(\lambda_1, \lambda_2)$. More details are available in AppendixA.

When the assortment matrix satisfies the assortment balancing conditions, 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 (11), the system defined by matching conditions (5), balancing conditions (6) and assortment conditions (10) has a unique solution:*

$$\forall (i,j) \in I^2 : \quad p_{ij} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \tag{12}$$

Proof. In AppendixC.1 □

[14] Note that the residents population shares $(\lambda_i)_{i \in (0,1)^n}$ change as ε tends to zero, and are hence not the same as in the original population state $s = (\theta_1, \theta_2, ..., \theta_n, \theta_\tau, \lambda_1, \lambda_2, ..., \lambda_n, \varepsilon)$.

[15] At the limit when ε goes to zero, we have for all $i \in [\![1..n]\!]$, $\phi_{\tau i}(\lambda_1, \lambda_2, ..., \lambda_n, 0) = p_{\tau\tau}$. Indeed, according to the balancing conditions, the probability of a resident to be matched with a mutant $p_{\tau i}$ is zero. Thus, the assortativity is independent of the resident types, and we have $\sigma \in [0,1]$.

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

[17] For convenience, we call λ_τ the population share of mutant, i.e. $\lambda_\tau = \varepsilon$.

Note that under uniform random matching, for all $(i,j) \in I^2$ $\phi_{ji} = 0$ and we obtain $p_{ij} = \lambda_i$, 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_{ii}:

$$\forall\, i \in I : \quad p_{ii} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} \tag{13}$$

The conditional probabilities p_{ii} 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.

The conditional probabilities for the case of two residents and one mutant are detailed in AppendixB.

We now derive a useful result for our analysis of evolutionary stability (section 4). Let $B^{NE}(s) \subseteq X^{n+1}$ denote the set of Bayesian Nash equilibria in population state $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \varepsilon)$, i.e. all solutions $(x^1, ..., x^n, x^\tau)$ of (4). This defines an equilibrium correspondence $B^{NE}(\theta_1, ..., \theta_n, \theta_\tau, \cdot) : (0,1)^{n+1} \rightrightarrows X^{n+1}$. This correspondence maps the population share of each type to the associated equilibria. Using the above definition of assortativity, it can be extended by continuity to $(0,1)^n \times [0,1)$ to cover the limit when the mutant share ε goes to zero. The following lemma is an generalization to the case of two types in Alger and Weibull (2013):

Lemma 1. *$B^{NE}(s)$ is compact for each $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \varepsilon) \in \Theta^{n+1} \times (0,1)^n \times [0,1)$. $B^{NE}(s) \neq \emptyset$ if for all $i \in I$ u_i are concave in their first arguments.*
The correspondence $B^{NE}(\theta_1, ..., \theta_n, \theta_\tau, \cdot) : (0,1)^n \times [0,1) \rightrightarrows X^{n+1}$ is upper hemi-continuous.

Proof. In AppendixC.2. □

3. Cohabitation of two residents

The core contribution of this paper consists of the consideration of the cohabitation of more than one resident type in the population. What happens then? If one type dominates the other, i.e. if individuals of type θ_1 get a higher payoff than individuals of type θ_2, it seems unlikely that θ_2 would survive. In evolutionary game dynamics, the evolution of strategies (and preferences) is dictated by what is called a replicator, and depends on the difference between each payoff obtained and the average payoff in the population. If the payoff of a given type is greater than the average payoff, then the population share of this type will increase. Thus, the two types should get the same payoff to coexist without one overcoming (or invading) one another. We call this condition Payoff Equality.

The Payoff Equality condition is similar to the concept of balanced configuration by Dekel et al. (2007). A configuration is a distribution of preferences in a population and the associated equilibrium. It is balanced when all types present receive the same fitness, i.e. the same payoff. In our case, the distribution of preferences is simply defined by the population share of each type. In the following section, we evaluate the conditions under which the Payoff Equality condition is satisfied.

3.1 Payoff Equality condition

Let's assume $(x^1, x^2) \in X^2$ is a BNE in the population state $s = (\theta_1, \theta_2, \lambda)$, with $\lambda \in (0,1)$ the share of θ_2. Noting Π_1 the average payoff of individuals θ_1 and Π_2 the average payoff of individuals

θ_2, the Payoff Equality condition is met when:

$$\Pi_1(x^1, x^2, \lambda) = \Pi_2(x^1, x^2, \lambda) \tag{14}$$

We can write the payoffs of θ_1 and θ_2 (Equation (1)) using the matching probabilities derived in (12)[18] and the notations $\pi^{ij} = \pi(x^i, x^j)$:[19]

$$\begin{aligned} \Pi_1(x^1, x^2, \lambda) \quad &= ((1-\lambda) + \lambda \cdot \phi_{12}) \cdot \pi^{11} + \lambda(1-\phi_{12}) \cdot \pi^{12} \\ \Pi_2(x^1, x^2, \lambda) \quad &= (1-\lambda)(1-\phi_{21}) \cdot \pi^{21} + (\lambda + (1-\lambda)\phi_{21}) \cdot \pi^{22} \end{aligned} \tag{15}$$

Under what conditions on the payoffs $(\pi^{11}, \pi^{12}, \pi^{21}, \pi^{22})$ and on $\lambda \in (0,1)$ are total payoffs equal?

Since we are in the context of only two residents, the assortment balancing condition (11) implies that $\phi_{12} = \phi_{21}$. We can then equate the expressions of the payoff functions of θ_1 and θ_2 (equation (15)), which gives, after rearranging, the following condition:

$$\lambda(1-\phi_{12})(\pi^{11} + \pi^{22} - \pi^{12} - \pi^{21}) = \pi^{11} - \pi^{21} - \phi_{12}(\pi^{22} - \pi^{21})$$

This equation can also be rewritten by reordering the terms to put forward the share of θ_1 in the population $(1-\lambda)$, as follows:

$$(1-\lambda)(1-\phi_{12})(\pi^{11} + \pi^{22} - \pi^{12} - \pi^{21}) = \pi^{22} - \pi^{12} - \phi_{12}(\pi^{11} - \pi^{12})$$

We define:

$$\begin{aligned} Q_\pi &= \pi^{11} - \pi^{21} - \phi_{12}(\pi^{22} - \pi^{21}) \\ R_\pi &= \pi^{22} - \pi^{12} - \phi_{12}(\pi^{11} - \pi^{12}) \\ S_\pi &= \pi^{11} + \pi^{22} - \pi^{12} - \pi^{21} \end{aligned} \tag{16}$$

Note that we have: $Q_\pi + R_\pi = (1-\phi_{12})S_\pi$.

We can rewrite the Payoff Equality condition to obtain two equivalent equations, one for λ and the other for $(1-\lambda)$:

$$\lambda(1-\phi_{12})S_\pi = Q_\pi \tag{17}$$

$$(1-\lambda)(1-\phi_{12})S_\pi = R_\pi \tag{18}$$

We have the following proposition:

Proposition 2 (Payoff Equality)**.** *Let $(x^1, x^2) \in X^2$ be a BNE in the population state $s = (\theta_1, \theta_2, \lambda)$, with $\lambda \in (0,1)$. The Payoff Equality condition $\Pi_1(x^1, x^2, \lambda) = \Pi_2(x^1, x^2, \lambda)$ is satisfied if and only if:*

1. *$Q_\pi = 0$ and $S_\pi = 0$, or*
2. *$Q_\pi \neq 0$ and $R_\pi \neq 0$ are of the same sign, and $\lambda(1-\phi_{12}) = Q_\pi / S_\pi$, or*
3. *$Q_\pi = 0, R_\pi = 0, S_\pi \neq 0$ and $\phi_{12} = 1$.*

Note that for all other possible cases, Payoff Equality is not satisfied.

[18] The matching probabilities with only two residents are detailed in the Appendix, Equation (B.2)

[19] Remember that the assortment functions depend on the population share λ, i.e. $\phi_{12} = \phi_{12}(\lambda)$ and $\phi_{21} = \phi_{21}(\lambda)$

Proof. In AppendixC.3

□

Here are a few examples for each case of the proposition:

1. (a) If $\pi^{12} = \pi^{11} = \pi^{22} = \pi^{21}$. For example, if $x^1 = x^2$ then all individuals play the same strategy. No matter the share of each type, they will get the same payoff.

   (b) If $\pi^{11} = \pi^{21}$, $\pi^{12} = \pi^{22}$, and if $\phi_{12} = 0$. All individuals obtain the same payoff when playing against an individual of type θ_1. They also obtain the same payoff when playing against a θ_2. Since there is no assortment, the matching probabilities are equal to the population share, i.e. all individuals have the same probabilities $(1 - \lambda)$ to play against a θ_1 and λ to play against a θ_2.

2. (a) If $\{\pi^{11} = \pi^{22}, \pi^{11} > \pi^{12}, \pi^{11} > \pi^{21}\}$ or if $\{\pi^{11} = \pi^{22}, \pi^{11} < \pi^{12}, \pi^{11} < \pi^{21}\}$, and if $\lambda = (\pi^{11} - \pi^{21})/S_\pi$ and $\phi_{12} \neq 1$. This is an interesting and somewhat counter-intuitive case because the assortment does not play a role in the equilibrium population share λ. In the special case where $\pi^{12} = \pi^{21}$, then λ should be equal to 0.5.

   (b) If $\{\pi^{12} = \pi^{21} = \pi^{11} \neq \pi^{22}\}$ or if $\{\pi^{11} \neq \pi^{22} = \pi^{12} = \pi^{21}\}$, and if $\lambda = -\phi_{12}/(1 - \phi_{12})$. In this case, the equilibrium population share does not depend on the payoffs. Note also that the assortment should be negative, $\phi_{12} < 0$, i.e. we are in a situation of heterophily.[20]

3. An assortment equal to 1 means that individuals θ_1 only meet individuals θ_1, getting π^{11}, while individuals θ_2 only meet individuals θ_2, getting π^{22}. If $Q_\pi = 0, R_\pi = 0$ and $\phi_{12} = 1$, then $\pi^{11} = \pi^{22}$ and all individuals get the same payoff.

4. If π^{11} and π^{12} are strictly greater than π^{21} and π^{22}, then there is no λ that satisfies Payoff Equality. No matter the population share, individuals θ_1 will always get more than individuals θ_2.

In any fitness game, Proposition 2 enables to determine if two strategies could be played or not by individuals of a heterogeneous population that satisfies Payoff Equality, just by computing Q_π, R_π and S_π.

In the standard framework of evolutionary game dynamics, the game is finite and individuals play pure strategies. At the equilibrium, the remaining strategies in the population should respect the Payoff Equality. Thus, Proposition 2 allows to quickly select or eliminate candidate strategies for an equilibrium in a population of two types. 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.

3.2 Cohabitation of two residents under uniformly-constant assortment

Having established the conditions for equality of the payoffs among two resident types in the population, we now consider the case of a uniformly-constant assortment matrix (cf. Definition 8), which is an extension of uniform random matching accounting for assortatively matched interactions. We show that *Hamiltonian strategies* (cf. Definition 5) play a key role in this context.

[20] See for instance Harrigan and Yap (2017) for a discussion of negative ties in networks.

We recall that $B^{NE}(s)$ is the set of solutions of the Bayesian Nash Equilibrium problem (Definition 6), with $s = (\theta_1, \theta_2, \lambda)$ the population state. Since we assume that the assortment function ϕ_{12} is constant, it is equal to the assortativity by continuity (cf. (9)): for all $\lambda \in (0, 1)$, $\phi_{12}(\lambda) = \sigma \in [0, 1]$. The following theorem proves that if individuals of a heterogeneous population play *Hamiltonian strategies*, then the Payoff Equality is satisfied.

Theorem 1 (Payoff Equality and Hamiltonian strategies)**.** *When the assortment function ϕ_{12} is constant, let $s = (\theta_1, \theta_2, \lambda)$ be a population where θ_1 individuals play x^1 , θ_2 individuals play x^2, such that $(x^1, x^2) \in B^{NE}(s) \subset X_\sigma^2$, and $\lambda = Q_\pi/(1-\sigma)S_\pi$ the share of θ_2 in the population.*

If $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, then the population $(\theta_1, \theta_2, \lambda)$ satisfies the Payoff Equality condition.

Proof. First, if $x^1 = x^2$ then all individuals play the same strategy and earn the same payoff. Now suppose that $x^1 \neq x^2$. Since θ_1 plays $x^1 \in X_\sigma$, θ_2 plays $x^2 \in X_\sigma$ with $x^1 \neq x^2$, and $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, we have:

$$\begin{cases} x^1 \in \underset{x \in X}{\operatorname{argmax}}\, u_\sigma(x, x^1) & \Rightarrow \quad \forall x \neq x^1 \in X,\ \pi(x^1, x^1) > (1-\sigma) \cdot \pi(x, x^1) + \sigma \cdot \pi(x, x) \\ x^2 \in \underset{y \in X}{\operatorname{argmax}}\, u_\sigma(y, x^2) & \Rightarrow \quad \forall y \neq x^2 \in X,\ \pi(x^2, x^2) > (1-\sigma) \cdot \pi(y, x^2) + \sigma \cdot \pi(y, y) \end{cases}$$

In particular, for $x = x^2$ and $y = x^1$, we have:

$$\begin{cases} \pi^{11} > (1-\sigma) \cdot \pi^{21} + \sigma \cdot \pi^{22} & \Rightarrow \quad Q_\pi > 0 \\ \pi^{22} > (1-\sigma) \cdot \pi^{12} + \sigma \cdot \pi^{11} & \Rightarrow \quad R_\pi > 0 \end{cases}$$

Consequently, since $\lambda = Q_\pi/((1-\sigma)S_\pi)$ by assumption, we are in case 2. of Proposition 2 and the Payoff Equality condition is satisfied. □

3.3 Some examples on finite games

We now look at several examples in two-strategies games to illustrate the Payoff Equality condition. We use the notation defined above $\pi^{ij} = \pi(x^i, x^j)$ to describe the payoff obtained by a player playing x^i when her opponent plays x^j. Moreover, we denote by π_{ij} the payoff when pure strategy i is played against pure strategy j. Also, we study in the following examples the case of a constant assortment function: for all $\lambda \in (0, 1)$, $\phi_{12}(\lambda) = \sigma$.

Coordination game

In the coordination game, players get reward only if they play the same strategy (i.e $\pi_{AA} > 0$, $\pi_{BB} > 0$ and $\pi_{AB} = \pi_{BA} = 0$). Suppose that individuals θ_1 are committed to strategy A while individuals θ_2 are committed to strategy B. Examples of such preferences include $u_{\theta_1}(x, y) = -(x - x_A)^2$ and $u_{\theta_2}(x, y) = -(x - x_B)^2$ where x_A and x_B denotes strategies A and B. When an individual θ_1 is matched with another individual of the same type θ_1, she gets the payoff $\pi^{AA} = \pi_{AA}$, and when she is matched with an individual θ_2, she gets $\pi^{AB} = \pi_{AB} = 0$. Thus, $S_\pi = \pi_{AA} + \pi_{BB} > 0$, $Q_\pi = \pi_{AA} - \sigma\pi_{BB}$ and $R_\pi = \pi_{BB} - \sigma\pi_{AA}$.

If $\pi_{AA} = \pi_{BB}$, then there are two possibilities to satisfy Payoff Equality. First, if $\sigma = 1$, for any population share $\lambda \in (0, 1)$, individuals only play against similar others, earning $\pi_{AA} = \pi_{BB}$ (we are then in case 3. of Proposition 2). Second, if $\sigma < 1$ and $\lambda = (\pi_{AA} - \sigma\pi_{BB})/[(1-\sigma)(\pi_{AA} + \pi_{BB})]$, we are in case 2. of Proposition 2. For instance under uniform random matching, $\sigma = 0$, and with $\lambda = 1/2$ all individuals earn $\pi_{AA}/2 = \pi_{BB}/2$ playing half of the time against similar others.

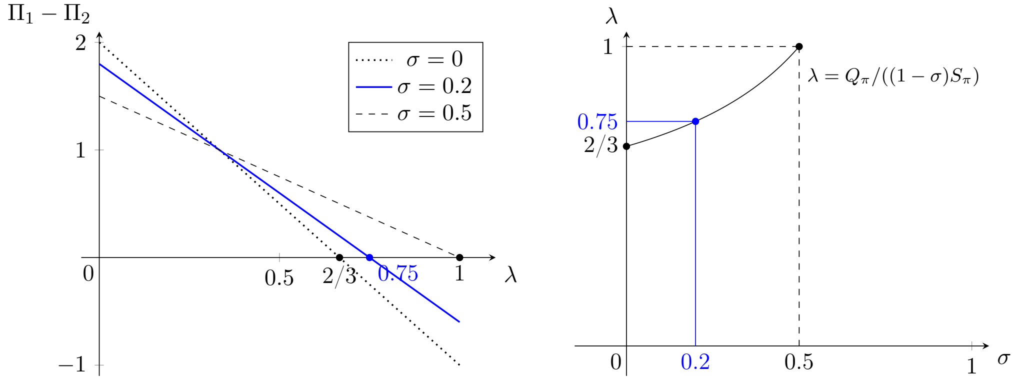
If $\pi_{AA} > \pi_{BB}$, then there exists a heterogeneous population that satisfies Payoff Equality only if $\sigma < \pi_{BB}/\pi_{AA}$, and $\lambda = (\pi_{AA} - \sigma\pi_{BB})/[(1-\sigma)(\pi_{AA} + \pi_{BB})]$ (case 2. of Proposition 2). If $\sigma \geq \pi_{BB}/\pi_{AA}$, then individuals θ_1 always earn more than individuals θ_2.

Reciprocally, if $\pi_{AA} < \pi_{BB}$, then there exists a heterogeneous population that satisfies Payoff Equality only if $\sigma < \pi_{AA}/\pi_{BB}$ and $\lambda = (\pi_{AA} - \sigma\pi_{BB})/[(1-\sigma)(\pi_{AA} + \pi_{BB})]$ (case 2. of Proposition 2).

Table 1: Coordination game example

	A	B
A	$(2,2)$	$(0,0)$
B	$(0,0)$	$(1,1)$

For example, let $\pi_{AA} = 2$ and $\pi_{BB} = 1$. A heterogeneous population satisfies Payoff Equality when $\sigma < 0.5$ and $\lambda = (2-\sigma)/[3(1-\sigma)]$. When $\sigma \geq 0.5$, individuals θ_1 always earn more than individuals θ_2. With $\sigma = 0.2$, $\lambda = 0.75$, and individuals θ_1 and θ_2 get the same total payoff $\Pi = 0.8$. Under uniform random matching, $\sigma = 0$, we get the classical mixed Nash equilibrium with $\lambda = 2/3$. The results are illustrated in Figure 1.



(a) Payoff difference

(b) λ satisfying PE as a function of σ

Figure 1: Payoff Equality (PE) in a coordination game between individuals θ_1 playing A and individuals θ_2 playing B (with λ the share of θ_2).

Prisoner's dilemma

In the prisoner's dilemma, players can either cooperate (C) or defect (D), getting $\pi_{CD} < \pi_{DD} < \pi_{CC} < \pi_{DC}$. Suppose a population of *homo kantiensis* (θ_1, playing C) and *homo oeconomicus* (θ_2, playing D). We consider three different examples: (a) $S_\pi < 0$, (b) $S_\pi = 0$ and (c) $S_\pi > 0$.

Table 2: Prisoner's dilemma examples

(a)

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

(b)

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

(c)

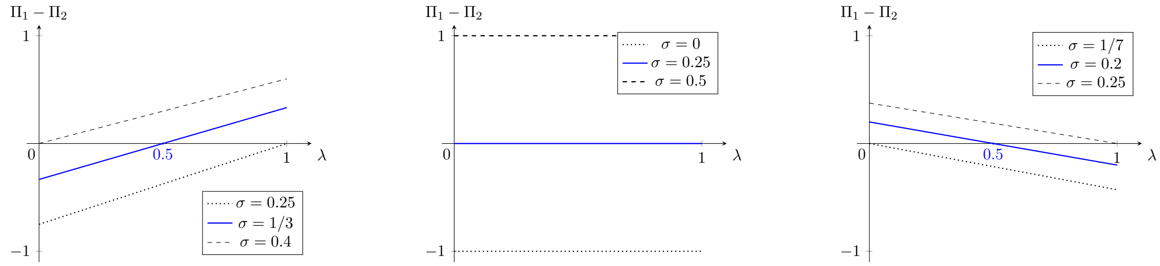
	C	D
C	$(4,4)$	$(0,4.5)$
D	$(4.5,0)$	$(1,1)$

(a) First, let $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 6$. We then have $S_\pi = -1 < 0$, $Q_\pi = -2 + 5\sigma$ and $R_\pi = 1 - 4\sigma$. Thus, there exists a heterogeneous population satisfying Payoff

Equality when $0.25 < \sigma < 0.4$ (see Figure 2a). With $\sigma = 1/3$, then $\lambda = 0.5$ and *homo kantiensis* and *homo oeconomicus* co-exist and get the same payoff equal to $\Pi = 8/3$. If the assortment is too low (e.g. under uniform random matching), 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 assortment value consistent with Payoff Equality is $\sigma = 0.25$ (see Figure 2b). But then, for any population share λ, *homo kantiensis* and *homo oeconomicus* earn the same payoff.

(c) Finally, let $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 4.5$. We have $S_\pi = 0.5 > 0$, $Q_\pi = -0.5 + 3.5\sigma$ and $R_\pi = 1 - 4\sigma$. Thus, there exists a heterogeneous population satisfying Payoff Equality when $1/7 < \sigma < 0.25$ (see Figure 2c). For example, when $\sigma = 0.2$, then $\lambda = 0.5$ and *homo kantiensis* and *homo oeconomicus* live together and get the same payoff equal to $\Pi = 2.4$. As above, the assortment plays a key role: if too low or too high, one type will predominate.



(a) $\pi_{CC} = 4$, $\pi_{CD} = 0$, $\pi_{DC} = 6$, $\pi_{DD} = 1$

(b) $\pi_{CC} = 4$, $\pi_{CD} = 0$, $\pi_{DC} = 5$, $\pi_{DD} = 1$

(c) $\pi_{CC} = 4$, $\pi_{CD} = 0$, $\pi_{DC} = 4.5$, $\pi_{DD} = 1$

Figure 2: Payoff difference in prisoner's dilemma between *homo kantiensis* (Π_1) and *homo oeconomicus* (Π_2)

Bergstrom (2003) and Allen and Nowak (2015) study the evolution of cooperative strategy under assortative matching in social dilemmas in an evolutionary game dynamics framework. Since at equilibrium strategies must respect the Payoff Equality, their results are consistent with ours. Allen and Nowak (2015) investigate a game between relatives (with constant assortment), finding that assortment favors cooperation in prisoner's dilemma. In a simplified version of the game, where payoffs ($\pi_{CD} = -c$, $\pi_{DD} = 0$, $\pi_{CC} = b - c$ and $\pi_{DC} = b$ with $b > c > 0$), they highlight that cooperation is favored when a condition similar to the Hamilton's rule is satisfied.[21] We obtain an analogous condition in this simplified game. Cooperation will outperform defection when $b\sigma > c$ (Case (b) of the example, i.e. $S_\pi = 0$). This condition is also derived by Bergstrom (2003) in prisoner's dilemma games with additive payoffs ($S_\pi = 0$).

4. Evolutionary stability

After having assessed the conditions under which individuals of two types can coexist in a population in Section 3, we study the evolutionary stability of this heterogeneous population against mutant invasions.

[21] 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).

4.1 Evolutionarily stable population

We extend the concept of *evolutionarily stable preference* (Alger and Weibull, 2013) introducing heterogeneity in resident individuals. An *evolutionarily stable population* should respect two conditions. First, all types should earn the same payoff to coexist, i.e. they satisfy the Payoff Equality condition (cf. (14)). Second, the population must resist the invasion of any other type, i.e. resident individuals earn a greater payoff than a small group of mutants. Formally:

Definition 9 (Evolutionarily stable population). A population made of n resident types $s = (\theta_1, ..., \theta_n, \lambda_1^\circ, ..., \lambda_n^\circ)$ is evolutionarily stable against a mutant type $\theta_\tau \in \Theta$ such that for all $i \in [\![1..n]\!]$ $\theta_\tau \neq \theta_i$ if:

1. Resident individuals earn the same payoff when there is no mutant: for all $(i, j) \in [\![1..n]\!]^2$, $\Pi_i(x^1, ..., x^n, \lambda_1^\circ, ..., \lambda_n^\circ) = \Pi_j(x^1, ..., x^n, \lambda_1^\circ, ..., \lambda_n^\circ)$ in all Bayesian Nash equilibria $(x^1, ..., x^n)$ in the population state $s = (\theta_1, ..., \theta_n, \lambda_1^\circ, ..., \lambda_n^\circ)$;
2. Resident individuals earn a greater payoff than a small share of mutants: there exists an $\bar{\varepsilon} > 0$ such that for all $i \in [\![1..n]\!]$: $\Pi_i(x^1, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon) > \Pi_\tau(x^1, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon)$ in all Bayesian Nash equilibria $(x^1, ..., x^n, x^\tau)$ in all states $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \varepsilon)$ with $\varepsilon \in (0, \bar{\varepsilon})$ and for all $i \in [\![1..n]\!]$ $|\lambda_i - \lambda_i^\circ| < \bar{\varepsilon}$.

Moreover, a population is evolutionarily stable if it is evolutionarily stable against all types $\theta_\tau \in \Theta$ such that for all $i \in [\![1..n]\!]$, $\theta_\tau \neq \theta_i$.

This definition is similar to the concept of *evolutionarily stable configuration* 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 payoff, 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 is 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. In the following, we will focus on the case of two resident types.

4.2 Evolutionary stability in finite game with uniformly-constant assortment

In this section, we consider the case of a uniformly-constant assortment matrix. Note that, by continuity, all non-diagonal elements of the assortment matrix are equal to the assortativity. In other words, for any population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \varepsilon)$:[22]

$$\forall (i, j) \in \{1, 2, \tau\}^2, i \neq j : \quad \phi_{ij} = \sigma \in [0, 1]$$

We study 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, i.e. $A = ((\pi_{ij}))_{(i,j) \in \{1,2\}^2}$. When players are permitted to used mixed strategies, the strategy set X is the segment $\Delta = \{z \in$

[22] λ is the relative share of θ_2 with respect to θ_1, i.e. $\lambda_1 = (1 - \lambda)(1 - \varepsilon)$ and $\lambda_2 = \lambda(1 - \varepsilon)$.

$\mathbb{R}_+^2 : z_1 + z_2 = 1\}$. The payoff obtained by an individual playing strategy $x \in X = \Delta$ when matched with an individual playing $y \in X$ is then: $\pi(x, y) = xAy$, where $\pi : X^2 \to \mathbb{R}$ is a bilinear function.

We first introduce some lemmas that will be useful to derive our main results. Let $(x^1, x^2, x^\tau) \in X^3$ be a Bayesian Nash equilibrium (cf. Definition 6) of the population $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \varepsilon)$. 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 when $x^1 \neq x^2$. For this purpose, let $\alpha_1, \alpha_2, \in [0, 1]$ be the probabilities that θ_1, θ_2 individuals attach to the first pure strategy: $x^1 = (\alpha_1, 1-\alpha_1)$ and $x^2 = (\alpha_2, 1-\alpha_2)$. Then, there exists $\gamma \in \mathbb{R}$ such that $x^\tau = \gamma x^1 + (1-\gamma)x^2 = (\gamma\alpha_1 + (1-\gamma)\alpha_2, 1 - \gamma\alpha_1 - (1-\gamma)\alpha_2)$.[23] For instance, when the mutants play like individuals θ_1, $\gamma = 1$, and when they play like individuals θ_2, $\gamma = 0$.

Suppose that the population $s = (\theta_1, \theta_2, \lambda)$ satisfies the Payoff Equality condition: $\Pi_1(x^1, x^2, \lambda) = \Pi_2(x^1, x^2, \lambda) = \Pi_\theta$. When $x^1 \neq x^2$, we can write the differences in payoff between the residents and the mutant at the limit (when ε goes to zero):

Lemma 2 (Difference in payoffs)**.** *For a population $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \varepsilon)$ engaged in a symmetric 2×2 fitness game such that the Payoff Equality condition is satisfied when $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, the difference in payoffs between the residents playing $x^1 \neq x^2$ and the mutant θ_τ at the limit (when ε goes to zero) is:*

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

Proof. In AppendixC.4. □

When studying the sign of this difference in payoffs, it is useful to understand what is the sign of $\gamma(1-\gamma)$. Without loss of generality and by symmetry we can assume $\alpha_2 < \alpha_1$, i.e. individuals θ_1 play the pure strategy 1 with a greater probability than individuals θ_2. Let α_τ be the probability that τ attaches to pure strategy 1, $\alpha_\tau = \gamma\alpha_1 + (1-\gamma)\alpha_2$. If $\alpha_\tau \in (\alpha_2, \alpha_1)$, 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 Fig. 3).

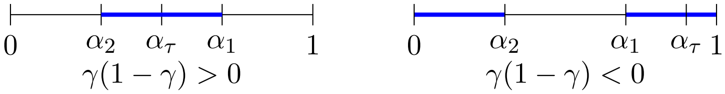


Figure 3: Sign of $\gamma(1-\gamma)$ depending on the probabilities attached to the first pure strategy

Finally, we also need to introduce one of the results of Alger and Weibull (2013). This result describes the strategies played by *homo hamiltonensis* (and more generally by *homo moralis*) in symmetric 2×2 fitness game. We recall that X_σ is the set of fixed-points of *homo hamiltonensis*, i.e. the strategies that *homo hamiltonensis* individuals play when they are alone in the population (called *Hamiltonian strategies*). As we show below, there is a close link between *Hamiltonian strategies*, and the strategies played by a heterogeneous *evolutionarily stable population* in the case of uniformly-constant assortment.

Lemma 3 (Proposition 2 of Alger and Weibull (2013))**.** *Let*

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

[23] The only case when this decomposition does not exist is when $\alpha_1 = \alpha_2$.

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*

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

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

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

We omit in what follows the case where $\sigma = 0$ (uniform random matching) to focus on the situations with assortative matching (for an analysis of the random matching case see for instance Dekel et al., 2007). Recall that $B^{NE}(s)$ is the set of solutions of the Bayesian Nash Equilibrium problem (6), with $s = (\theta_1, \theta_2, \lambda)$ the population state. 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. The following proposition details when two resident types cannot be part of an *evolutionarily stable population*:

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 AppendixC.5.

□

The proposition shows that if one resident does not play a *Hamiltonian strategy* when the type set is rich in a symmetric 2×2 fitness game under uniformly-constant and strictly positive assortment, then the population is not evolutionarily stable.[24]

But what happens when the two residents play *Hamiltonian strategies*? Could this population be evolutionarily stable? The answer is yes, when X_σ is a singleton and when $X_\sigma = \{0, 1\}$.[25] Let Θ_{12} be the set of mutants τ that are behaviorally indistinguishable from residents θ_1 and θ_2:

$$\Theta_{12} = \left\{\theta_\tau \in \Theta : \exists\, x \in X \,\text{such that}\, (x^1, x, x) \,\text{or}\, (x, x^2, x) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, 0)\right\}$$

We have the following theorem:

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*

[24] The proof develops a stronger argument than "not evolutionarily stable". It shows that there exists a mutant type such that the mutants earn a strictly greater payoff than the residents in all Bayesian Nash equilibria in a neighborhood of the entry point. Alger and Weibull (2013) call this property *evolutionary unstability*.

[25] In the last possible case, $X_\sigma = [0, 1]$, any strategy gives the same payoff meaning that there does not exist an *evolutionarily stable population*, even for single-type population

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. 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 from Theorem 1 because $\lambda = Q_\pi/((1-\sigma)S_\pi)$ by assumption. Then, from Lemma 2, we know that the difference in payoffs between the residents and mutants θ_τ at the limit is:

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

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 $S_\pi = (\alpha_1 - \alpha_2)^2 \left(\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21}\right)$ (C.6), we have $\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21} > 0$. Thus, we are in case 1 of Lemma 3, 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 $\langle (x_t^1, x_t^2, x_t^\tau) \rangle_{t \in \mathbb{N}}$ 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 $\Pi_2(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t) > \Pi_\tau(x_t^1, x_t^2, x_t^\tau, \lambda_t, \varepsilon_t)$. □

This theorem proves the existence of a heterogeneous *evolutionarily stable population* in all symmetric 2×2 fitness games under uniformly-constant and strictly positive assortment as long as it is possible to find two distinct types that are playing the same *Hamiltonian strategy.*[26] However, a heterogeneous *evolutionarily stable population* of two types playing diverse strategies is only possible when $X_\sigma = \{0, 1\}$.

Even though the strategies played by two types of an *evolutionarily stable population* are *Hamiltonian strategies*, there are differences between a heterogeneous *evolutionarily stable population* and a population of *homo hamiltonensis.* We discuss these differences in section 5.1. Moreover, this link with *Hamiltonian strategies* is only valid under uniformly-constant assortment matrix. We discuss the non uniformly-constant assortment case in section 5.2. In the following section, we focus on the case where two types play diverse strategies, studying the same examples as in section 3.3.

4.3 Some examples of finite games

In the following examples, we study the evolutionary stability of a heterogeneous population of two types playing diverse strategies, under uniformly-constant assortment: for all $(i, j) \in I^2$ such that $i \neq j$ and for all $(\lambda, \epsilon) \in (0, 1) \times [0, 1)$, $\phi_{ij}(\lambda, \epsilon) = \sigma$.

Coordination game

In the coordination game we considered in section 3.3, we had $\pi_{AA} = 2$ and $\pi_{BB} = 1$. For all $i \in \{1, 2, \tau\}$, let α_i be the probability that individuals θ_1, θ_2 and τ attach to strategy A.

Let individuals θ_1 be committed to strategy A ($\alpha_1 = 1$) and individuals θ_2 to strategy B ($\alpha_2 = 0$). We saw that it was possible to have Payoff Equality when $\sigma < 0.5$. For instance, for an assortment $\sigma = 0.2$ and a population share $\lambda = 0.75$, individuals θ_1 and θ_2 get the same payoff equal to $\Pi_\theta = 0.8$. Moreover, since the residents play the pure strategies, we have $S_\pi = \pi_{AA} + \pi_{BB} = 3 > 0$ and $\alpha_\tau = \gamma$ ($\alpha_\tau = \gamma\alpha_1 + (1-\gamma)\alpha_2$). From Lemma 2, we know that the difference in payoffs between the residents and the mutant at the limit is: $\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)S_\pi > 0$. As illustrated in Figure 4a, the difference in payoffs is strictly positive for all $\gamma \in (0, 1)$. Therefore, using similar arguments than in the proof of Theorem 2, we can conclude that this population is evolutionarily stable.

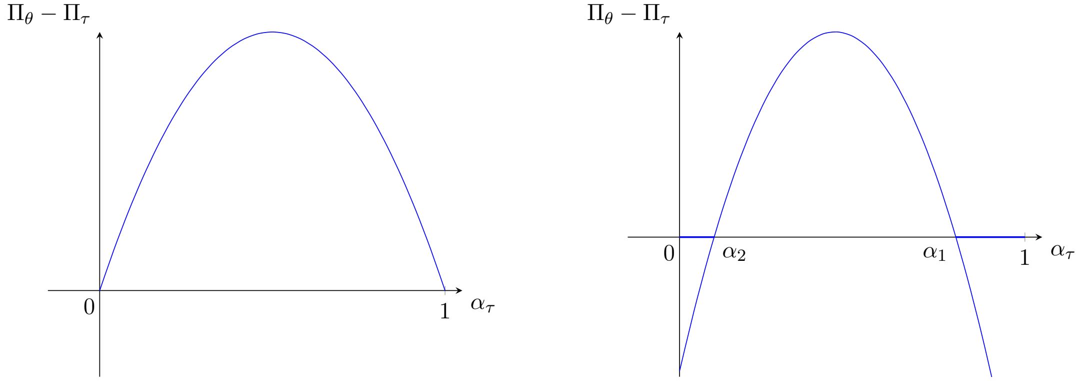
However, if the residents do not play the pure strategies, then the population is not evolutionarily stable. For instance, let $\alpha_1 = 0.8$ and $\alpha_2 = 0.1$. With the same assortment $\sigma = 0.2$ and population share $\lambda = 0.75$, individuals θ_1 and θ_2 get the same payoff. Yet, as illustrated in Figure 4b, if the mutants play $\alpha_\tau \in [0, 0.1) \cup (0.8, 1]$, then they will earn more than the residents at the limit since $\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)S_\pi$ with $\gamma = (\alpha_\tau - \alpha_2)/(\alpha_1 - \alpha_2)$.

More generally, in a coordination game under uniformly-constant assortment where $\pi_{AA} > \pi_{BB}$, there exists an *evolutionarily stable population* of two types playing diverse strategies if and only if each resident is committed to a pure strategy and $\sigma < \pi_{BB}/\pi_{AA}$. In this range, the Payoff Equality is satisfied with $\lambda = (\pi_{AA} - \sigma\pi_{BB})/\left[(1-\sigma)(\pi_{AA} + \pi_{BB})\right]$, and any mutant would earn less since $S_\pi > 0$.

Prisoner's dilemma

We study the evolutionarily stability of a population made of individuals *homo oeconomicus* that always defect and of individuals *homo kantiensis* that always cooperate, going back to our three

[26] As mentioned above, the only exception is when $X_\sigma = [0, 1]$, any strategy gives the same payoff, and no *evolutionarily stable population* exists, even for single-type population.



(a) Evolutionarily stable population
($\alpha_1 = 1$, $\alpha_2 = 0$)

(b) Non evolutionarily-stable population
($\alpha_1 = 0.8$, $\alpha_2 = 0.1$)

Figure 4: Payoff difference between a resident population and the mutants in coordination game ($\pi_{AA} = 2$, $\pi_{BB} = 1$), when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under uniformly-constant assortment $\sigma = 0.2$.

examples: (a) $S_\pi < 0$, (b) $S_\pi = 0$ and (c) $S_\pi > 0$.

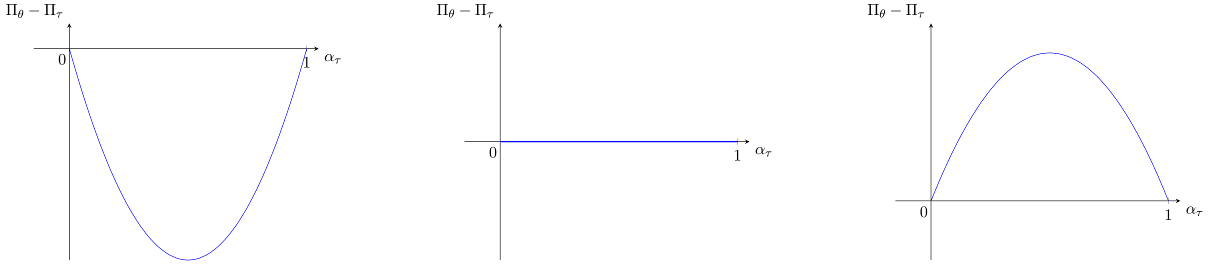
(a) First, let $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 6$.
With a uniformly-constant assortment $\sigma = 1/3$, then with $\lambda = 0.5$ the population satisfies the Payoff Equality condition and $\Pi_\theta = 8/3$.
However, we have $S_\pi = -1 < 0$, and since the difference in payoffs between the residents and the mutant at the limit is: $\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)S_\pi$ (Lemma 2), any mutant would earn more than the residents at the limit as illustrated in Figure 5a. Hence we can conclude that the population is not evolutionarily stable.

(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 $R_\pi = 1 - 4\sigma$. This is an additive game ($S_\pi = 0$), and as discussed in section 3.3, the only uniformly-constant assortment allowing Payoff Equality is $\sigma = 0.25$. With this value, any $\lambda \in (0, 1)$ satisfies Payoff Equality. On the other hand, any mutant would also earn the same payoff (see Figure 5b). In fact, this case contradicts the assumption that β_σ is a singleton in Theorem 2 since any strategy is a best-reply to another for *homo hamiltonensis*.

(c) Finally, let $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 4.5$.
With a uniformly-constant assortment $\sigma = 0.2$, then with $\lambda = 0.5$ the population satisfies the Payoff Equality condition and $\Pi_\theta = 2.4$.
Moreover, we have $S_\pi = 0.5 > 0$, and the difference in payoffs between the residents and the mutants at the limit is: $\Pi_\theta - \Pi_\tau = \sigma\gamma(1-\gamma)S_\pi$ (Lemma 2). Thus, for all $\gamma \in (0, 1)$, $\Pi_\theta - \Pi_\tau > 0$ (see Figure 5c) and, using similar arguments than in Theorem 2, we can conclude that this population is evolutionarily stable.

The following proposition details when a heterogeneous *evolutionarily stable population* of two types playing diverse strategies (cooperate and defect) in the prisoner's dilemma under uniformly-constant assortment exists:

Proposition 4 (Evolutionary stability of *homo oeconomicus* and *homo kantiensis*)**.** *In a prisoner's dilemma under uniformly-constant assortment where* Θ *is rich, there exists a heterogeneous evolutionarily stable population of homo oeconomicus and homo kantiensis if and only if* $\pi_{CC} + \pi_{DD} - \pi_{DC} - \pi_{CD} > 0$ *and* $(\pi_{DC} - \pi_{CC})/(\pi_{DC} - \pi_{DD}) < \sigma < (\pi_{DD} - \pi_{CD})/(\pi_{CC} - \pi_{CD})$.



(a) Non evolutionarily-stable ($\pi_{CC} = 4$, $\pi_{CD} = 0$, $\pi_{DC} = 6$, $\pi_{DD} = 1$, $\sigma = 1/3$, $\lambda = 0.5$)

(b) Non evolutionarily-stable ($\pi_{CC} = 4$, $\pi_{CD} = 0$, $\pi_{DC} = 5$, $\pi_{DD} = 1$, $\sigma = 0.25$)

(c) Evolutionarily stable ($\pi_{CC} = 4$, $\pi_{CD} = 0$, $\pi_{DC} = 4.5$, $\pi_{DD} = 1$, $\sigma = 0.2$, $\lambda = 0.5$)

Figure 5: Payoff difference between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoner's dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under uniformly-constant assortment.

Proof. In AppendixC.6.

□

We will further discuss this result in section 5, but it is worth insisting on the importance of assortative matching in making this heterogeneous population possible. It is only when the assortment is strictly positive that individuals *homo kantiensis* can survive without being dominated by individuals *homo oeconomicus*. Hence, assortment is critical to better understand cooperative behaviors, a result already highlighted by Eshel and Cavalli-Sforza (1982), Bergstrom (2003) or Allen and Nowak (2015) among others.[27]

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 look at the case of non uniformly-constant assortment using a few examples. Finally, we discuss what determines the types of preferences favored by evolution in our framework.

5.1 Heterogeneous population vs homogeneous single-type resident 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.

[27] Cooperative behaviors can also arise in repeated games thanks to reciprocity (see e.g. Nowak and Sigmund, 2005; Bowles and Gintis, 2004) and in public good games when participation is optional (Hauert et al., 2002).

However, such a population is not evolutionarily stable in most cases.[28] 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:

$$x_h \in \underset{x\in X}{\operatorname{argmax}}\{p_{11}\left((1-\sigma)\pi(x,x_h)+\sigma\pi(x,x)\right)+p_{21}\left((1-\sigma)\pi(x,x_\sigma^2)+\sigma\pi(x,x)\right)\} \tag{19}$$

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 (19). 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 (19) when:

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

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.[29] Therefore, (x_σ^1, x_σ^2) 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_σ^1 and x_σ^2. Can they also play a mixed strategy? Let $x_h = (\alpha_h, 1-\alpha_h) = \alpha_h x_\sigma^1 + (1-\alpha_h)x_\sigma^2$ a mixed strategy ($\alpha_h \in (0,1)$), x_h is solution of (19) when:

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

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_σ^2) 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.[30] In other words, *homo hamiltonensis* individuals cannot be part of a heterogeneous

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

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

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

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 2). 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. On the other hand, proposition 4 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 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 $\pi_{CC} + \pi_{DD} - \pi_{DC} - \pi_{CD} < 0$ but not when $\pi_{CC} + \pi_{DD} - \pi_{DC} - \pi_{CD} > 0$ (Lemma 3). 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.[31] 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 The case of state-dependent assortment

Throughout most of this paper, we assumed a uniformly-constant assortment matrix highlighting some relevant results. In this section, we discuss the more general case of non uniformly-constant assortment studying the same prisoner's dilemma as in section 3.3 and 4.3. We find that a non uniformly-constant assortment not only allows diverse *evolutionarily stable populations*, but also render these population more robust to mutant invasion.

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

Motivation

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).[32] Therefore, going beyond the classical case of an assortative matching that is uniform across all types in the population and independent of the share in the population, we evaluate the impact of relaxing the assumption of a uniformly constant assortment on the equilibria and their evolutionary stability.

Payoffs under non uniformly-constant assortment

We first write the payoffs of each type at the limit when the mutant share goes to zero to show the differences with the uniformly-constant assortment case. Let $(x^1, x^2, x^\tau) \in B^{NE}(s)$, with $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ the population state. We note $\Pi_1 = \Pi_1(x^1, x^2, x^\tau, \lambda, 0)$, $\Pi_2 = \Pi_2(x^1, x^2, x^\tau, \lambda, 0)$ and $\Pi_\tau = \Pi_\tau(x^1, x^2, x^\tau, \lambda, 0)$ the average payoffs obtained by individuals of types θ_1, θ_2 and τ at the limit (i.e. when the share of the mutant tends to zero). Moreover, we set $\phi_{12} = \phi_{12}(\lambda, 0) = \phi_{21}(\lambda, 0)$ for the assortment between types θ_1 and θ_2 at the limit. Finally, we call $\Gamma = \lim_{\varepsilon \to 0}(\phi_{\tau 1} - \phi_{\tau 2})/\varepsilon$. We can then express the payoffs of each type at the limit as a function of the population share and assortment functions using equations (B.2):

$$\begin{cases} \Pi_1 = (1 - \lambda + \lambda\phi_{12}) \cdot \pi^{11} + \lambda(1 - \phi_{12}) \cdot \pi^{12} \\ \Pi_2 = (1 - \lambda)(1 - \phi_{12}) \cdot \pi^{21} + (\lambda + (1 - \lambda)\phi_{12}) \cdot \pi^{22} \\ \Pi_\tau = ((1 - \lambda)(1 - \sigma) - \lambda(1 - \lambda)\Gamma) \cdot \pi^{\tau 1} + (\lambda(1 - \sigma) + \lambda(1 - \lambda)\Gamma) \cdot \pi^{\tau 2} + \sigma \cdot \pi^{\tau\tau} \end{cases}$$

In the uniformly-constant assortment case, we had $\phi_{12} = \sigma$ and $\Gamma = 0$. As discussed in AppendixB, the limit Γ can be interpreted as the matching probability difference between mutants and residents of the two types: $\Gamma = \lim_{\varepsilon \to 0}(p_{\tau 2} - p_{\tau 1})/\varepsilon$. 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 $\Gamma \neq 0$. Moreover, when the assortment functions $\phi_{\tau 1}$ and $\phi_{\tau 2}$ are differentiable in $\varepsilon = 0$, we have $\Gamma = \phi'_{\tau 1}(\lambda, 0) - \phi'_{\tau 2}(\lambda, 0)$[33]. Therefore, Γ is the marginal assortment difference between mutants and residents of the two types.

Consider the case of a prisoner's dilemma, suppose that the residents satisfy the Payoff Equality and for all $i \in \{1, 2, \tau\}$, let $\alpha_i \in [0, 1]$ be the probability that individuals θ_1, θ_2 and τ attach to the strategy "cooperate". If $\alpha_1 \neq \alpha_2$, there exists $\gamma \in \mathbb{R}$ such that $\alpha_\tau = \gamma\alpha_1 + (1 - \gamma)\alpha_2$ (cf. section 4.2). It is then possible to write the difference in payoff between the residents and the mutants when the share of the mutant goes to zero (see the proof of Lemma 2 in AppendixC.4 for detailed calculations):

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

On the survival of selfish and moral individuals

We consider the example studied above (Sections 3.3 and 4.3). The objective is to analyze if non uniformly-constant assortment enables the evolutionary stability of a heterogeneous population made

[32] 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.

[33] Noting $\phi'_{\tau i}(\lambda, 0) = \partial\phi_{\tau i}(\lambda, 0)/\partial\varepsilon$

of *homo kantiensis* (θ_1) and *homo oeconomicus* (θ_2). Given the great number of cases offered by the relaxation of the uniformly-constant assortment hypothesis, we consider here a specific case:

1. We suppose that $\phi_{12}(\lambda, 0) = 0.24\lambda + 0.08$. Thus, when the share of *homo oeconomicus* λ goes to zero, $\phi_{12}(0,0) = \phi_{21}(0,0) = 0.08$. This means that when the type θ_2 individuals (*homo oeconomicus*) are mutants in a population made of two types $s = (\theta_1, \theta_2, 0)$, their assortativity is 0.08. In other words, for $s = (\theta_1, \theta_2, 0)$, the probability for a *homo oeconomicus* to meet another *homo oeconomicus* is $p_{22} = 0.08$. Reciprocally, when θ_1 (*homo kantiensis*) is a mutant in a population made of two types $s = (\theta_1, \theta_2, 1)$, its assortativity is 0.32. Consequently, the shape of $\phi_{12}(\cdot, 0)$ 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. We also suppose that $\Gamma = (1-\sigma)/\lambda$. This shape allows $p_{1\tau}$ and $p_{2\tau}$ to belong to $[0,1]$. Moreover, it means that $p_{1\tau} = 0$ and $p_{2\tau} = 1 - \sigma$, i.e. a mutant either meet a *homo oeconomicus* or another mutant, which increases the likelihood that the population of *homo kantiensis* and *homo oeconomicus* is evolutionarily stable.

With these assumptions, we find for each case previously studied an *evolutionarily stable population* of *homo kantiensis* and *homo oeconomicus*. This contrasts with the case of a uniformly-constant assortment, in which there is no *evolutionarily stable population* when S_π is negative. However, note that if $\sigma = 1$ and if the mutants play the strategy "cooperate", they will always earn more than the residents. Hence, there is a maximum value of σ allowing the evolutionary stability of the population under non uniformly-constant assortment.

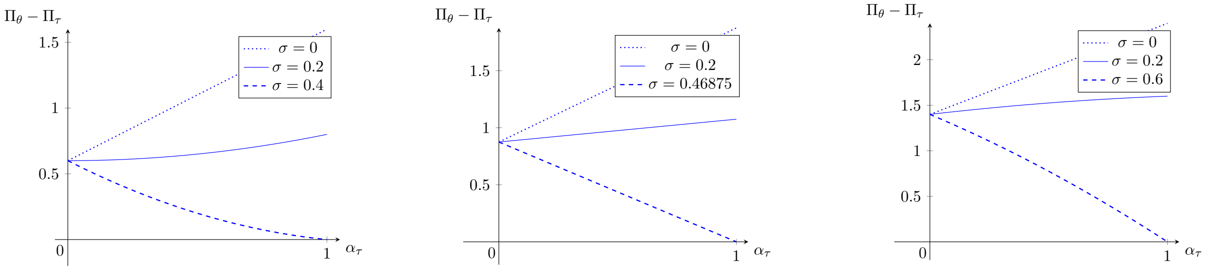
(a) When $S_\pi < 0$ with $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 6$. For $\lambda = 5/6$, $\phi_{12} = 0.28$, *homo kantiensis* and *homo oeconomicus* individuals get the same total payoff $\Pi_\theta = 1.6$. Moreover, for $\sigma < 0.4$, the residents earn a strictly greater payoff than any mutant at the limit, and thus following the same arguments as in Theorem 2, the population is evolutionarily stable (see Figure 6a).

(b) When $S_\pi = 0$ with $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 5$. Recall that the Payoff Equality is satisfied if and only if $\phi_{12} = 0.25$. This value is reached for $\lambda = 17/24$. Then, *homo kantiensis* and *homo oeconomicus* individuals get the same total payoff $\Pi_\theta = 1.875 > \Pi_\tau$ for any mutant when $\sigma < 0.46875$ (see Figure 6b). Thus an *evolutionarily stable population* exists when $\sigma < 0.46875$.

(c) Finally, when $S_\pi > 0$ with $\pi_{CD} = 0$, $\pi_{DD} = 1$, $\pi_{CC} = 4$ and $\pi_{DC} = 4.5$. For $\lambda = 0.5$, $\phi_{12} = 0.2$ and *homo kantiensis* and *homo oeconomicus* individuals get the same total payoff $\Pi_\theta = 2.4$ strictly greater than Π_τ for any mutant when $\sigma < 0.6$ and thus the population is evolutionarily stable (see Figure 6c).

Note that the share of *homo kantiensis* (1-λ) increases with $S_\pi = \pi_{CC} - \pi_{CD} + \pi_{DD} - \pi_{DC}$. When $S_\pi < 0$, there is about 17% of *homo kantiensis* in the population while this share rises up to 50% when $S_\pi > 0$. This comes as no surprise. Indeed, $\pi_{CC} - \pi_{CD}$ can be interpreted as the gain minus the cost of cooperation, and $\pi_{DC} - \pi_{DD}$ the gain minus the cost of defection. In our examples, we changed only the value of π_{DC}. A decrease in π_{DC} increases the relative benefit of cooperation and thus the share of *homo kantiensis*.

Finally, both *homo kantiensis* and *homo oeconomicus* 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 payoff of the population since Π_θ increases with the share of *homo kantiensis*. On the other hand, individuals *homo oeconomicus* help to resist the invasion of mutants. In fact, suppose that



(a) $S_\pi < 0$, $\lambda = 5/6$ **(b)** $S_\pi = 0$, $\lambda = 17/24$ **(c)** $S_\pi > 0$, $\lambda = 0.5$

Figure 6: Payoff difference between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoner's dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ)

mutants are matched with *homo kantiensis* with greater probability than with *homo oeconomicus*. For instance, let $\Gamma = -(1-\sigma)/(1-\lambda)$. Then, in the case of $\sigma = 0$, the mutants would at least earn $\Pi_\tau = 4$ and the heterogeneous resident population 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.

The diversity of heterogeneous evolutionarily stable population

Although we focused on a population of *homo kantiensis* and *homo oeconomicus*, there also exists heterogeneous *evolutionarily stable population* in which the residents do not necessarily play the pure strategies "cooperate" and "defect" under a non uniformly-constant assortment. Thus, they are not necessarily *homo kantiensis* or *homo oeconomicus*. Going back to our three examples:

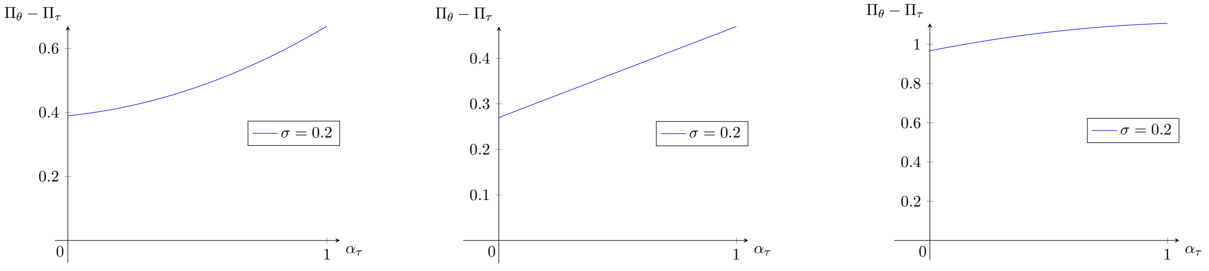
(a) Suppose that individuals θ_1 cooperate with probability $\alpha_1 = 0.78$ and that individuals θ_2 cooperate with probability $\alpha_2 = 0.1$. For $\lambda = 5/6$, $\phi_{12} = 0.28$, and individuals θ_1 and θ_2 get the same average payoff $\Pi_\theta = 1.78984$. When $\sigma = 0.2$, the residents earn more than any mutant at the limit and the population is therefore evolutionarily stable (see Figure 7a).

(b) When individuals θ_1 cooperate with probability $\alpha_1 = 0.8$ and individuals θ_2 cooperate with probability $\alpha_2 = 0.4$, for $\lambda = 17/24$, $\phi_{12} = 0.25$, we have that individuals θ_1 and θ_2 get the same average payoff $\Pi_\theta = 2.55$. Hence, when $\sigma = 0.2$, the residents earn more than any mutant at the limit rendering the population evolutionarily stable (see Figure 7b).

(c) Finally, suppose that individuals θ_1 cooperate with probability $\alpha_1 = 0.85$ and individuals θ_2 cooperate with probability $\alpha_2 = 0.15$. For $\lambda = 0.5$, $\phi_{12} = 0.2$, individuals θ_1 and θ_2 get the same average payoff $\Pi_\theta = 2.38725$. Therefore, for $\sigma = 0.2$, the residents earn more than any mutant at the limit and the resident population is evolutionarily stable (see Figure 7c).

Relaxing the assumption of uniformly-constant assortment, we find a variety of heterogeneous *evolutionarily stable populations*. We notably observe that resident individuals can play strategies outside the set of *Hamiltonian strategies* and that mixed strategies can be observed in a heterogeneous *evolutionarily stable population*.

Note also that these populations are more robust than under uniformly-constant assortment. For instance, when $\sigma = 0.2$, the residents earn a strictly greater payoff than all the mutants, including the behaviorally-alike, in all the cases (see Figures 6 and 7). This also provides an evolutionary justification in favor of the heterogeneity of preferences. Indeed, if there is a single resident type in the population, then this type is *homo moralis* with a morality coefficient $\kappa = 0.2$ (or a behaviorally-alike). However, if the assortativity σ evolves, then all resident individuals become vulnerable to the



(a) $S_\pi < 0$, $\lambda = 5/6$ **(b)** $S_\pi = 0$, $\lambda = 17/24$ **(c)** $S_\pi > 0$, $\lambda = 0.5$

Figure 7: Payoff difference between a resident population playing mixed strategies ((a) $(\alpha_1, \alpha_2) = (0.78, 0.1)$; (b) $(\alpha_1, \alpha_2) = (0.8, 0.4)$; (c) $(\alpha_1, \alpha_2) = (0.85, 0.15)$) and the mutants in prisoner's dilemma, when the mutant share tends to zero, depending on the strategy played by mutants (α_τ)

entry of mutants. For instance, in the case S_π positive and $\sigma = 0.2$, the *Hamiltonian strategies* are defect and cooperate. Now, if the assortativity goes to zero, the only evolutionarily stable strategy is defect, and *homo moralis* with morality $\kappa = 0.2$ is no more evolutionarily stable.[34] Conversely, in a heterogeneous *evolutionarily stable population*, types who would not be evolutionarily stable alone mutually contribute to resist the invasion of mutants, and they are less sensitive to variations in the assortativity. More specifically, the identified population remains evolutionarily stable as long as the assortativity does not exceed a threshold that depends on the population state and the game payoffs (see Figures 6).[35] In particular, they remain evolutionarily stable when the assortativity goes to zero.

Nevertheless, it is worth mentioning that the shapes of ϕ_{12} and Γ were set arbitrarily. These examples aim at introducing the case of non uniformly-constant assortment. A more in-depth analysis of this type of situation is needed to derive more generic results and to better understand the conditions under which heterogeneous *evolutionarily stable populations* exist in this case.

5.3 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 $\pi_{CC} + \pi_{DD} - \pi_{CD} - \pi_{DC}$ and the value of assortativity σ (Proposition 4). As for coordination games, in the example in section 4.3.1, a population composed of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable because *homo oeconomicus* individuals can also play a mixed strategy, incompatible with evolutionary

[34] Because one of his Bayesian Nash equilibria is the cooperation strategy, and the definition of evolutionary stability encompasses all Bayesian Nash equilibria.

[35] If the assortativity is mutant-dependent and can take any value ($\sigma \in [0, 1]$), then there does not exist any *evolutionarily stable population*, as proved by Newton (2017) in the classical setting of one resident.

stability (Proposition 3 and Theorem 2).[36] 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 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 Proposition 4, 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.[37] 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 precision and better understand the individual motives behind the observed decisions.

[36] Individuals *homo kantiensis* (θ_1) always play the socially-optimal strategy A. On the other hand, individuals *homo oeconomicus* (θ_2) can play strategies A, B, or the mixed strategy $(1/6, 5/6)$ (similar arguments as in section 5.1.1).

[37] 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.

6. Conclusion

In many countries and contexts we see some people consistently deciding not to vote, not to donate to charity or not to cooperate in public good games while some other prefer to do so. Following this empirical observation of various behaviors among individuals in a population, 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 assortment functions matrix to model 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* 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, exploring the case of non uniformly-constant assortment, which we briefly discussed in section 5.2, 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.

AppendixA. The Algebra of assortative matching

In a population state $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \lambda_\tau)$, the assortment matrix is $\Phi = ((\phi_{ij}(\lambda_1, ..., \lambda_n, \lambda_\tau)))_{(i,j)\in I^2}$ such as for all $(i,j) \in I^2$, $\phi_{ij}(\lambda_1, ..., \lambda_n, \lambda_\tau) = p_{ii} - p_{ij}$.[38] To be well defined, the matching process must satisfy two sets of conditions:

- The matching conditions: for all $i \in I$, $\sum\limits_{j\in I} p_{ji} = 1$ (5)
- The balancing conditions: for all $(i,j) \in I^2$, $\lambda_j \cdot p_{ij} = \lambda_i \cdot p_{ji}$ (6)

These conditions imply another set of conditions on the assortment functions ϕ_{ij}, which we call *assortment balancing conditions.*

Assortment balancing condition

Property 1 (Assortment balancing condition). The assortment matrix satisfies the *assortment balancing conditions* (11) 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.

$$\begin{aligned}
&\lambda_j \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{ik}\right) - \phi_{ij}\right] - \lambda_i \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{jk}\right) - \phi_{ji}\right] \\
=&\sum_{k\in I} \lambda_j \lambda_k p_{ii} - \sum_{k\in I} \lambda_j \lambda_k p_{ik} - \lambda_j p_{ii} + \lambda_j p_{ij} - \sum_{k\in I} \lambda_i \lambda_k p_{jj} + \sum_{k\in I} \lambda_i \lambda_k p_{jk} + \lambda_i p_{jj} - \lambda_i p_{ji} \\
=&\lambda_j p_{ii} - \sum_{k\in I} \lambda_j \lambda_i p_{ki} - \lambda_j p_{ii} + \lambda_i p_{ji} - \lambda_i p_{jj} + \sum_{k\in I} \lambda_i \lambda_j p_{kj} + \lambda_i p_{jj} - \lambda_i p_{ji} \\
=&\lambda_i \lambda_j \left[\sum_{k\in I} p_{kj} - \sum_{k\in I} p_{ki}\right] \\
=&0
\end{aligned}$$

□

The assortment balancing conditions impose a particular relationship between the assortment functions. For example, as noted by Bergstrom (2003) in the case of assortative encounters between two types, the assortment $\phi_{12} = p_{11} - p_{12}$ defined between a type θ_1 and a type θ_2 is equal to the assortment $\phi_{21} = p_{22} - p_{21}$ defined between θ_2 and θ_1.

[38] $I = (\llbracket 1..n \rrbracket \cup \{\tau\})$

Property 2 (Assortment matrix in a population of two types). When $s = (\theta_1, \theta_2, \lambda_1, \lambda_2)$ with $(\lambda_1, \lambda_2) \in (0,1)^2$, if the matching process satisfies the matching and balancing conditions, then the assortment matrix Φ is symmetric, i.e. we have $\phi_{12}(\lambda_1, \lambda_2) = \phi_{21}(\lambda_1, \lambda_2)$.

Proof. When $s = (\theta_1, \theta_2, \lambda_1, \lambda_2)$, the assortment balancing conditions boils down to $\lambda_2(\lambda_2 - 1)\phi_{12} = \lambda_1(\lambda_1 - 1)\phi_{21}$. Since $\lambda_2 = 1 - \lambda_1$, we get $\phi_{12} = \phi_{21}$. □

When a third type θ_τ is part of the population, Property 2 does not hold anymore, i.e. for all $\varepsilon > 0$ we do not necessarily have $\phi_{12}(\lambda_1, \lambda_2, \varepsilon) = \phi_{21}(\lambda_1, \lambda_2, \varepsilon)$. However, at the limit when the share of the third type goes to zero, the residents are matched between them, as if there was no mutants, and thus we have $\phi_{12} = \phi_{21}$. Formally:

Property 3 (Assortment matrix in a population of two residents and one mutant). When $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \varepsilon)$ with $(\lambda_1, \lambda_2, \varepsilon) \in (0,1)^3$, if the matching process satisfies the matching and balancing conditions, then we have $\lim_{\varepsilon \to 0} \phi_{12}(\lambda_1, \lambda_2, \varepsilon) = \lim_{\varepsilon \to 0} \phi_{21}(\lambda_1, \lambda_2, \varepsilon)$.

Proof. When $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \varepsilon)$, the assortment balancing conditions are:

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

Rewriting the first equation, we get:

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

Note that for all $(i,j) \in \{1, 2, \tau\}^2$, $\phi_{ij} = p_{ii} - p_{ij}$ is bounded and belongs to $[-1, 1]$. Let λ be the limit of λ_2 when ε goes to zero. We have $\lim_{\varepsilon \to 0} \lambda_2(1 - \lambda_2) = \lim_{\varepsilon \to 0} \lambda_1(1 - \lambda_1) = \lambda(1 - \lambda)$, and thus $\lim_{\varepsilon \to 0} \phi_{12}(\lambda_1, \lambda_2, \varepsilon) = \lim_{\varepsilon \to 0} \phi_{21}(\lambda_1, \lambda_2, \varepsilon)$. □

Using the definition of assortativity, the assortment functions $\phi_{ij} : (0,1)^{n+1} \to [-1, 1]$ can be extended by continuity to $(0,1)^n \times [0,1)$ to cover the limit when the mutant share ε goes to zero. Therefore, Property 3 can be rewritten as $\phi_{12}(\lambda_1, \lambda_2, 0) = \phi_{21}(\lambda_1, \lambda_2, 0)$. For coherence, we will impose that $\phi_{12}(\lambda_1, \lambda_2, 0) = \phi_{12}(\lambda_1, \lambda_2)$. Indeed, a population consisting of two types could also be described as a population of three types when the share of the third type is zero.

Matching probabilities

It is possible to write the conditional probabilities p_{ij} in function of the assortment functions ϕ_{ij} and the population shares. Let (S) be the system of equations defined by matching conditions (5), balancing conditions (6) and assortment conditions (10):

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

When the assortment matrix Φ satisfies the assortment balancing conditions, the system (S) has a unique solution:

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

Proof. In AppendixC.1 □

The existence and uniqueness of solution A.2 also means that there are $(n+1)^2$ linearly-independent equations in the system (S) because there are $(n+1)^2$ unknowns p_{ij}. We have $(n+1)$ matching conditions (5), $\binom{n+1}{2}$ balancing conditions (6) and $2 \times \binom{n+1}{2}$ assortment conditions (10) (the diagonal of the assortment matrix adds no conditions since the $\phi_{ii} = p_{ii} - p_{ii}$). Thus, there are $(n+1)^2 + \binom{n+1}{2}$ equations in (S). However, the $\binom{n+1}{2}$ assortment balancing conditions allow to express $\binom{n+1}{2}$ assortment functions as a function of the others which gives $(n+1)^2$ linearly-independent equations in (S).

AppendixB. The case of two resident types (n=2)

We apply the model to the case of three types: two resident types (θ_1 and θ_2) and one mutant θ_τ. The three types and their respective shares define a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \varepsilon)$, where $(\lambda_1, \lambda_2, \varepsilon) \in (0,1)^3$ are the population shares of θ_1, θ_2 and θ_τ.

We can explicit the conditional probabilities computed in (A.2):

$$\begin{aligned}
p_{11} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \varepsilon \cdot \phi_{1\tau} \\
p_{12} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \varepsilon \cdot \phi_{1\tau} - \phi_{12} \\
p_{1\tau} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \varepsilon \cdot \phi_{1\tau} - \phi_{1\tau} \\
p_{21} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \varepsilon \cdot \phi_{2\tau} - \phi_{21} \\
p_{22} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \varepsilon \cdot \phi_{2\tau} \\
p_{2\tau} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \varepsilon \cdot \phi_{2\tau} - \phi_{2\tau} \\
p_{\tau 1} &= \varepsilon + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2} - \phi_{\tau 1} \\
p_{\tau 2} &= \varepsilon + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2} - \phi_{\tau 2} \\
p_{\tau\tau} &= \varepsilon + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2}
\end{aligned} \tag{B.1}$$

We can then calculate the limits of the conditional probabilities when the mutant share ε tends to zero. First note that for all $(i,j) \in \{1,2,\tau\}^2$, ϕ_{ij} is bounded, and thus $\lim_{\varepsilon\to 0} \varepsilon\phi_{ij} = 0$. Also, the definition of assortativity (8) implies that: for all $i \in I$, $\lim_{\varepsilon\to 0} \phi_{\tau i} = \sigma$.

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

From Property 3, we have: $\phi_{12}(1-\lambda, \lambda, 0) = \phi_{21}(1-\lambda, \lambda, 0) = \phi_{12}(\lambda)$.

Finally, we need to compute the limits of $\phi_{1\tau}$ and $\phi_{2\tau}$. Using the assortment balancing conditions (detailed in the proof of Property 3), we have:

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

The limits of $\phi_{1\tau}$ and $\phi_{2\tau}$ when ε goes to zero are:

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

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

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

$$\begin{aligned}
\lim_{\varepsilon\to 0}\, p_{11} &= (1-\lambda) + \lambda \cdot \phi_{12}(\lambda)\\
\lim_{\varepsilon\to 0}\, p_{12} &= (1-\lambda) \cdot (1-\phi_{12}(\lambda))\\
\lim_{\varepsilon\to 0}\, p_{1\tau} &= (1-\lambda) \cdot (1-\sigma) - \lambda \cdot (1-\lambda) \cdot \Gamma\\
\lim_{\varepsilon\to 0}\, p_{21} &= \lambda \cdot (1-\phi_{12}(\lambda))\\
\lim_{\varepsilon\to 0}\, p_{22} &= \lambda + (1-\lambda) \cdot \phi_{12}(\lambda)\\
\lim_{\varepsilon\to 0}\, p_{2\tau} &= \lambda \cdot (1-\sigma) + \lambda \cdot (1-\lambda) \cdot \Gamma\\
\lim_{\varepsilon\to 0}\, p_{\tau 1} &= 0\\
\lim_{\varepsilon\to 0}\, p_{\tau 2} &= 0\\
\lim_{\varepsilon\to 0}\, p_{\tau\tau} &= \sigma
\end{aligned} \tag{B.2}$$

Note that when ε goes to zero, we have $p_{\tau 1} = p_{\tau 2} = 0$, and individuals of type θ_1 and θ_2 behave as if individuals θ_τ were not in the population. Bayesian Nash equilibria of θ_1 and θ_2 are then the same as in Definition 2. Moreover, the conditional probabilities p_{11}, p_{21}, p_{12} and p_{22} are consistent with the classical setting (Bergstrom, 2003; Alger and Weibull, 2013).

When the assortment matrix is uniformly constant, we have $\phi_{12}(\lambda) = \sigma$ and $\Gamma = 0$. The limit Γ can be interpreted as the matching probability difference between mutants and residents of the two types: $\Gamma = \lim_{\varepsilon\to 0}(p_{\tau 2} - p_{\tau 1})/\varepsilon$. 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 differentiable in $\varepsilon = 0$, we have $\Gamma = \phi'_{\tau 1}(\lambda, 0) - \phi'_{\tau 2}(\lambda, 0)$[39] because $\phi_{\tau 1}(1-\lambda, \lambda, 0) = \phi_{\tau 2}(1-\lambda, \lambda, 0) = \sigma$. Therefore, Γ is the marginal assortment difference between mutants and residents of the two types.

[39] Noting $\phi'_{\tau i}(\lambda, 0) = \partial\phi_{\tau i}(\lambda, 0)/\partial\varepsilon$

AppendixC. Proofs of Lemmas and Propositions

AppendixC.1 Proof of Proposition 1

Let (S) be the system of equations defined by matching conditions (5), balancing conditions (6) and assortment conditions (10):

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

Suppose there exists matching probabilities p_{ij} solutions of the system (S). Since $\sum\limits_{k\in I} p_{ki}=1$, we have $\sum\limits_{k\in I}\lambda_i p_{ki}=\lambda_i$ for all $i\in I$. Using the balancing conditions, we get $\lambda_i-\sum\limits_{k\in I}\lambda_k p_{ik}=0$. Moreover, since $\sum\limits_{k\in I}\lambda_k=1$, we have $p_{ii}=\sum\limits_{k\in I}\lambda_k p_{ii}$. Adding these two equations, we obtain $p_{ii}=\lambda_i+\sum\limits_{k\in I}\lambda_k(p_{ii}-p_{ik})$ for all $i\in I$, i.e. $p_{ii}=\lambda_i+\sum\limits_{k\in I}\lambda_k\phi_{ik}$. Since for all $(i,j)\in I^2$, $p_{ij}=p_{ii}-\phi_{ij}$, we get $p_{ij}=\lambda_i+\sum\limits_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}$. We have proven that if a solution of (S) exists, then it must be $p_{ij}=\lambda_i+\sum\limits_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}$ (12).

We now show that $q_{ij}=\lambda_i+\sum\limits_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}$ (12) is solution of (S) using the assortment balancing conditions:

$$\begin{cases}\forall\, j\in I,\ \sum\limits_{i\in I} q_{ij}=\sum\limits_{i\in I}\left[\lambda_i+\sum\limits_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}\right]=1+\sum\limits_{i\in I}\frac{\lambda_i}{\lambda_j}\left[\sum\limits_{k\in I}\lambda_k\phi_{jk}-\phi_{ji}\right]=1+\frac{1}{\lambda_j}\left[\sum\limits_{k\in I}\lambda_k\phi_{jk}-\sum\limits_{i\in I}\lambda_i\phi_{ji}\right]=1\\ \forall\,(i,j)\in I^2,\ \lambda_j q_{ij}-\lambda_i q_{ji}=\lambda_j\lambda_i+\lambda_j\left[\sum\limits_{k\in I}\lambda_k\phi_{ik}-\phi_{ij}\right]-\lambda_i\lambda_j-\lambda_i\left[\sum\limits_{k\in I}\lambda_k\phi_{jk}-\phi_{ji}\right]=0\\ \forall\,(i,j)\in I^2,\ q_{ii}-q_{ij}=\lambda_i+\sum\limits_{k\in I}\lambda_k\phi_{ik}-\lambda_i-\sum\limits_{k\in I}\lambda_k\phi_{ik}+\phi_{ij}=\phi_{ij}\end{cases}$$

AppendixC.2 Proof of Lemma 1

This proof extends the proof provided by Alger and Weibull (2013) for one resident type to the case of $n>1$ resident types.

First, from Definition 6 we have that, in a population state $s=(\theta_1,...,\theta_n,\theta_\tau,\lambda_1,...,\lambda_n,\varepsilon)$, $(x^1,x^2,...,x^n,x^\tau)\in X^{n+1}$ is a type-homogeneous Bayesian Nash equilibrium if::

$$\forall i\in I:\quad x^i\in\underset{x\in X}{\operatorname{argmax}}\ \sum_{j\in I}\left(p_{ij}\cdot u_i(x,x^j)\right)\tag{C.1}$$

Hence, rewriting with the assortment matrix Φ, we get:

$$\forall i\in I:\quad x^i\in\underset{x\in X}{\operatorname{argmax}}\ \sum_{j\in I}\left(\left[\lambda_i+\left(\left(\sum_{k\in I}\lambda_k\phi_{ik}\right)-\phi_{ij}\right)\right]\cdot u_i(x,x^j)\right)\tag{C.2}$$

Since for all $i\in I, u_i$ is continuous and X is compact, then Weierstrass's maximum theorem implies that the right hand side in C.2 defines a non-empty and compact set for all $(\lambda_1,...,\lambda_n,\varepsilon)\in$

$(0,1)^n \times [0,1)$. The solutions to C.2 can thus be written as the set of fixed points of a compact valued and upper hemi-continuous correspondence (Aliprantis and Border, 2006) $B_\lambda : X^{n+1} \rightrightarrows X^{n+1}$. This set is therefore closed, and $B^{NE}(s)$ is compact for each $s = (\theta_1, ..., \theta_n, \theta_\tau, \lambda_1, ..., \lambda_n, \varepsilon) \in \Theta^{n+1} \times (0,1)^n \times [0,1)$.

Second, since for all $i \in I, u_i$ is concave in the first argument then B_λ is convex-valued and has a fixed point by Kakutani's fixed point theorem which establishes the fact that $B^{NE}(s)$ is non-empty.

Third, fixing the $(\theta_i)_{i\in I}$, we can write the maximands in C.2 as $U_i(x, x^1, x^2, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon)$. We note $U_i{}^*(x^1, x^2, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon) = \max_{x\in X} U_i(x, x^1, x^2, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon)$. Since the $(U_i)_{i\in I}$ are continuous, Berge's maximum theorem implies that their maximum over X, $U_i{}^*$, are also continuous. Moreover, by definition of $B^{NE}(s)$, we have, $(x^1, x^2, ..., x^n, x^\tau) \in B^{NE}(s)$ if and only if for all $i \in I$:

$$U_i{}^*(x^1, x^2, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon) - U_i(x, x^1, x^2, ..., x^n, x^\tau, \lambda_1, ..., \lambda_n, \varepsilon) \geq 0 \quad \forall x \in X \tag{C.3}$$

For all $i \in [\![1..n]\!]$ let $< \lambda_{i,t} >_{t\in\mathbb{N}} \rightarrow \lambda_i^0$ and let $< \varepsilon_t >_{t\in\mathbb{N}} \rightarrow \varepsilon^0$. If $(x_t^1, x_t^2, ..., x_t^n, x_t^\tau) \in B^{NE}(s)$ and for all $i \in I$, $x_t^i \rightarrow x^{i,0}$, we have by continuity on the left-hand side in C.3, for all $i \in I$:

$$U_i{}^*(x^{1,0}, x^{2,0}, ..., x^{n,0}, x^{\tau,0}, \lambda_1^0, ..., \lambda_n^0, \varepsilon^0) - U_i(x, x^{1,0}, x^{2,0}, ..., x^{n,0}, x^{\tau,0}, \lambda_1^0, ..., \lambda_n^0, \varepsilon^0) \geq 0 \quad \forall x \in X \tag{C.4}$$

This last results proves that $(x^{1,0}, x^{2,0}, ..., x^{n,0}, x^{\tau,0}) \in B^{NE}(s)$ and therefore that the correspondence $B^{NE}(\theta_1, ..., \theta_n, \theta_\tau, \cdot) : (0,1)^n \times [0,1) \rightrightarrows X^{n+1}$ is upper hemi-continuous.

AppendixC.3 Proof of Proposition 2

First, let's assume that Payoff Equality is satisfied. We distinguish two cases:

When $\phi_{12} = 1$:
Since we have Payoff Equality, equations (17) and (18) are satisfied and we necessarily have that $Q_\pi = R_\pi = 0$ (since $(1 - \phi_{12}) = 0$). Then either $S_\pi = 0$ and we are in case 1. of the proposition, or $S_\pi \neq 0$ and we are in case 3. of the proposition.

When $\phi_{12} \neq 1$:
Payoff equality (equations (17) and (18)) implies only two possibilities.
Either $Q_\pi = 0$ and then $S_\pi = 0$ and we are in the case 1. of the proposition.
Or $Q_\pi \neq 0$ and then since $(1 - \phi_{12}) > 0$ and $\lambda > 0$, Q_π and S_π are of the same sign. Moreover, since $\lambda < 1$, R_π and S_π have the same sign. Finally, dividing (17) by $S_\pi \neq 0$ we have: $\lambda(1 - \phi_{12}) = Q_\pi / S_\pi$ and we are in case 2. of the proposition.

For the converse, using similar arguments, if one of the three cases (1.,2.,3. of Proposition 1) is true then the condition stated in equation (17) is satisfied and we have Payoff Equality.

AppendixC.4 Proof of Lemma 2

Let $(x^1, x^2, x^\tau) \in X^3$ be a Bayesian Nash equilibrium of the population $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \varepsilon)$ and suppose that the population $s = (\theta_1, \theta_2, \lambda)$ satisfies the Payoff Equality condition: $\Pi_1(x^1, x^2, \lambda) = \Pi_2(x^1, x^2, \lambda) = \Pi_\theta$.

For all $(i, j) \in I^2$, we note $\pi(x^i, x^j) = \pi^{ij}$. We can write the payoffs obtained by each type when

ε goes to zero using the conditional probabilities B.2:

$$\begin{cases} \Pi_1(x^1,x^2,x^\tau,\lambda,0) = \Pi_1 = (1-\lambda+\lambda\phi_{12})\cdot\pi^{11} + \lambda(1-\phi_{12})\cdot\pi^{12} \\ \Pi_2(x^1,x^2,x^\tau,\lambda,0) = \Pi_2 = (1-\lambda)(1-\phi_{12})\cdot\pi^{21} + (\lambda+(1-\lambda)\phi_{12})\cdot\pi^{22} \\ \Pi_\tau(x^1,x^2,x^\tau,\lambda,0) = \Pi_\tau = [(1-\lambda)(1-\sigma) - \lambda(1-\lambda)\Gamma]\cdot\pi^{\tau 1} + [\lambda(1-\sigma)+\lambda(1-\lambda)\Gamma]\cdot\pi^{\tau 2} + \sigma\cdot\pi^{\tau\tau} \end{cases}$$

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 when matched with an individual playing y is then: $\pi(x,y) = xAy$ We can rewrite the payoffs in function of the matrix payoff A:

$$\begin{cases} \Pi_1 = x^1\left[(1-\lambda)(1-\phi_{12})Ax^1 + \lambda(1-\phi_{12})Ax^2\right] + \phi_{12}x^1Ax^1 \\ \Pi_2 = x^2\left[(1-\lambda)(1-\phi_{12})Ax^1 + \lambda(1-\phi_{12})Ax^2\right] + \phi_{12}x^2Ax^2 \\ \Pi_\tau = x^\tau\left[((1-\lambda)(1-\sigma) - \lambda(1-\lambda)\Gamma)Ax^1 + (\lambda(1-\sigma)+\lambda(1-\lambda)\Gamma)Ax^2\right] + \sigma x^\tau Ax^\tau \end{cases}$$

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 $x^\tau = (\alpha_\tau, 1-\alpha_\tau)$. When $x^1 \neq x^2$, there exists $\gamma \in \mathbb{R}$ such that $x^\tau = \gamma x^1 + (1-\gamma)x^2 = (\gamma\alpha_1 + (1-\gamma)\alpha_2, 1-\gamma\alpha_1-(1-\gamma)\alpha_2)$.

From Payoff Equality, we know that $\Pi_1 = \Pi_2 = \Pi_\theta$. Thus, $\gamma\Pi_1 + (1-\gamma)\Pi_2 = \Pi_\theta$. We can then write the difference between the payoff of the residents and the payoff of the mutants as follows:

$$\begin{aligned} \Pi_\theta - \Pi_\tau &= \gamma\Pi_1 + (1-\gamma)\Pi_2 - \Pi_\tau \\ &= [\gamma\phi_{12} - \gamma^2\sigma - \gamma(1-\lambda)(\phi_{12}-\sigma) + \gamma\lambda(1-\lambda)\Gamma]\cdot x^1Ax^1 \\ &+ [-\gamma(1-\gamma)\sigma - \gamma\lambda(\phi_{12}-\sigma) - \gamma\lambda(1-\lambda)\Gamma]\cdot x^1Ax^2 \\ &+ [-\gamma(1-\gamma)\sigma - (1-\gamma)(1-\lambda)(\phi_{12}-\sigma) + (1-\gamma)\lambda(1-\lambda)\Gamma]\cdot x^2Ax^1 \\ &+ [(1-\gamma)\phi_{12} - (1-\gamma)^2\sigma - (1-\gamma)\lambda(\phi_{12}-\sigma) - (1-\gamma)\lambda(1-\lambda)\Gamma]\cdot x^2Ax^2 \end{aligned}$$

Rearranging, we get:

$$\begin{aligned} \Pi_\theta - \Pi_\tau = &[\gamma(1-\gamma)\sigma + \gamma\lambda(\phi_{12}-\sigma) + \gamma\lambda(1-\lambda)\Gamma]\cdot[x^1Ax^1 - x^1Ax^2 - x^2Ax^1 + x^2Ax^2] \\ &+ [(1-\gamma-\lambda)(\phi_{12}-\sigma) - \lambda(1-\lambda)\Gamma]\cdot[x^2A(x^2-x^1)] \end{aligned}$$

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

$$\begin{cases} x^1Ax^1 = \alpha_1^2\pi_{11} + \alpha_1(1-\alpha_1)(\pi_{21}+\pi_{12}) + (1-\alpha_1)^2\pi_{22} \\ x^1Ax^2 = \alpha_1\alpha_2\pi_{11} + \alpha_1(1-\alpha_2)\pi_{12} + (1-\alpha_1)\alpha_2\pi_{21} + (1-\alpha_1)(1-\alpha_2)\pi_{22} \\ x^2Ax^1 = \alpha_1\alpha_2\pi_{11} + \alpha_2(1-\alpha_1)\pi_{12} + (1-\alpha_2)\alpha_1\pi_{21} + (1-\alpha_1)(1-\alpha_2)\pi_{22} \\ x^2Ax^2 = \alpha_2^2\pi_{11} + \alpha_2(1-\alpha_2)(\pi_{21}+\pi_{12}) + (1-\alpha_2)^2\pi_{22} \end{cases} \tag{C.5}$$

Therefore:

$$\begin{aligned} &x^1Ax^1 - x^1Ax^2 - x^2Ax^1 + x^2Ax^2 = (\alpha_1-\alpha_2)^2(\pi_{11}+\pi_{22}-\pi_{12}-\pi_{21}) \\ &x^2A(x^2-x^1) = (\alpha_1-\alpha_2)[\alpha_2(\pi_{12}-\pi_{11}) + (1-\alpha_2)(\pi_{22}-\pi_{21})] \end{aligned} \tag{C.6}$$

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

$$\begin{aligned}\Pi_\theta - \Pi_\tau = &[\gamma(1-\gamma)\sigma + \gamma\lambda(\phi_{12}-\sigma) + \gamma\lambda(1-\lambda)\Gamma] \cdot (\alpha_1 - \alpha_2)^2 \cdot (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21}) \\ &+ [(1-\gamma-\lambda)(\phi_{12}-\sigma) - \lambda(1-\lambda)\Gamma] \cdot (\alpha_1-\alpha_2) \cdot [\alpha_2(\pi_{12}-\pi_{11}) + (1-\alpha_2)(\pi_{22}-\pi_{21})]\end{aligned} \tag{C.7}$$

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

$$\Pi_\theta - \Pi_\tau = \gamma(1-\gamma)\sigma(\alpha_1-\alpha_2)^2 (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21}) \tag{C.8}$$

AppendixC.5 Proof of Proposition 3

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 Payoff Equality condition, it is not evolutionarily stable. Thus, we consider next a population that respects the Payoff Equality condition (Proposition 2).

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 from Lemma 2 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 (C.6), we know that $S_\pi = (\alpha_1-\alpha_2)^2 (\pi_{11} + \pi_{22} - \pi_{12} - \pi_{21})$.[40] Hence, rewriting the expression above, we have:

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

We consider the three different cases of Lemma 3:

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.3). 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. or 3. 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. $x = (\eta, 1-\eta)$

[40] 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.

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 (C.5), 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 3 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.3) 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 $\langle(x_t^1,x_t^2,\hat{x}_t)\rangle_{t\in\mathbb{N}}$ 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 $\Pi_2(x_t^1,x_t^2,\hat{x},\lambda_t,\varepsilon_t) < \Pi_\tau(x_t^1,x_t^2,\hat{x},\lambda_t,\varepsilon_t)$.

AppendixC.6 Proof of Proposition 4

First note that *homo oeconomicus* individuals always defect while *homo kantiensis* individuals always cooperate. If there exists a heterogeneous *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis*, we know from Proposition 3 that $\{0,1\} \in X_\sigma$. From Lemma 3, we either have $X_\sigma = \{0,1\}$ and $\pi_{CC}+\pi_{DD}-\pi_{DC}-\pi_{CD}>0$ or $X_\sigma=[0,1]$ and $\pi_{CC}+\pi_{DD}-\pi_{DC}-\pi_{CD}=0$. But in the latter case, the game is additive and the population cannot be evolutionarily stable. Thus, $X_\sigma=\{0,1\}$ and $\pi_{CC}+\pi_{DD}-\pi_{DC}-\pi_{CD}>0$. This also means that $S_\pi>0$. Since $\sigma<1$ (else the Payoff Equality can not be respected since *homo kantiensis* would earn more than *homo oeconomicus* by only meeting similar others), from Proposition 2 we also have $Q_\pi>0$ and $R_\pi>0$. Let *homo kantiensis* be type θ_1, then $Q_\pi = \pi_{CC}-\pi_{DC}-\sigma(\pi_{DD}-\pi_{DC})$ and $R_\pi=\pi_{DD}-\pi_{CD}-\sigma(\pi_{CC}-\pi_{CD})$. Since we have $\pi_{CD}<\pi_{DD}<\pi_{CC}<\pi_{DC}$, we obtain $(\pi_{DC}-\pi_{CC})/(\pi_{DC}-\pi_{DD})<\sigma<(\pi_{DD}-\pi_{CD})/(\pi_{CC}-\pi_{CD})$.

If $\pi_{CC} + \pi_{DD} - \pi_{DC} - \pi_{CD} > 0$, then from Lemma 3 we have $X_\sigma \subseteq \{0, 1\}$. Moreover, individuals *homo oeconomicus* always defect, while *homo kantiensis* always cooperate. Thus, $S_\pi = \pi_{CC} + \pi_{DD} - \pi_{DC} - \pi_{CD} > 0$. Since $(\pi_{DC} - \pi_{CC})/(\pi_{DC} - \pi_{DD}) < \sigma < (\pi_{DD} - \pi_{CD})/(\pi_{CC} - \pi_{CD})$, we have $Q_\pi = \pi_{CC} - \pi_{DC} - \sigma(\pi_{DD} - \pi_{DC}) > 0$ and $R_\pi = \pi_{DD} - \pi_{CD} - \sigma(\pi_{CC} - \pi_{CD}) > 0$. From Proposition 2, we know that $\lambda = Q_\pi/((1 - \sigma)S_\pi) \in (0, 1)$ satisfies Payoff Equality. Moreover, using the same arguments than in the proof of Proposition 3 (case 1 when $x^1 \neq x^2$), we can show that $X_\sigma = \{0, 1\}$. Consequently, from Theorem 2, we can conclude that the population of *homo oeconomicus* and *homo kantiensis* with $\lambda = Q_\pi/((1 - \sigma)S_\pi)$ is evolutionarily stable.

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.

Charalambos Aliprantis and Kim C Border. *Infinite dimensional analysis.* Springer, 2006.

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.

Samuel Bowles and Herbert Gintis. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical population biology*, 65(1):17–28, 2004.

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.

Paul W Eastwick and Eli J Finkel. Sex differences in mate preferences revisited: Do people know what they initially desire in a romantic partner? *Journal of Personality and Social Psychology*, 94 (2):245, 2008.

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

Ilan Eshel and Luigi Luca Cavalli-Sforza. Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences*, 79(4):1331–1335, 1982.

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.

Chaim Fershtman and Yoram Weiss. Social rewards, externalities and stable preferences. *Journal of Public Economics*, 70(1):53–73, 1998.

Benjamin Golub and Matthew O Jackson. How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338, 2012.

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.

Nicholas Harrigan and Janice Yap. Avoidance in negative ties: Inhibiting closure, reciprocity, and homophily. *Social Networks*, 48:126–141, 2017.

Christoph Hauert, Silvia De Monte, Josef Hofbauer, and Karl Sigmund. Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570):1129–1132, 2002.

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.

Theo Hitiris and John Posnett. The determinants and effects of health expenditure in developed countries. *Journal of Health Economics*, 11(2):173–181, 1992.

Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, 2003.

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.

Levent Koçkesen, Efe A Ok, and Rajiv Sethi. The strategic advantage of negatively interdependent preferences. *Journal of Economic Theory*, 92(2):274–299, 2000.

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.

Jonathan Newton. The preferences of Homo Moralis are unstable under evolving assortativity. *International Journal of Game Theory*, 46(2):583–589, 2017.

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 and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, 437(7063):1291, 2005.

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.

William H Sandholm. Preference evolution, two-speed dynamics, and rapid social change. *Review of Economic Dynamics*, 4(3):637–679, 2001.

William H Sandholm. *Population games and evolutionary dynamics.* MIT press, 2010.

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.

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.