DEMB Working Paper Series

N. 84

The Interplay of Cultural Aversion and Assortativity for the Emergence of Cooperation

Ennio Bilancini*, Leonardo Boncinelli**, Jiabin Wuz***

April 2016

* University of Modena and Reggio Emilia
Address: Viale Berengario 51, 41121 Modena, Italy
email: ennio.bilancini @unimore.it

**University of Florence
Address: Via delle Pandette 9, 50127, Florence, Italy
email: leonardo.boncinelli @unifi.it

***University of Oregon,
Address: 1285 University of Oregon, Eugene, OR, USA
email: jwu5@uoregon.edu

ISSN: 2281-440X online

The Interplay of Cultural Aversion and Assortativity for the Emergence of Cooperation

Ennio Bilancini* Leonardo Boncinelli† Jiabin Wu‡

April 30, 2016

Abstract

This paper investigates the emergence of cooperation in a heterogeneous population. The population is divided into two cultural groups. Agents in the population are randomly matched in pairs to engage in a prisoner dilemma. The matching process is *assortative*, that is, cooperators are more likely to be matched with cooperators, defectors are more likely to be matched with defectors. When two agents of different cultures are matched, they suffer a cost due to their cultural differences. We call such a cost *cultural aversion*. We find that when cultural aversion is sufficiently strong, perfect correlation between culture and behavior emerges: all agents from one cultural group cooperate, while all agents from the other cultural group defect.

JEL classification code: C72; C73; Z10.

Keywords: prisoner dilemma; assortativity; cultural aversion; cooperation; type-monomorphic

*Dipartimento di Economia "Marco Biagi", Università degli Studi di Modena e Reggio Emilia, Viale Berengario 51, 43 ovest, 41121 Modena, Italia. Tel.: +39 059 205 6843, fax: +39 059 205 6947, email: ennio.bilancini@unimore.it.

†Dipartimento di Scienze per l'Economia e l'Impresa, Università degli Studi di Firenze, Via delle Pandette 9, 50127 Firenze, Italia. Tel.: +39 055 2759578, fax: +39 055 2759910, email: leonardo.boncinelli@unifi.it.

‡Department of Economics, University of Oregon, 1285 University of Oregon, Eugene, OR, USA. Tel: +1 (541) 346-5778. Email: jwu5@uoregon.edu

1 Introduction

It is matter of fact that societies in almost every age and place, and especially modern societies, are comprised of groups which differ between each other on a cultural basis. Often these differences – think, for instance, of language and religion – entail a cost that individuals suffer when interactions occur between members of different groups.

The literature studying under which conditions cooperation can emerge in societies has so far given little consideration to the role of costly interactions between different cultural groups.[1] This paper attempts to address this issue.

We consider a model in which a heterogeneous population of agents are randomly matched in pairs to engage in a prisoner dilemma. Each agent carries one of two different cultural types and the population is divided into two cultural groups. Each cultural type is substantiated by a set of identity traits, like language and religion. These traits are essential because they represent the identity of a culture (Akerlof and Kranton, 2000, 2005). At the same time, these traits may originate a cost when members of different cultural groups interact with each other: think of communication costs due to different languages, or coordination costs due to differences in work times for religious habits (e.g., Muslims observe Ramadan, Jews rest on Saturday, Catholics rest on Sunday). Such a cost can also be psychological. Starting with the seminal work of Becker (1957) on taste-based discrimination, numerous researches have found that mistrust, animosity and negative attitude across cultural groups have significant impact on trade and investment (Guiso et al., 2009, Michaels and Zhi, 2010, Fisman et al., 2014); on labor market outcomes (Becker, 1993, Bertrand and Mullainathan, 2004, Bandiera et al., 2009); on financial activities such as mergers and bank loans (Giannetti and Yafeh, 2012, Ahern et al., 2012). We call the cost/disutility of cross-cultural interaction *cultural aversion* in the rest of the paper.

Each agent also carries an auxiliary trait, which is characterized by his action to cooperate or defect in the prisoner dilemma. The auxiliary traits have cultural nature as well, but they do not convey an identity.

Changes in traits such as preferences, customs and faiths usually take place across generations over time, while actions evolve over a much shorter time horizon. In this paper, we take identity traits as given for agents, and we analyze the evolution over auxiliary traits. This means that we are considering a time horizon that is not long enough to allow identity traits to vary, but sufficiently long for selection to be active on auxiliary traits.

[1] The importance of cultural factors for the evolution of cooperation has been studied, among others, by Henrich and Boyd (2001), Boyd et al. (2003), Henrich (2004), Boyd and Richerson (2009), and Boyd et al. (2011). None of these or similar studies, however, consider the cost of interactions due to cultural mismatch.

The central question we investigate is whether cultural aversion between groups is relevant for cooperation. At a first glance, one may think that cultural aversion does not affect, *per se*, the relative advantage of cooperation over defection, and hence it cannot have any effect on the evolution of cooperation. If this is the case, cultural aversion simply reduces individual payoffs depending on the frequency of type-mismatches (i.e., the frequency of matches between agents belonging to different cultural groups); therefore, interventions aimed at reducing cultural aversion are clearly the best policy for societal interests. However, it turns out that is not necessarily the case if we allow for action-assortativity in the matching process, i.e., cooperators are more likely to interact with cooperators, and defectors with defectors. Action-assortativity describes people's generally tendency to interact more with those who act like them than with those who behave differently. In social psychology, social matching theory (Walster et al., 1966) argues that people tend to form successful partnerships with those who share similar levels of social desirability. In sociology, there is a long tradition to study behavioral homophily (see McPherson et al., 2001, for a survey), which finds that people associate themselves with those who share similar behavioral patterns. Although assortativity is commonly observed in human societies along many other dimensions including race, gender, language, religion, dress and origin (see, among others, McPherson et al., 2001, Ruef et al., 2003, Currarini et al., 2009, 2010, Bramoullé et al., 2012), action-assortativity plays a unique role in the situation we consider: given action-assortativity, cooperation and defection can work as instruments to avoid type-mismatches, and cultural aversion can be, to some extent, beneficial to society.

More specifically, we find that larger cultural aversion works in favor of states where there is perfect correlation between culture and behavior, that we call *type-monomorphic* states: all agents of one cultural group cooperate, and all agents of the other cultural group defect. This result crucially depends on the interplay between action-assortativity and cultural aversion; given the presence of action-assortativity, cultural aversion provides each agent an incentive to conform to the action that is mostly played by his own group members to avoid costly type-mismatches. The magnitude of such an incentive depends on both the degree of action-assortativity and the degree of segregation of the two cultural groups in actions. When the two cultural groups are sufficiently segregated in actions, that is, the majority of one group choose to cooperate, while the majority of the other group choose to defect, action-assortativity increases the cost of an agent choosing an action differing from what most his own group members choose; this effect lessens with a smaller degree of assortativity and a less pronounced segregation, but it vanishes only if the degree of assortativity goes to zero or if segregation is nil.

Our results have important implications in terms of welfare. When action-assortativity is sufficiently strong such that full cooperation can be achieved in the absence of cultural aversion, the presence of cultural aversion can reduce cooperation. Obviously, full cooperation without any cultural aversion is the first best from a societal point of view. However, and notably, if cultural aversion cannot be reduced to zero, then the type-monomorphic state where only the larger cultural group cooperates can entail a larger total surplus than the monomorphic state where everybody cooperates. In fact, the reduction of benefits from cooperation (due to the fact that part of the population defects to avoid type-mismatches) can be more than compensated by the reduction in the frequency of costly type-mismatches (thanks to the perfect correlation between action and culture). So, somewhat surprisingly it may happen that a small *increase* in cultural aversion can increase total welfare, when the society is moved from a monomorphic to a type-monomorphic state. Of course, as cultural aversion increases even more, total welfare is bound to decrease indefinitely, unless action-assoratativity is total. Furthermore, when action-assortativity is sufficiently weak such that full defection is obtained in the absence of cultural aversion, a positive degree of cultural aversion can increase cooperation by allowing the society to move to a type-monomorphic state. Importantly, societal welfare can increase in this way only if cultural aversion is not too large.

Beyond the baseline model described above, we develop four extensions with the aim of investigating if, and to what extent, our findings about the interplay between cultural aversion and action-assortativity are affected by other elements that can reasonably play a role.

First, we consider the possibility of assortativity also in identity traits. In this case the matching process also exhibits type-assortativity (Alger and Weibull, 2013), that is, agents from the same cultural group has a higher probability to be matched with their own group members. We find that allowing for some degree of type-assortativity does not change our main results qualitatively. Moreover, type-assortativity alone cannot account for the phenomenon of perfect correlation between culture and behaviors, even when cultural aversion is present, because cooperation and defection can no longer serve as instruments to help the agents to avoid type-mismatches. This demonstrates the importance of action-assortativity in our model.

Second, we consider the existence of asymmetries in the degree of cultural aversion (as suggested by Bisin et al., 2004), with agents of one type suffering more when type-mismatches occur. Here as well our results are qualitatively maintained, with the additional insight that an increase in the degree of assortativity in actions can make the system switch from a state

where the majority cooperates and the minority defects to a state where the majority defects and the minority cooperates. In terms of welfare, this generates a non-monotonic pattern.

Third, we consider the interplay of assortativity, cultural aversion, and a *legal institution*, which is modeled as a compensation scheme (Tabellini et al., 2008): when a defector meets a cooperator, and the legal system detects the incident, the former has to pay a compensation to the latter. We find that the legal institution itself cannot produce type-monomorphic states even if cultural aversion is in place. Moreover, when the legal institution is weak (i.e., low compensation) and/or cultural aversion is weak, the legal institution complements assortative matching to achieve full cooperation. On the other hand, when the legal institution is strict (i.e., high compensation), and cultural aversion is strong, a somewhat counter-intuitive result occurs: for a low degree of action-assortativity full cooperation is obtained, while a high degree of action-assortativity leads to type-monomorphic states. The rationale is that when action-assortativity is sufficiently high, the legal institution would not represent a very effective punishment for defectors, since they rarely end up matched with cooperators; instead, differentiating in actions is still an effective way to avoid costly type-mismatches. This result may help to explain why in many developed countries with strong formal institutions, minorities still suffer from high crime rates. For example, in the United States, certain ethnic minority groups concentrate in inner cities, where crime rates are significantly higher compared to the average crime rate nationwide. Moreover, members of these groups are the main victims of the crimes occurred in these hypersegregated areas (Massey, 1995).

Finally, we consider the case where action-assortativity is state-dependent, that is, the likelihood of assortative matching in actions depends on the fraction of the population that cooperates. We show that the introduction of state-dependent assortativity does not affect the substance of our results, although the analysis becomes more complex inducing us to focus on sufficient conditions for evolutionary stability. As an example, we also provide specific results regarding the stranger-in-the-night matching process (Bergstrom, 2013) which gives rise to state-dependent assortativity.

The paper is organized as follows. Section 2 discusses the related literature. Section 3 describes the basic model and provides the main result. Section 4 conducts welfare analysis. Section 5 extends the basline model. Section 6 provides a discussion and concludes.

2 Related literature

The works most closely related to our study are perhaps Bergstrom (2003) and Bergstrom (2013), who consider matching assortativity and show that, when assortativty is in actions,

it can crucially allow for the evolution of cooperation. More precisely, we follow the idea developed in Bergstrom (2003) who defines the *index of assortativity* of a matching process between two cooperators (and similarly for two defectors) as the difference between the probability that a cooperator is matched with another cooperator and the probability that a defector is matched with a cooperator. Importantly, Bergstrom (2013) also considers a state-dependent index of assortativty, exploring in particular the so called stranger-in-the-night matching process. Basically, we add on Berstrom's models by introducing two exogenous cultures (in the form of two distinct types) and allowing for cultural aversion.[2]

Alger (2010) and Alger and Weibull (2010, 2012) apply the index of assortativity not to actions but to types, in order to study the evolution of preferences for altruism. The work of Alger and Weibull (2013) is also related to the current paper. They consider a heterogeneous population in which agents with different cultural types carry different preference traits and they are matched assortatively according to their types. They show that moral preferences (i.e., agents whose preferences attach some extra value to the act of cooperating), and hence cooperation, can spread in the society. This is so because sufficiently strong type-assortativity allows agents of the moral type to internalize part of the benefits of cooperation, so that they obtain a higher payoff than the selfish type agents. One important difference between our approach and theirs is that we focus on action-assortativity, showing that the interplay between action-assortativity and cultural aversion gives rise to a unique phenomenon: type-monomorphic states. Another important difference is that we do not allow evolution on types, but only on actions. To put it differently, we do not allow for the evolution of preferences, as we focus on a shorter time horizon.[3]

The coexistence of cooperators and defectors together with some form of separation[4] between agents that adopt different actions has been already obtained theoretically in Bilancini and Boncinelli (2009), where the possibility to refuse to interact with those who defected in

[2]The effects of cultural aversion have been studied also in the context of social coordination between risk-dominant and payoff-dominant conventions (Bilancini and Boncinelli, 2015). The main finding is that, if cultural aversion is strong enough, then both conventions survive in the long run, with perfect correlation between culture and convention.

[3]The literature on indirect evolutionary approach considers a time horizon that is long enough for selection to be active on types. See, among others, Güth and Yaari (1992), Güth (1995), Bester and Güth (1998), McNamara et al. (1999), Sethi and Somanathan (2001), Ok and Vega-Redondo (2001), Van Veelen (2006), Dekel et al. (2007), Heifetz et al. (2007b,a), Kuran and Sandholm (2008), Akçay et al. (2009) and Wu (2015).

[4]In the study of the evolution of cooperation through group selection (see, among others, Traulsen and Nowak, 2006, Bowles, 2006, Van Veelen, 2009, Choi and Bowles, 2007) a kind of separation between agents is considered, which is however quite different from action-assortativty: individuals interact mostly within their group but their actions affect the likelihood that the group survives against other groups.

the past leads to an interaction structure where there is a segregated group of cooperators who leaves out all defectors. Wang et al. (2012) provide supporting experimental evidence. Rezaei and Kirley (2012) find similar results in a setup where links are automatically severed upon defection.

Finally, the literature on cultural transmission[5], which focuses on the socialization of culture from one generation to another one, is also partly related to our paper. See Cavalli-Sforza and Feldman (1981), Boyd and Richerson (1988), Bisin and Verdier (2001), Bisin et al. (2004), and Tabellini et al. (2008), among many others, for an explicit modelling of an inter-generational cultural transmission process, where cultural types evolve in an overlapping generation model. With respect to these models, we consider a shorter time-horizon: long enough to have auxiliary traits (actions) evolve endogenously, but not too long as to have cultural traits that convey an identity (types) and cultural aversion (cost of type-mismatch) exogenously given.

3 The baseline model

The model and the following analysis are built upon Bergstrom (2003). As a distinctive feature in our model, agents are heterogeneous in cultural types, and interactions between agents of different cultural types are costly.

3.1 The Prisoner Dilemma game

Consider a large population of agents who are repeatedly and randomly matched in pairs to engage in some pairwise interaction. There are two actions available to the agents, labelled by C and D. Action C stands for cooperation, it costs c to the agent who adopts it, and gives a benefit $b > c$ to the partner interacting with the agent. Action D stands for defection, it costs nothing to the agent who adopts it, and gives no benefit to the partner. The resulting payoff matrix is given by

[5]See Richerson and Boyd (2008) for a comprehensive review of the role of culture in the evolution of human behavior. See also Bowles (1998) for a discussion of the impact of institutions on the evolution of preferences.

	C	D
C	$b-c, b-c$	$-b, c$
D	$c, -b$	$0, 0$

which is known as the *Prisoner Dilemma with additive payoffs*.

3.2 Heterogeneous types and cultural aversion

Each agent in the population carries one of two cultural types, labelled by x and y. Hence, the population is divided into two cultural groups. An agent suffers a cost of cultural aversion d in case of type-mismatch, i.e., for interacting with an agent of a cultural type different from his own. The proportion of x-type agents over the whole population is β. Without loss of generality suppose that $\beta \in (0.5, 1)$. A population state is characterized by $s = (s_x, s_y) \in [0,1]^2$, where s_x denotes the fraction of x-type agents that are cooperators, and s_y is interpreted analogously.

Let us define:

$$\begin{aligned}
\eta_{x|C}(s) &= \frac{\beta s_x}{\beta s_x + (1-\beta)s_y}; \\
\eta_{y|C}(s) &= \frac{(1-\beta)s_y}{\beta s_x + (1-\beta)s_y}; \\
\eta_{x|D}(s) &= \frac{\beta(1-s_x)}{\beta(1-s_x) + (1-\beta)(1-s_y)}; \\
\eta_{y|D}(s) &= \frac{(1-\beta)(1-s_y)}{\beta(1-s_x) + (1-\beta)(1-s_y)}.
\end{aligned}$$

where $\eta_{x|C}(s)$ is the fraction of cooperators that are x-types in state s; $\eta_{x|D}(s)$, $\eta_{y|C}(s)$ and $\eta_{y|D}(s)$ are analogously interpreted.

3.3 Two-pool assortative matching process

In our daily lives, we tend to interact with those who act like us and avoid those who behave differently. To capture such a tendency, we adopt the two-pool assortative matching process with uniform assortativity by (Cavalli-Sforza and Feldman, 1981). Two-pool assortative

matching process is a random matching process such that every agent in the population is matched, with probability $p \in [0,1]$, with an agent choosing the same strategy as he does (i.e., he draws from an assortative pool), and with probability $1-p$ with a randomly selected agent (i.e., he draws from a random pool consisting of all individuals who who did not match from an assortative pool).

Given a population state s, the probabilities of matching among the agents are given as

$$\begin{aligned} Pr(C|C) &= p + (1-p)(\beta s_x + (1-\beta)s_y); \\ Pr(D|C) &= (1-p)(\beta(1-s_x) + (1-\beta)(1-s_y)); \\ Pr(D|D) &= p + (1-p)(\beta(1-s_x) + (1-\beta)(1-s_y)); \\ Pr(C|D) &= (1-p)(\beta(s_x) + (1-\beta)(s_y)); \end{aligned}$$

where $Pr(C|C)$ denote the probability that a cooperator is matched with another cooperator, $Pr(D|C)$ denote the probability that a cooperator is matched with a defector. $Pr(D|D)$ and $Pr(C|D)$ are analogously interpreted.

One can observe that $Pr(C|C)$ and $Pr(D|D)$ are increasing in p, while $Pr(D|C)$ and $Pr(C|D)$ are decreasing in p. Hence, probability p captures how unlikely that a cooperator is matched with a defector. Probability p is referred to as the index of assortativity in Bergstrom (2003).

Note that when $p = 0$, the matching process is reduced to a uniformly random matching process. When $p = 1$, the cooperators and the defectors are completely segregated from each other.

For now, we assume that p is a constant across population state as in Bergstrom (2003). In Subsection 5.4, we generalize the analysis to the case in which the index of assortativity is not uniform across states.

3.4 Evolutionarily Stable States

As justified in the Introduction, we assume that auxiliary traits (actions) evolve faster than identity traits (cultural types). Therefore, we focus on the evolution of actions taking types as given. We do so under the standard assumption of payoff monotonicity (Weibull, 1995), simply comparing the expected payoffs of cooperators and defectors.

In a setting like ours, where the distribution of types is held fixed while the distribution of actions is not, there is no agreement in the literature on the appropriate notion of evolu-

tionary stability.[6] Therefore, we take a conservative position opting for a quite demanding definition: one that ensures that a state is stable under any reasonable dynamics which satisfies payoff monotonicity.

We will say that a state s is ***evolutionarily stable*** if there exists an invasion barrier $\bar{\epsilon} > 0$ such that, for every pair $\epsilon_x, \epsilon_y \geq 0$, with $0 < \epsilon_x + \epsilon_y < \bar{\epsilon}$, describing the fraction of x-type mutants and y-type mutants, respectively, and for every pair (σ_x, σ_y) describing the fraction of cooperators among x-type mutants and y-type mutants, respectively, we have that mutants perform worse than the incumbents of the same type; in particular, if $\epsilon_x > 0$ and $\sigma_x \neq s_x$ then the average payoff of x-type mutants must be strictly lower than the average payoff of x-type incumbents, and if $\epsilon_y > 0$ and $\sigma_y \neq s_y$ then the average payoff of y-type mutants must be strictly lower than the average payoff of y-type incumbents. If a state is evolutionary stable, the fraction of mutants of each type will decrease over time in any payoff monotone dynamics.

We denote with $\pi(C, x|s)$ the expected payoff in population state s of a cooperator that is an x-type. We define $\pi(D, x|s)$, $\pi(C, y|s)$ and $\pi(D, y|s)$ analogously. For the ease of exposition, we define $\kappa = \beta s_x + (1 - \beta) s_y$.

For a x type cooperator, if he enters the assortative pool, he always gets a payoff of $b - c$ from the interaction. However, he may encounter a y type cooperator with probability $\eta_{y|C}(s)$, which costs him a penalty of d. On the other hand, if he instead enters the random pool, he only encounters another cooperator with probability κ. Moreover, among all the agents he can encounter in the random pool, $1 - \beta$ of them are y type agents. So he receives a penalty of d with probability $1 - \beta$. Therefore, the expected payoff of a x type agent is given by

$$\pi(C, x|s) = p(b - d\eta_{y|C}(s)) + (1 - p)[\kappa b - d(1 - \beta)] - c.$$

Similarly, we can write down the expected payoffs of the other three types of agents:

$$\begin{aligned}
\pi(D, x|s) &= -pd\eta_{y|D}(s) + (1 - p)[\kappa b - d(1 - \beta)], \\
\pi(C, y|s) &= p(b - d\eta_{x|C}(s)) + (1 - p)[\kappa b - d\beta] - c, \\
\pi(D, y|s) &= -pd\eta_{x|D}(s) + (1 - p)[\kappa b - d\beta].
\end{aligned}$$

[6] Evolutionary stability in incomplete information games with fixed distribution of types is studied in Ely and Sandholm (2005), who consider best-response dynamics, Cressman (2003, Section 4.7.2) and Amann and Possajennikov (2009), who apply replicator dynamics.

We observe that

$$\pi(C, x|s) - \pi(D, x|s) = p(b - d(\eta_{y|C}(s) - \eta_{y|D}(s))) - c,$$
$$\pi(C, y|s) - \pi(D, y|s) = p(b - d(\eta_{x|C}(s) - \eta_{x|D}(s))) - c.$$

Using the differences above, the following result can be proved:

PROPOSITION **1.**

- $s = (1, 1)$ *is an evolutionarily stable state if and only if* $p \geq \frac{c}{b-\beta d}$*;*
- $s = (0, 0)$ *is an evolutionarily stable state if and only if* $p \leq \frac{c}{b+\beta d}$*;*
- $s = (1, 0)$ *and* $s = (0, 1)$ *are evolutionarily stable states if and only if* $\frac{c}{b+d} \leq p \leq \frac{c}{b-d}$.

There are no other states that can be evolutionarily stable.

Figure 1 provides a graphical illustration of Proposition 1. As one can see, the range of values that p can take in $[0, 1]$ can be partitioned in five regions. If $p \in [0, c/(b + d))$, then $s = (0, 0)$ is the unique evolutionarily stable state. If $p \in [c/(b + d), c/(b + \beta d)]$, then $s = (0, 0)$, $s = (1, 0)$ and $s = (0, 1)$ are all and only the evolutionarily stable states. If $p \in (c/(b + \beta d, c/(b - \beta d))$, then $s = (1, 0)$ and $s = (0, 1)$ are all and only the evolutionarily stable states. If $p \in [c/(b - \beta d), c/(b - d)]$, then $s = (1, 1)$, $s = (1, 0)$ and $s = (0, 1)$ are all and only the evolutionarily stable states. Finally, if $p \in (c/(b - d), 1]$, then $s = (1, 1)$ is the unique evolutionarily stable state.

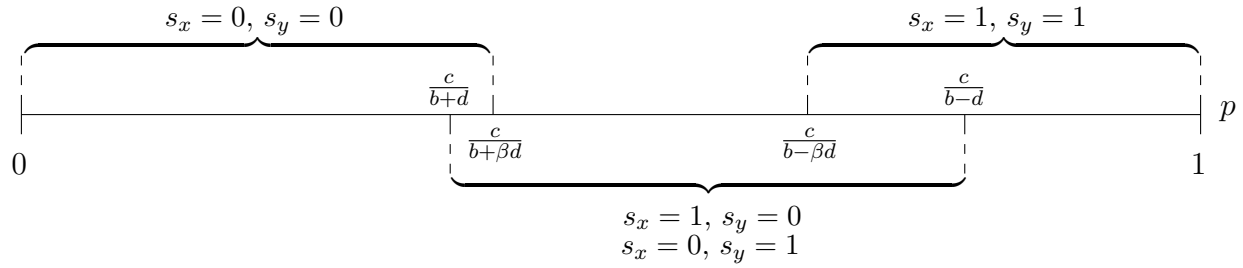


Figure 1: Evolutionarily stable states as a function of p. The picture is drawn assuming $b = 2$, $c = 1$, $d = 3/4$ and $\beta = 2/3$.

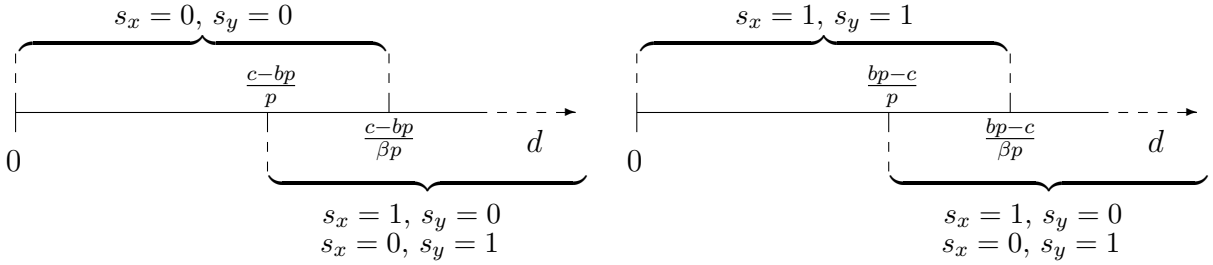
Note that when $d = 0$, if $p > \frac{c}{b}$, $s = (1, 1)$ is uniquely evolutionarily stable; if $p < \frac{c}{b}$, $s = (0, 0)$ is uniquely evolutionarily stable. In other words, without cultural aversion, people

either all cooperate when the matching is highly assortative, or all defect when the matching is less assortative.

However, when the cultural aversion d is positive, type-monomorphic states $s = (0, 1)$ and $s = (1, 0)$, in which partial cooperation is sustained, emerge as the unique evolutionarily stable states in the region of $p \in (c/(b + \beta d, c/(b - \beta d))$.

The existence of type-monomorphic states hinges on the interplay between cultural aversion and assortativity in actions. Cultural aversion introduces an incentive for the agents from the same cultural group to coordinate on the action played by most of their group members, because the presence of assortativity in actions help them to avoid being matched with agents from the other group. Eventually, the agents with different cultural types are sorted to coordinate on different actions. This sorting result is the main novelty of this model compared to the literature on the evolution of cooperation in prisoner dilemmas.

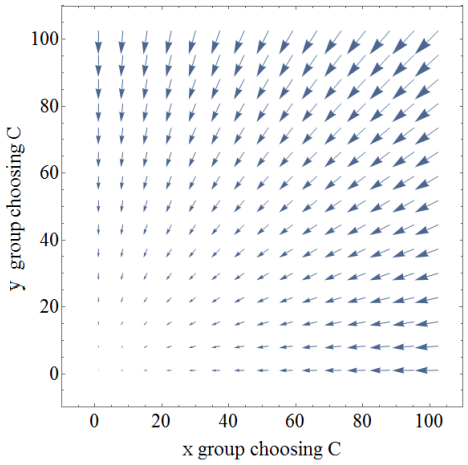
Figure 2 depicts the evolutionarily stable states as a function of d. When assortativity level is low ($p < \frac{c}{b}$), increasing cultural aversion helps to foster cooperation in a population consisting of only defectors. On the other hand, when assortativity is high ($p > \frac{c}{b}$), increasing cultural aversion induces defection in a population consisting of only cooperators. In Section 4, we provide a detailed analysis of the welfare effect induced by the cost of cultural aversion d.



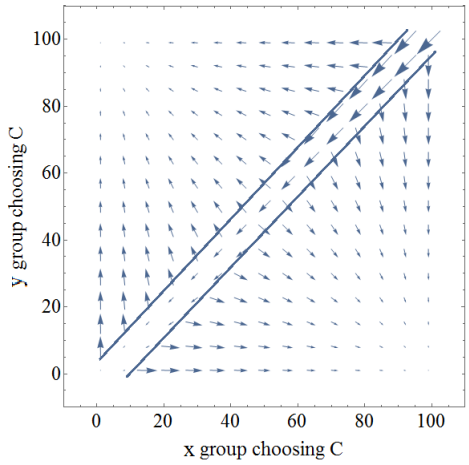
(a) The picture is drawn assuming $p < c/b$. (b) The picture is drawn assuming $p > c/b$.

Figure 2: Evolutionarily stable states as a function of d.

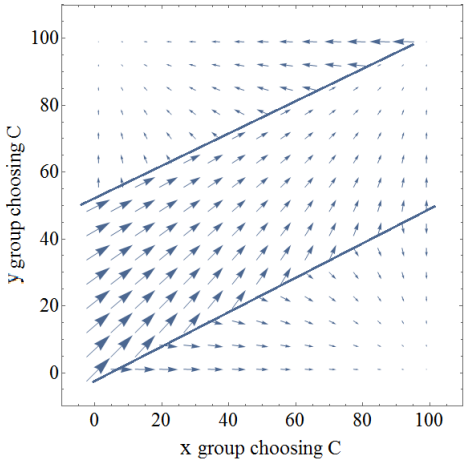
To give an idea of the kind of payoff-monotone dynamics at play here, in Figure 3 we depict the phase diagram under best response dynamic, for a numerical example. In particular, we set $\beta = 0.6$, $b = 3$, $c = 1$, $d = 1$ and p takes value from $0.2, 0.35, 0.4$ and 0.5. The x-axis represents the proportion of x group agents choosing to cooperate (s_x). The y-axis represents the proportion of y group agents choosing to cooperate (s_y). When $p = 0.2$, best response dynamic converges to $s = (0, 0)$ from any initial state. When $p = 0.5$, best



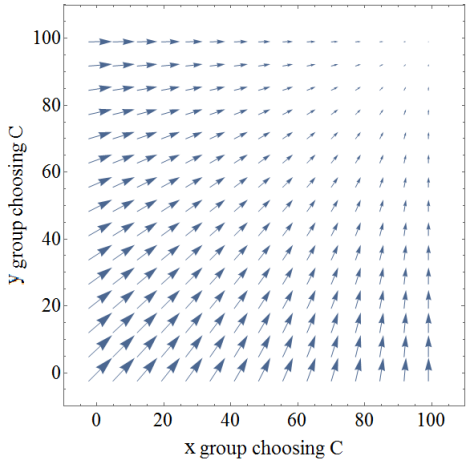
(a) $p = 0.2$



(b) $p = 0.35$



(c) $p = 0.4$



(d) $p = 0.5$

Figure 3: Plots of Best Responses

response dynamic converges to $s = (1, 1)$ from any initial state. When $p = 0.35$, The simplex of population states is divided into three regions, which define the basins of attractions of $s = (0, 0)$, $(0, 1)$, $(1, 0)$. The dynamic would converge to one of these three states depending on the initial states. When $p = 0.4$, The simplex of population states is divided into three regions, which define the basins of attractions of $s = (1, 1)$, $(0, 1)$, $(1, 0)$. The dynamic would converge to one of these three states depending on the initial states.

One final comment regards the role of the relative size of cultural groups. Note that, in the presence of cultural aversion, also β affects how action-assortativity determines the emergence of cooperation. In particular, as β gets closer to 1, i.e., as the two cultural groups have more and more unequal sizes, a greater action-assortativity is necessary for $s = (1, 1)$ to be evolutionarily stable and a smaller one for $s = (0, 0)$, while the stability of type-monomorphic states is unaffected. So, in a sense, the asymmetry in the size of cultural groups makes the emergence of type-monomorphic states more likely, at least for intermediate values of assorativity.

4 Welfare

We have shown how a larger cultural aversion can promote more or less cooperation in a society, depending on which state – either full cooperation or full defection – emerges in the absence of cultural aversion, which in turn depends on the degree of action-assortativity. The effects of cultural aversion, however, are not limited to changes in the extent of cooperation; they also involve modifications to the cost and frequency of type-mismatches. In order to evaluate the overall impact of cultural aversion, we focus our attention on societal welfare, which is simply the sum over of individual payoffs over the whole population.

We denote the societal welfareof a state $s = (s_x, s_y)$ with $W(s_x, s_y)$. In a monomorphic state societal welfare is $W(1, 1) = b - c - 2\beta(1 - \beta)d$ if everybody cooperates, while it is $W(0, 0) = -2\beta(1 - \beta)d$ if everybody defects. Instead,in a type-monoporphic state we have either that $W(1, 0) = \beta(b - c) - 2\beta(1 - \beta)(1 - p)d$ or that $W(0, 1) = (1 - \beta)(b - c) - 2\beta(1 - \beta)(1 - p)d$ depending on whether, respectively, the majority cooperates or the minority cooperates.

In any range of parameters where only one state is evolutionarily stable, the minimal cultural aversion is clearly optimal for any $p > 0$, since type-mismatches are costly, ceteris paribus, from a societal point of view. However, $d = 0$ is not necessarily the most desirable situation. Indeed, besides the cost of type mismatches, there are other two effects of cultural aversion on welfare that are not necessarily negative. These effects exist when cultural

aversion is strong enough that also type-monomorphic states are evolutionarily stable. One effect is always beneficial and is the incentive to coordinate on the same action that is adopted by one's own type, which can greatly reduce the number of type-mismatches; the strength of such effect crucially, and positively, depends on the degree of action-assortativity. The second effect is the result of the first: in order to avoid type mismatches an agent can be induced to adopt a different action; evidently, this effect is beneficial when the change is from defection to cooperation and detrimental when the change goes in the opposite direction.

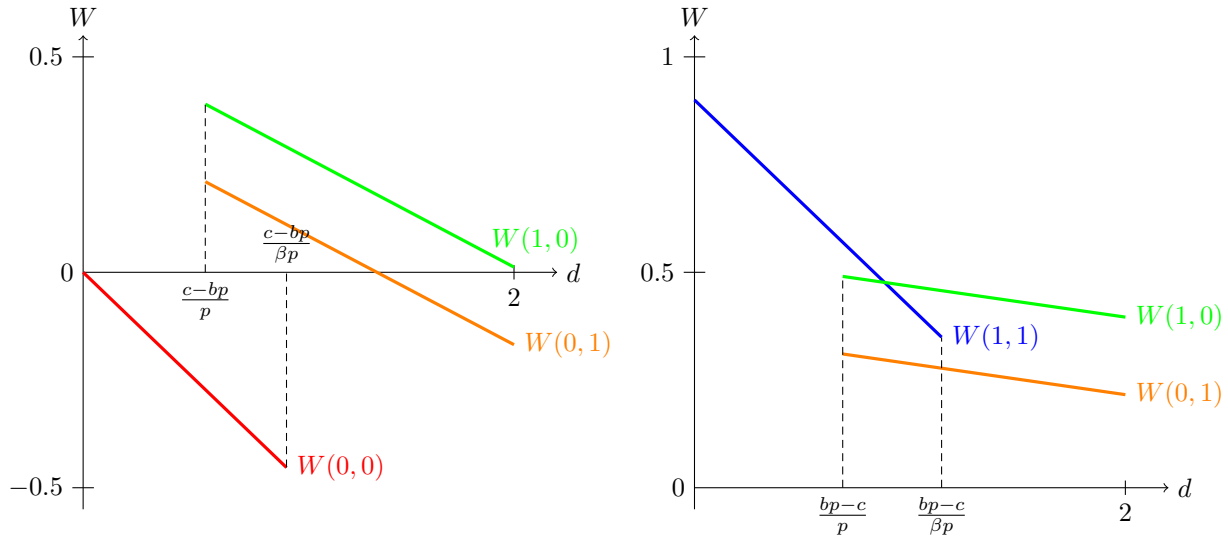


Figure 4: Welfare of evolutionarily stable states as a function of d. The pictures are drawn assuming $b = 2.1$, $c = 1.2$, $\beta = 0.6$, and $p = 0.45$ on the left, while $p = 0.85$ on the right.

When $p < c/b$, the monomorphic state that is evolutionarily stable is full defection, so a type-monomorphic state can be preferred as it allows more cooperation (see Figure 4, left panel, for an example). In particular, when both kinds of states are evolutionarily stable we have that $W(1,0) > W(0,1) > W(0,0)$ for all feasible values of d. So, if d is not too large, the type-monorphic state with some cultural aversion is preferable to the monomorphic state where everybody defects but there is no cultural aversion at all – and, of course, the type-monomorphic state where the majority cooperates is the most preferred. The gains from cooperation, although only a part of the population is involved, more than offset the increased cost of type mismatches, which however are reduced in number.

Instead, when $p > c/b$, it is obviously optimal to have $d = 0$, because the monomorphic state where everybody cooperates is evolutionarily stable – and full cooperation in the absence of cultural aversion is, by construction, the first best. However, if some cultural

aversion is present, and it is sufficiently strong to make also type-monomorphic states evolutionarily stable, it may happen that $W(1,0) > W(0,1) > W(1,1)$ (see Figure 4, right panel, for an example). If d is not too large, the lost benefits of cooperation (due to the fact that part of the population now defects to avoid type mismatches) can be more than offset by the reduced number of type-mismatches, even if the cost of a single type mismatch can be greater.

The following proposition summarizes:

PROPOSITION **2.** *Suppose that* $(s_x = 1, s_y = 0)$ *and* $(s_x = 0, s_y = 1)$ *are evolutionarily stable states. It follows that* $W(1,0) > W(0,1)$*. In addition, if they are not the only evolutionarily stable states, then we have that either:*

- $(s_x = 0, s_y = 0)$ *is also evolutionarily stable, with* $W(1,0) > W(0,1) > W(0,0)$*;*
- $(s_x = 1, s_y = 1)$ *is also evolutionarily stable, with* $W(1,0) > W(1,1)$ *if and only if* $dp > \frac{(b-c)}{2\beta}$*, and* $W(0,1) > W(1,1)$ *if and only if* $dp > \frac{(b-c)}{2(1-\beta)}$*.*

Proposition 2 indicates that cultural aversion and action-assortativity have a non-trivial interplay for what concerns their effect on welfare. In particular, when $d > 0$ it is possible that p has a non-monotonic impact on welfare. Figure 5 illustrates an examples of this case. We stress that for $d = 0$ this can not happen.

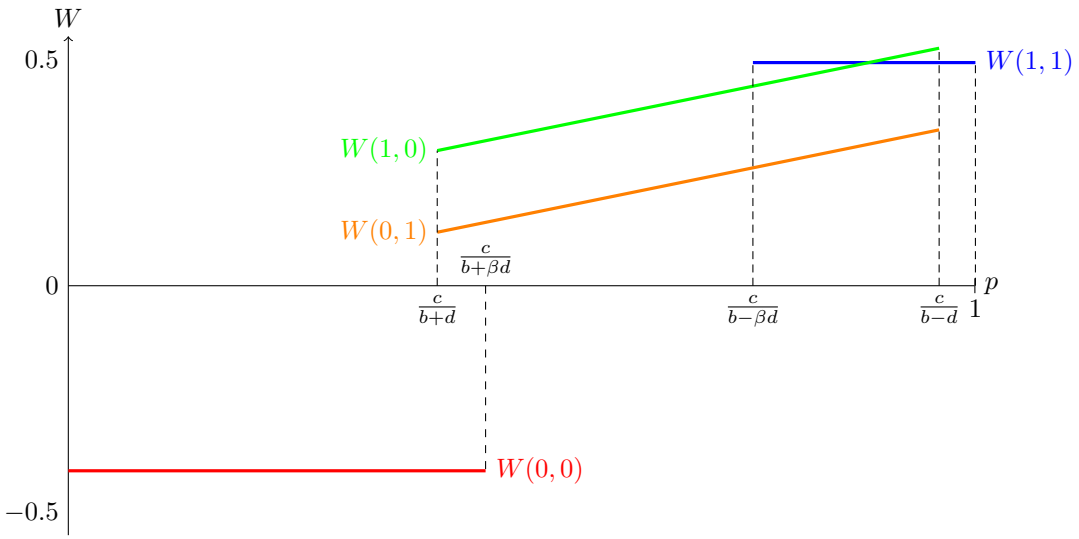


Figure 5: Welfare of evolutionarily stable states as a function of p. The picture is drawn assuming $b = 2.1$, $c = 1.2$, $d = 0.85$ and $\beta = 0.6$.

5 Model extensions

5.1 Asymmetric cultural aversion

Cultural aversion may be asymmetric. For example, as measured in Bisin et al. (2004), Jews' cultural intolerance towards Catholics and Protestants is much stronger than Catholics and Protestants' cultural intolerance towards each other. In this section, we study the consequences of how asymmetric cultural aversion.

Let d_x denote the cultural aversion of group x – i.e., the cost to be matched with a member of group y – and d_y denote the cultural aversion of group y – i.e., the cost to be matched with a member of group x – and suppose that $d_x \neq d_y$. Intuitively, $d_x > d_y$ captures the situation in which a conservative majority is hostile against a minority who is not so unhappy to work with the majority. Conversely, $d_x < d_y$ captures the situation in which a relatively open-minded majority faces an inward-looking minority.

Under asymmetric cultural aversion, the following result holds:

PROPOSITION **3.**

- $s = (1,1)$ *is an evolutionarily stable state if and only if* $p \geq \frac{c}{b - \max\{\beta d_y, (1-\beta) d_x\}}$;
- $s = (0,0)$ *is an evolutionarily stable state if and only if* $p \leq \frac{c}{b + \max\{\beta d_y, (1-\beta) d_x\}}$;
- $s = (1,0)$ *is an evolutionarily stable state if and only if* $\frac{c}{b+d_x} \leq p \leq \frac{c}{b-d_x}$;
- $s = (0,1)$ *is an evolutionarily stable state if and only if* $\frac{c}{b+d_y} \leq p \leq \frac{c}{b-d_y}$.

There are no other states that can be evolutionarily stable.

Compared to Proposition 1, Proposition 3 provides an additional insight to the problem. That is, asymmetry in cultural aversion serves as an equilibrium selection mechanism between the two type-mornomorphic states. For instance, when $d_x > d_y$ we have that for $c/(b+d_x) < p < c/(b+d_y)$ the type-monomorphic state $s = (1,0)$ is evolutionarily stable, while $s = (0,1)$ is not.

Also, Proposition 3 indicates that when action-assortativity increases, the group with stronger cultural aversion is likely to cooperate first. However, as action-assortativity keeps increasing, the group with weaker cultural aversion finds cooperation beneficial. This leads to defection in the group with stronger cultural aversion because of its members' strong desire to differentiate themselves from the other group.

5.2 Assortativity in cultural types

Besides the tendency to interact more with people who act similarly to us, we tend to interact more with those who are similar to ourselves along the cultural dimensions, such as language, religion, dress and origin. To capture the assortativity in both actions and cultural types, we develop the analysis in a variant of the two-pool assortative process, that we call *dual two-pool assortative process*. Every agent in the population enters with probability p a pool where all agents play his same action, and with probability $1-p$ he enters a random pool consisting of all individuals who did not enter the action-assortative pool. After that, every agent, whatever pool has entered, enters with probability q a sub-pool where all agents have his same type, and with probability $1-q$ he enters a random pool consisting of all individuals who did not enter the type-assortative pool.

We observe that p can still be understood as the index of assortativity in actions, while q can analogously be understood as the index of assortativity in cultural types.

We remind that $\kappa = \beta s_x + (1-\beta)s_y$. Expected payoffs at population state s are given by:

$$\begin{aligned}
\pi(C,x|s) &= pqb + p(1-q)(b - \eta_{y|C}(s)d) + (1-p)q\kappa b + (1-p)(1-q)(\kappa b - (1-\beta)d) - c,\\
\pi(D,x|s) &= pq0 + p(1-q)(-\eta_{y|D}(s)d) + (1-p)q\kappa b + (1-p)(1-q)(\kappa b - (1-\beta)d),\\
\pi(C,y|s) &= pqb + p(1-q)(b - \eta_{x|C}(s)d) + (1-p)q\kappa b + (1-p)(1-q)(\kappa b - \beta d) - c,\\
\pi(D,y|s) &= pq0 + p(1-q)(-\eta_{x|D}(s)d) + (1-p)q\kappa b + (1-p)(1-q)(\kappa b - \beta d).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\pi(C,x|s) - \pi(D,x|s) &= p(b - d(\eta_{y|C}(s) - \eta_{y|D}(s))(1-q)) - c,\\
\pi(C,y|s) - \pi(D,y|s) &= p(b - d(\eta_{x|C}(s) - \eta_{x|D}(s))(1-q)) - c.
\end{aligned}$$

Using the differences above, the following result can be proved:

Proposition **4.**

- $s = (1,1)$ *is an evolutionarily stable state if and only if* $p \geq \frac{c}{b-(1-q)\beta d}$*;*
- $s = (0,0)$ *is an evolutionarily stable state if and only if* $p \leq \frac{c}{b+(1-q)\beta d}$*;*
- $s = (1,0)$ *and* $s = (0,1)$ *are evolutionarily stable states if and only if* $p \geq \frac{c}{b+(1-q)d}$ *and* $p < \frac{c}{b-(1-q)d}$*.*

There are no other states that can be evolutionarily stable.

Proposition 4 is qualitatively similar to Proposition 1. From this we can conclude that our results, which are based on assortativity in actions, are robust to the presence of assortativity in types as well. However, we observe that the higher is q, the smaller is the interval of values of p for which type-monomorphic states are evolutionarily stable (and, also, the larger are the interval of values of p for which monomorphic states are evolutionarily stable). Indeed, a higher q reduces the net effect of actions as instruments to avoid type-mismatches.

Note that Proposition 4 implies that when $p = 0$, $s = (0, 0)$ is the only evolutionarily stable state. When $d = 0$, $s = (0, 0)$ and $s = (1, 1)$ are the only possible evolutionarily stable states. Therefore, assortativity in cultural types alone, without either cultural aversion or assortativity in actions, cannot produce type-mornomorphic stable states.

Let us conclude this discussion of the role of type-assortativity by pointing out a relationship that we do not considered here, but that is potentially interesting. One might argue that cultural aversion and the degree of type-assortativity are not independent one of the other. In particular, q might be an increasing function of d: the higher the cost of type-mismatches, the more effective will be the mechanisms developed by each culture to avoid interactions with members of other cultures. In such a case, which states are evolutionarily stable, as d varies, depends on how sensitive q is to changes in d. In principle, all our results remain valid, provided that q is sufficiently insensitive. We consider the study of specific mechanisms linking q to d as a promising direction for future research.

5.3 Legal institutions

There is a consensus that formal institutional arrangements such as legal sanctions are important for sustaining economic exchange and promoting trade and development in modern civic societies (North, 1981). However, in most of the developed countries nowadays, conflicts between different ethnic, racial and cultural groups are still commonly seen. So, in this subsection, we would like to investigate how a legal institution interacts with cultural aversion and action-assortativity in shaping people's behaviors.

Consider the following legal institution (see Tabellini et al., 2008, for a similar model of legal institution). Whenever a pair of matched agents is such that one cooperates and the other defects, the legal institution has a positive probability $r \in (0, 1)$ to detect the pair and learn this fact. If the pair is detected, the defecting agent faces a fine $f > c$ and the cooperating agent receives an equal amount of compensation f. The detection rate r measures the strength of the legal institution.

Note that this legal institution can also be regarded as a court system. Any agent can take his/her opponent to the court. When a cooperating agent takes a cooperating agent to

the court, their material payoffs are unchanged. When a cooperating agent takes a defecting agent to the court or the other way around, the defecting agent pays rf and the cooperating agent receives rf. When a defecting agent takes a defecting agent to the court, both incur rf. Given this court system, defecting agents never have incentive to take their opponents to the court, while cooperating agents have incentive to take their opponents to the court if and only if their opponents are defecting.

We remind that $\kappa = \beta s_x + (1-\beta)s_y$. Expected payoffs at population state s are given by:

$$\begin{aligned}
\pi(C,x|s) &= p(b - d\eta_{y|C}(s)) + (1-p)[\kappa b - d(1-\beta) + (1-\kappa)rf] - c,\\
\pi(D,x|s) &= -pd\eta_{y|D}(s) + (1-p)[\kappa(b-rf) - d(1-\beta)],\\
\pi(C,y|s) &= p(b - d\eta_{x|C}(s)) + (1-p)[\kappa b - d\beta + (1-\kappa)rf] - c,\\
\pi(D,y|s) &= -pd\eta_{x|D}(s) + (1-p)[\kappa(b-rf) - d\beta],
\end{aligned}$$

which lead to:

$$\begin{aligned}
\pi(C,x|s) - \pi(D,x|s) &= p(b - d(\eta_{y|C}(s) - \eta_{y|D}(s)) - rf) + rf - c,\\
\pi(C,y|s) - \pi(D,y|s) &= p(b - d(\eta_{x|C}(s) - \eta_{x|D}(s)) - rf) + rf - c.
\end{aligned}$$

Using the differences above, the following result can be proved:

PROPOSITION **5.** *If $rf < c$, we have that:*

- $(s_x = 1, s_y = 1)$ *is an evolutionarily stable state if and only if* $p \geq \frac{c-rf}{b-d\beta-rf}$ *and* $b - d\beta - rf > 0$*;*
- $(s_x = 0, s_y = 0)$ *is an evolutionarily stable state if and only if* $p \leq \frac{c-rf}{b+d\beta-rf}$*;*
- $(s_x = 1, s_y = 0)$ *and* $(s_x = 0, s_x = 1)$ *are both evolutionarily stable states if and only if, either* $\frac{c-rf}{b+d-rf} \leq p \leq \frac{c-rf}{b-d-rf}$ *in case* $b - d - rf > 0$*, or* $\frac{c-rf}{b+d-rf} \leq p$ *in case* $b - d - rf \leq 0$.

There are no other states that can be evolutionarily stable.

Compared to Proposition 1, Proposition 5 shows that a relatively weak legal institution ($qf < c$) complements action-assortativity in achieving cooperation. More specifically, given the presence of the legal institution, the thresholds on p for switching from the monomorphic state where everybody defects to a type-monomorphic state, and for switching from a type-monomorphic state to the monomorphic state where everybody cooperates, are both lower.

However, things turn out to change qualitatively when the legal institution is strong, as implied by the following result:

PROPOSITION **6.** *If $rf > c$, we have that:*[7]

- *if $b - d \geq c$, then $(s_x = 1, s_y = 1)$ is the unique evolutionarily stable state;*
- *if $b - d < c$, then:*
  - *$(s_x = 1, s_y = 1)$ is an evolutionarily stable state if and only if $p \leq \frac{rf-c}{rf-(b-d\beta)}$;*
  - *$(s_x = 1, s_y = 0)$ and $(s_x = 0, s_x = 1)$ are both evolutionarily stable states if and only if $p \geq \frac{rf-c}{rf-(b-d)}$;*
  - *there are no other states that can be evolutionarily stable.*

The first part of Proposition 6 shows that when cultural aversion is sufficiently weak, a sufficiently strict legal institution can completely substitute the role of action-assortativity and induce full cooperation.

On the other hand, the second part of Proposition 6 illustrates a counter-intuitive result: When cultural aversion is sufficiently strong, for a low degree of action-assortativity the only stable state is the monomorphic one where everybody cooperates, while a high degree of action-assortativity leads to the stability of type-monomorphic states. The rationale is that when action-assortativity is sufficiently high, the legal institution would not represent a very effective punishment for defectors, since they rarely end up matched with cooperators; instead, differentiating in actions is still an effective way to avoid costly type-mismatches.

People usually blame bad behaviors on weak formal institutional arrangement. However, in reality, we still observe high crime rates in certain minority groups in developed countries. Proposition 6 provides a possible explanation for why a strict legal institution is sometimes ineffective (Massey, 1995).

Finally, Proposition 5 and 6 indicate that a legal institution itself cannot produce type-monomorphic states if either cultural aversion or action-assortativity is absent ($d = 0$ or $p = 0$). This demonstrates the importance of cultural aversion and action-assortativity in our model.

5.4 State-dependent assortativity

So far we have considered a particular matching process, the two-pool matching process, that entails a state-independent index of assortativity. However, the index of assortativity

[7]Fundamentally, the same result holds for $rf = c$ too, provided that $p > 0$ and $b - d \neq c$. Indeed, it can be shown that, if $rf = c = b - d$ and $p > 0$ then $(s_x = 1, s_y = 1)$, $(s_x = 1, s_y = 0)$ and $(s_x = 0, s_x = 1)$ are all and only the evolutionarily stable states, while if $rf = c$ and $p = 0$ then no evolutionarily stable state exists.

can be, in general, state-dependent. In this subsection we show that the introduction of state-dependent assortativity does not affect the gist of our results.

We denote with $p(C|D,s)$ the probability that a defector will encounter a cooperator in state s. We define $p(C|C,s)$, $p(D|C,s)$ and $p(D|D,s)$ analogously. So, we have that:

$$\begin{aligned}
\pi(C,x|s) &= p(C|C,s)(b-d\eta_{y|C}(s))+p(D|C,s)(-d\eta_{y|D}(s))-c,\\
\pi(D,x|s) &= p(C|D,s)(b-d\eta_{y|C}(s))+p(D|D,s)(-d\eta_{y|D}(s)),\\
\pi(C,y|s) &= p(C|C,s)(b-d\eta_{x|C}(s))+p(D|C,s)(-d\eta_{x|D}(s))-c,\\
\pi(D,y|s) &= p(C|D,s)(b-d\eta_{x|C}(s))+p(D|D,s)(-d\eta_{x|D}(s)).
\end{aligned}$$

In this setup the index of assortativity in actions is given by:

$$a(s)=p(C|C,s)-p(C|D,s).$$

We observe that $p(C|C,s)-p(C|D,s)=p(D|D,s)-p(D|C,s)$. Therefore,

$$\begin{aligned}
\pi(C,x|s)-\pi(D,x|s) &= a(s)(b-d(\eta_{y|C}(s)-\eta_{y|D}(s)))-c,\\
\pi(C,y|s)-\pi(D,y|s) &= a(s)(b-d(\eta_{x|C}(s)-\eta_{x|D}(s)))-c.
\end{aligned}$$

Using the differences above, the following result can be proved:

PROPOSITION **7.**

- $(s_x=1,s_y=1)$ *is an evolutionarily stable state if* $a(s)>\frac{c}{b-\beta d}$*;*
- $(s_x=0,s_y=0)$ *is an evolutionarily stable state if* $a(s)<\frac{c}{b+\beta d}$*;*
- $(s_x=1,s_y=0)$ *and* $(s_x=0,s_y=1)$ *are evolutionarily stable states if* $\frac{c}{b+d}<a(s)<\frac{c}{b-d}$.

There are no other states that can be evolutionarily stable.

The results in Proposition 7 involve inequalities where the index of assortativity is asked to be larger or smaller than some thresholds. Since the index $a(\cdot)$ is now state-dependent, the possibility for a state to be evolutionarily stable hinges on the actual specification of $a(\cdot)$. In particular, specific values of the index of assortativity should be derived from the underlying matching process. In the following we explore a specific matching process that is alternative to the two-pool matching process and which entails a state-dependent index of assortativity: the strangers-in-the-night matching process with uniform assortativity (Bergstrom, 2013).

The strangers-in-the-night matching process is such that every agent in the population is randomly matched with another agent, and the pair is actually formed to play the game with probability n if the two agents are *alike* in actions, and with probability m if the two agents are *different* in actions. We assume, as typical, that $n > m$.

To adapt a strangers-in-the-night matching process with uniform assortativity in actions to our setup, we particularizing eq. 20 in Bergstrom (2013) to the following, for every state s:

$$a(s) = \frac{[\beta s_x + (1-\beta)s_y][\beta(1-s_x) + (1-\beta)(1-s_y)](n^2 - m^2)}{[\beta s_x + (1-\beta)s_y][\beta(1-s_x) + (1-\beta)(1-s_y)](n-m)^2 + nm}. \tag{1}$$

The next result follows directly from Proposition 7 in the light of (1).

PROPOSITION **8.**

- $(s_x = 0, s_y = 0)$ *is always an evolutionarily stable state;*
- $(s_x = 1, s_y = 0)$ *and* $(s_x = 0, s_y = 1)$ *are evolutionarily stable states if* $\frac{c}{b+d} < \frac{\beta(1-\beta)(n^2-m^2)}{\beta(1-\beta)(n-m)^2+nm} < \frac{c}{b-d}$.

There are no other states that can be evolutionarily stable.

It can be useful to contrast this result with Bergstrom (2013, Theorem 6), where coexistence of cooperators and defectors is also obtained in an evolutionarily stable state. Our analysis indicates that, when type-heterogeneity is added together with cultural aversion in case of type-mismatch, coexistence of cooperation and defection is no longer evolutionarily stable, unless it comes with the separation of cooperators and defectors on the basis of their cultural types.

One further comment is worth doing here. Action-assortativty is meant to describe people's general tendency to interact more with those who act like them. In our analysis, as in the one by Begstrom, action-assortativity is taken as primitive. However, one may think of mechanisms that can justify the existence and working of action-assortativity. Let us provide a couple of examples.

Assortative matching can be generated by a process of repeated social interactions. Suppose that the agents come with a selcted action, either cooperate or defect. Then, they engage in n rounds of social interactions, where n can be undestood as a measure of the intensity of social interactions. In the first round, all agents are uniformly randomly matched in pairs. If both agents in a pair agree to keep the pair, they play the game and get the payoffs. Otherwise, the agents would go back into the matching pool for the next round. The same process repeats for the second round and so on. In the last round, the remaining

agents are uniformly randomly matched and they have to accept their matches. Different intensities of social interactions can generate different degrees of assortativity. Intuitively, larger n induces higher degree of assortativity because cooperators matched with defectors in early rounds would prefer breaking the pair and enter the next round in the hope to match another cooperator. Note that this would be true also if agents are impatient, provided that a greater n implies more frequent pairings.

Assortative matching can also be regarded as an institutional arrangement. Gunnthorsdottir et al. (2010) propose the so called "group-based meritocracy mechanism" (GBM) as opposed to "voluntary contribution mechanism" (VCM). GBM matches people according to their contributions in a public provision game and they show that GBM induces higher contributions than VCM in a laboratory setting. Nax et al. (2014) argue that assortative matching as a meritocratic institution has been adopted since early human civilizations. Nowadays, Chinese civil service exam, honorary circles, bonus wage schemes can all be considered as assortative matching. Rigos and Nax (2015) endogenize assortative as a variable determined by democratic consensus.

6 Final remarks

In this paper we have explored the interplay between assortativty in actions and cultural aversion. We have found that, in the presence of action-assortativty, cultural aversion works in favor of states where there is perfect correlation between culture and behavior. This happens because, thanks to action-assortativity, cooperation and defection can work as instruments to avoid interaction with individuals of another culture.

Further, our results account for the coexistence of a group of cooperators and a group of defectors, as well as the segregation in actions of the two cultural groups, which is a pattern that can be observed in reality. For example, social scientists have long documented and analyzed the hyper-segregation phenomenon in the United States (Massey and Denton, 1993, Cutler et al., 1999). Certain cultural groups experienced high levels of segregation and developed a persistent "ghetto culture", which is associated with poor work ethics (Hannerz, 1969, Lewis, 1969, Wilson, 1987, 1996, Lemman, 1991, Bonney, 1975, Sáez-Martı and Zenou, 2012). Our model provides an explanation for the emergence such a "ghetto culture" as a consequence of strategic choice: because of action-assortativity, shirking in effort (defect) prevents a member of these cultural groups from being estranged from their own cultural group members. However, at the same time, choosing to defect separates them from good job opportunities (matching with cooperators), which induces high unemployment and poverty.

References

Ahern, K. R., D. Daminelli, and C. Fracassi (2012). Lost in translation? the effect of cultural values on mergers around the world. *Journal of Financial Economics*.

Akçay, E., J. Van Cleve, M. W. Feldman, and J. Roughgarden (2009). A theory for the evolution of other-regard integrating proximate and ultimate perspectives. *Proceedings of the National Academy of Sciences 106*(45), 19061–19066.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *Quarterly journal of Economics*, 715–753.

Akerlof, G. A. and R. E. Kranton (2005). Identity and the economics of organizations. *Journal of Economic perspectives*, 9–32.

Alger, I. (2010). Public goods games, altruism, and evolution. *Journal of Public Economic Theory 12*(4), 789–813.

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

Alger, I. and J. W. Weibull (2012). A generalization of hamilton's rule: Love others how much? *Journal of Theoretical Biology 299*, 42–54.

Alger, I. and J. W. Weibull (2013). Homo moralis: Preference evolution under incomplete information and assortative matching. *Econometrica 81*(6), 2269–2302.

Amann, E. and A. Possajennikov (2009). On the stability of evolutionary dynamics in games with incomplete information. *Mathematical Social Sciences 58*(3), 310–321.

Bandiera, O., I. Barankay, and I. Rasul (2009). Social connections and incentives in the workplace: Evidence from personnel data. *Econometrica 77*(4), 1047–1094.

Becker, G. S. (1957). The economics of discrimination.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, 385–409.

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

Bergstrom, T. C. (2013). Measures of assortativity. *Biological Theory 8*(2), 133–141.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg more employable than Jamal? a field experiment on labor market discrimination. *American Economic Review 94*(4), 991–1013.

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

Bilancini, E. and L. Boncinelli (2009). The co-evolution of cooperation and defection under local interaction and endogenous network formation. *Journal of Economic Behavior & Organization 70*(1), 186–195.

Bilancini, E. and L. Boncinelli (2015). Social coordination with locally observable types.

Bisin, A., G. Topa, and T. Verdier (2004). Religious intermarriage and socialization in the united states. *Journal of political Economy 112*(3), 615–664.

Bisin, A. and T. Verdier (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic theory 97*(2), 298–319.

Bonney, N. (1975). Work and ghetto culture. *British Journal of Sociology*, 435–447.

Bowles, S. (1998). Endogenous preferences: The cultural consequences of markets and other economic institutions. *Journal of economic literature 36*(1), 75–111.

Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science 314*(5805), 1569–1572.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences 100*(6), 3531–3535.

Boyd, R. and P. J. Richerson (1988). *Culture and the evolutionary process*. University of Chicago Press.

Boyd, R. and P. J. Richerson (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences 364*(1533), 3281–3288.

Boyd, R., P. J. Richerson, and J. Henrich (2011). Rapid cultural adaptation can facilitate the evolution of large-scale cooperation. *Behavioral ecology and sociobiology 65*(3), 431–444.

Bramoullé, Y., S. Currarini, M. O. Jackson, P. Pin, and B. W. Rogers (2012). Homophily and long-run integration in social networks. *Journal of Economic Theory 147*(5), 1754–1786.

Cavalli-Sforza, L. L. and M. W. Feldman (1981). *Cultural transmission and evolution: a quantitative approach*. Number 16. Princeton University Press.

Choi, J.-K. and S. Bowles (2007). The coevolution of parochial altruism and war. *Science 318*(5850), 636–640.

Cressman, R. (2003). *Evolutionary dynamics and extensive form games*, Volume 5. MIT Press.

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

Currarini, S., M. O. Jackson, and P. Pin (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences 107*(11), 4857–4861.

Cutler, D. M., E. L. Glaeser, and J. L. Vigdor (1999). The rise and decline of the american ghetto. *Journal of Political Economy 107*(3).

Dekel, E., J. C. Ely, and O. Yilankaya (2007). Evolution of preferences. *The Review of Economic Studies 74*(3), 685–704.

Ely, J. C. and W. H. Sandholm (2005). Evolution in bayesian games i: theory. *Games and Economic Behavior 53*(1), 83–109.

Fisman, R., Y. Hamao, and Y. Wang (2014). Nationalism and economic exchange: Evidence from shocks to sino-japanese relations. *Review of Financial Studies*, hhu017.

Giannetti, M. and Y. Yafeh (2012). Do cultural differences between contracting parties matter? evidence from syndicated bank loans. *Management Science 58*(2), 365–383.

Guiso, L., P. Sapienza, and L. Zingales (2009). Cultural biases in economic exchange? *Quarterly Journal of Economics*, 1095–1131.

Gunnthorsdottir, A., R. Vragov, S. Seifert, and K. McCabe (2010). Near-efficient equilibria in contribution-based competitive grouping. *Journal of Public Economics 94*(11), 987–994.

Güth, W. (1995). An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *International Journal of Game Theory 24*(4), 323–344.

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

Hannerz, U. (1969). Soulside: Inquiries into ghetto culture and community.

Heifetz, A., C. Shannon, and Y. Spiegel (2007a). The dynamic evolution of preferences. *Economic Theory 32*(2), 251–286.

Heifetz, A., C. Shannon, and Y. Spiegel (2007b). What to maximize if you must. *Journal of Economic Theory 133*(1), 31–57.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization 53*(1), 3–35.

Henrich, J. and R. Boyd (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of theoretical biology 208*(1), 79–89.

Kuran, T. and W. H. Sandholm (2008). Cultural integration and its discontents. *The Review of Economic Studies 75*(1), 201–228.

Lemman, N. (1991). The promised land: The great black migration and how it changed america. new york: Alfred a.

Lewis, O. (1969). Culture of poverty. In D. P. Moynihan (Ed.), *On Understanding Poverty*, pp. 201–213. London., Basic Books, Inc.

Massey, D. S. (1995). Getting away with murder: Segregation and violent crime in urban america. *University of Pennsylvania Law Review*, 1203–1232.

Massey, D. S. and N. A. Denton (1993). *American apartheid: Segregation and the making of the underclass*. Harvard University Press.

McNamara, J. M., C. E. Gasson, and A. I. Houston (1999). Incorporating rules for responding into evolutionary games. *Nature 401*(6751), 368–371.

McPherson, M., L. Smith-Lovin, and J. M. Cook (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, 415–444.

Michaels, G. and X. Zhi (2010). Freedom fries. *American Economic Journal: Applied Economics 2*(3), 256–281.

Nax, H. H., R. O. Murphy, and D. Helbing (2014). Stability and welfare of'merit-based'group-matching mechanisms in voluntary contribution game. *Available at SSRN 2404280*.

North, D. C. (1981). *Structure and change in economic history*. Norton.

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

Rezaei, G. and M. Kirley (2012). Dynamic social networks facilitate cooperation in the n-player prisoners dilemma. *Physica A: Statistical Mechanics and its Applications 391*(23), 6199–6211.

Richerson, P. J. and R. Boyd (2008). *Not by genes alone: How culture transformed human evolution*. University of Chicago Press.

Rigos, A. and H. H. Nax (2015). Assortativity evolving from social dilemmas. Technical report, Department of Economics, University of Leicester.

Ruef, M., H. E. Aldrich, and N. M. Carter (2003). The structure of founding teams: Homophily, strong ties, and isolation among us entrepreneurs. *American sociological review*, 195–222.

Sáez-Martı, M. and Y. Zenou (2012). Cultural transmission and discrimination. *Journal of Urban Economics 72*(2), 137–146.

Sethi, R. and E. Somanathan (2001). Preference evolution and reciprocity. *Journal of economic theory 97*(2), 273–297.

Tabellini, G. et al. (2008). The scope of cooperation: Values and incentives. *Quarterly Journal of Economics 123*(3), 905–950.

Traulsen, A. and M. A. Nowak (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences 103*(29), 10952–10955.

Van Veelen, M. (2006). Why kin and group selection models may not be enough to explain human other-regarding behaviour. *Journal of theoretical biology 242*(3), 790–797.

Van Veelen, M. (2009). Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong. *Journal of Theoretical Biology 259*(3), 589–600.

Walster, E., V. Aronson, D. Abrahams, and L. Rottman (1966). Importance of physical attractiveness in dating behavior. *Journal of personality and social psychology 4*(5), 508.

Wang, J., S. Suri, and D. J. Watts (2012). Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences 109*(36), 14363–14368.

Weibull, J. (1995). Evolutionary game theory.

Wilson, W. (1996). The world of the new urban poor. alfred a.

Wilson, W. J. (1987). *The truly disadvantaged: The inner city, the underclass, and public policy*. University of Chicago Press.

Wu, J. (2015). Social connections and cultural heterogeneity.

A Appendix - Proofs

In the following we collect the proofs of all Propositions in the paper. For the sake of brevity, wherever possible we avoid repeating similar arguments developed in other proofs, limiting ourselves to highlight what adjustments have be done to prove the desired results (see, e.g., the proof of Proposition 1).

A.1 Proof of Proposition 1

The validity of Proposition 1 follows from the proof of Proposition 3 if we set $d_x = d_y = d$.

A.2 Proof of Proposition 2

Suppose that $(s_x = 1, s_y = 0)$ and $(s_x = 0, s_y = 1)$ are evolutionarily stable states. We simply note that $W(1,0) - W(0,1) = (2\beta - 1)(b - c) > 0$ for $\beta > 0.5$, which shows the first claim of the proposition.

Suppose that $(s_x, s_y) = (0,0)$ is also evolutionarily stable. To show the validity of the first bullet, it is enough to observe that $W(0,1) - W(0,0) = (1-\beta)(b-c) + 2\beta(1-\beta)pd > 0$, since $b > c$.

Suppose instead that $(s_x, s_y) = (1,1)$ is also evolutionarily stable. We note that $W(1,0) - W(1,1) = -(1-\beta)(b-c) + 2\beta(1-\beta)dp > 0$ requires $dp > \frac{(b-c)}{2\beta}$, while $W(0,1) - W(1,1) = -\beta(b-c) + 2\beta(1-\beta)dp > 0$ requires $dp > \frac{(b-c)}{2(1-\beta)}$, i.e., the two conditions in the second bullet.

A.3 Proof of Proposition 3

Preliminarily, we argue that to prove that a monomorphic or a type-monomorphic state is be evolutionarily stable, it is sufficient to check for pure invasions, i.e., $\sigma_x, \sigma_y \in \{0,1\}$, that is for invasions such that all x-type mutants are either only cooperators or only defectors and similarly for y-type mutants.

To see why, consider a generic state $s = (s_x, s_y)$ and an invasion of $\tilde{\epsilon}_x$ and $\tilde{\epsilon}_y$ mutants such that $0 < \tilde{\sigma}_x, \tilde{\sigma}_y < 1$. Hence, the new state $s' = (s'_x, s'_y)$ is such that $s'_x = s_x(1 - \tilde{\epsilon}_x/\beta) + \tilde{\sigma}_x\tilde{\epsilon}_x/\beta$ and $s'_y = s_x(1 - \tilde{\epsilon}_y/(1-\beta)) + \tilde{\sigma}_y\tilde{\epsilon}_y/(1-\beta)$. Consider (without loss of generality) the case where $\tilde{\sigma}_x > s_x$ and $\tilde{\sigma}_y < s_y$. Since the $\tilde{\epsilon}_x$ mutants are $\tilde{\sigma}_x\tilde{\epsilon}_x$ cooperators and $(1-\tilde{\sigma}_x)\tilde{\epsilon}_x$ defectors, and similarly for the $\tilde{\epsilon}_y$ mutants, we can rewrite the new state as $s'_x = s_x + \epsilon_x/\beta$ and $s'_y = s_y + \epsilon_y/(1-\beta)$, where $\epsilon_x = (\tilde{\sigma}_x - s_x)\tilde{\epsilon}_x$ and $\epsilon_y = (\tilde{\sigma}_y - s_y)\tilde{\epsilon}_y$. We observe that the expected payoff of $\tilde{\epsilon}_x - \epsilon_x$ mutants, whose fraction of cooperators is s_x, is the same as the incumbents of type x. Similarly, the expected payoff of $\tilde{\epsilon}_y - \epsilon_y$ mutants, whose fraction of cooperators is s_y, is the same as the incumbents of type y. Therefore, the success of the invasion crucially depends on whether ϵ_x cooperators and ϵ_y defectors earn a greater expected payoff than their respective incumbent types. From this observation follows that the original invasion is successful if and only if a smaller pure invasion of ϵ_x and ϵ_y mutants, with $\sigma_x = 1$ and $\sigma_y = 0$, is successful.

We consider the state $s = (s_x, s_y) = (1,1)$, and we suppose that a small fraction $\epsilon = \epsilon_x + \epsilon_y$ of mutants invades and a state $(s_x - \frac{\epsilon_x}{\beta}, s_y - \frac{\epsilon_y}{1-\beta})$ is reached. The following expressions (2) and (3) are, respectively, the relative gain that y-type cooperators have over y-type defectors, and the relative gain that x-type cooperators have over x-type defectors:

$$\pi(C, y|(s_x - \frac{\epsilon_x}{\beta}, s_y - \frac{\epsilon_y}{1-\beta})) - \pi(D, y|(s_x - \frac{\epsilon_x}{\beta}, s_y - \frac{\epsilon_y}{1-\beta})) = p\left(b - \frac{\beta - \epsilon_x}{1 - \epsilon_x - \epsilon_y}d_y + \frac{\epsilon_x}{\epsilon_x + \epsilon_y}d_y\right) - c; \quad (2)$$

$$\pi(C,x|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(D,x|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))=p\left(b-\frac{1-\beta-\epsilon_y}{1-\epsilon_x-\epsilon_y}d_x+\frac{\epsilon_y}{\epsilon_x+\epsilon_y}d_x\right)-c. \quad (3)$$

The worst case for expression (2) to be positive is when $\epsilon_x = 0$ and $\epsilon_y = \epsilon$. Analogously, the worst case for expression (3) to be positive is when $\epsilon_y = 0$ and $\epsilon_x = \epsilon$. Hence, it is easy to check that there exists an invasion barrier $\bar{\epsilon} > 0$ such that for any (ϵ_x, ϵ_y), with $\epsilon_x \geq 0$, $\epsilon_y \geq 0$, and $0 < \epsilon_x + \epsilon_y < \bar{\epsilon}$, expressions (2) and (3) are both positive if and only if $p \geq \frac{c}{b-\max\{\beta d_y,(1-\beta)d_x\}}$. This shows the validity of the first bullet in the statement of the proposition.

We now consider the state $s = (s_x, s_y) = (0, 0)$, and we suppose that a small fraction $\epsilon = \epsilon_x + \epsilon_y$ of mutants invades and a state $(s_x + \frac{\epsilon_x}{\beta}, s_y + \frac{\epsilon_y}{1-\beta})$ is reached. The following expressions (4) and (5) are, respectively, the relative gain that y-type defectors have over y-type cooperators, and the relative gain that x-type defectors have over x-type cooperators:

$$\pi(D,y|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,y|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(-b-\frac{\beta-\epsilon_x}{1-\epsilon_x-\epsilon_y}d_y+\frac{\epsilon_x}{\epsilon_x+\epsilon_y}d_y\right)+c; \quad (4)$$

$$\pi(D,x|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,x|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(-b-\frac{1-\beta-\epsilon_y}{1-\epsilon_x-\epsilon_y}d_x+\frac{\epsilon_y}{\epsilon_x+\epsilon_y}d_x\right)+c. \quad (5)$$

The worst case for expression (4) to be positive is when $\epsilon_x = 0$ and $\epsilon_y = \epsilon$. Analogously, the worst case for expression (5) to be positive is when $\epsilon_y = 0$ and $\epsilon_x = \epsilon$. Hence, it is easy to check that there exists an invasion barrier $\bar{\epsilon} > 0$ such that for any (ϵ_x, ϵ_y), with $\epsilon_x \geq 0$, $\epsilon_y \geq 0$, and $0 < \epsilon_x + \epsilon_y < \bar{\epsilon}$, expressions (4) and (5) are both positive if and only if $p \leq \frac{c}{b+\max\{\beta d_y,(1-\beta)d_x\}}$, which shows the validity of the second bullet in the statement of the proposition.

We then consider the state $s = (s_x, s_y) = (1, 0)$, and we suppose that a small fraction $\epsilon = \epsilon_x + \epsilon_y$ of mutants invades and a state $(s_x - \frac{\epsilon_x}{\beta}, s_y + \frac{\epsilon_y}{1-\beta})$ is reached. The following expressions (6) and (7) are, respectively, the relative gain that y-type defectors have over y-type cooperators, and the relative gain that x-type cooperators have over x-type defectors:

$$\pi(D,y|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,y|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(-b+\frac{\beta-\epsilon_x}{\beta-\epsilon_x+\epsilon_y}d_y-\frac{\epsilon_x}{1-\beta+\epsilon_x-\epsilon_y}d_y\right)+c; \quad (6)$$

$$\pi(C,x|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(D,x|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(b+\frac{1-\beta-\epsilon_y}{1-\beta-\epsilon_y+\epsilon_x}d_x-\frac{\epsilon_y}{\beta-\epsilon_x+\epsilon_y}d_x\right)-c. \quad (7)$$

The worst case for expression (6) to be positive is when $\epsilon_y = 0$ and $\epsilon_x = \epsilon$. To see why, plug $\epsilon_y = \epsilon - \epsilon_x$ into RHS of expression (6) and differentiate it with respect to ϵ_x. One can check that when ϵ is sufficiently small, the derivative is negative. Analogously, the worst case for expression (7) to be positive is when $\epsilon_x = 0$ and $\epsilon_y = \epsilon$. Hence, it is easy to check that there exists an invasion barrier $\bar{\epsilon} > 0$ such that for any (ϵ_x, ϵ_y), with $\epsilon_x \geq 0$, $\epsilon_y \geq 0$, and $0 < \epsilon_x + \epsilon_y < \bar{\epsilon}$, expressions (6) and (7) are both positive for any $\epsilon > 0$ small enough if and only if $\frac{c}{b-d_x} \geq p \geq \frac{c}{b+d_x}$, which shows the validity of the third bullet in the statement of the proposition.

Lastly, we consider the state $s = (s_x, s_y) = (0, 1)$, and we suppose that a small fraction $\epsilon = \epsilon_x + \epsilon_y$ of mutants invades and a state $(s_x + \frac{\epsilon_x}{\beta}, s_y - \frac{\epsilon_y}{1-\beta})$ is reached. The following expressions (8) and (9) are, respectively, the relative gain that y-type cooperators have over y-type defectors, and the relative gain that x-type defectors have over x-type cooperators:

$$\pi(C,y|(s_x+\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(D,y|(s_x+\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))=p\left(b+\frac{\beta-\epsilon_x}{\beta-\epsilon_x+\epsilon_y}d_y-\frac{\epsilon_x}{1-\beta+\epsilon_y-\epsilon_x}d_y\right)-c, \quad (8)$$

$$\pi(D,x|(s_x+\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(C,x|s_x+\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))=p\left(-b+\frac{1-\beta-\epsilon_y}{1-\beta-\epsilon_y+\epsilon_x}d_x-\frac{\epsilon_y}{\beta-\epsilon_x+\epsilon_y}d_x\right)+c, \tag{9}$$

The worst case for expression (8) to be positive is when $\epsilon_y = 0$ and $\epsilon_x = \epsilon$. Analogously, the worst case for expression (9) to be positive is when $\epsilon_x = 0$ and $\epsilon_y = \epsilon$. Hence, it is easy to check that there exists an invasion barrier $\bar{\epsilon} > 0$ such that for any (ϵ_x, ϵ_y), with $\epsilon_x \geq 0$, $\epsilon_y \geq 0$, and $0 < \epsilon_x + \epsilon_y < \bar{\epsilon}$, expressions (8) and (9) are both positive if and only if $\frac{c}{b-d_y} \geq p \geq \frac{c}{b+d_y}$, which shows the validity of the third bullet in the statement of the proposition.

Finally, we show that no other state can ever be evolutionarily stable. Ad absurdum, suppose that a state (s_x, s_y) is evolutionarily stable and that $s_x \in (0,1)$ (this is without loss of generality). In such a state the expected payoff of x-type cooperators must be equal to the expected payoff of x-type defectors. Now consider an invasion of ϵ mutants, all being x-type cooperators, i.e., $\epsilon_x = \epsilon$ and $\epsilon_y = 0$. Denote with $(s_x+\frac{\epsilon}{\beta}, s_y)$ the resulting state. From $\pi(C,x|(s_x,s_y)) = \pi(D,x|(s_x,s_y))$, it follows that $p(b-d_x(\eta_{y|C}(s_x,s_y)-\eta_{y|D}(s_x,s_y)))-c=0$. Since $\eta_{y|C}(s_x+\frac{\epsilon}{\beta},s_y)-\eta_{y|D}(s_x+\frac{\epsilon}{\beta},s_y) < \eta_{y|C}(s_x,s_y)-\eta_{y|D}(s_x,s_y)$, we have that, for any $\epsilon > 0$, an x-type cooperator obtains a strictly greater payoff than a x-type defector, i.e.:

$$\pi(C,x|(s_x+\frac{\epsilon}{\beta},s_y))-\pi(D,x|(s_x+\frac{\epsilon}{\beta},s_y))=p(b-d_x(\eta_{y|C}(s_x+\frac{\epsilon}{\beta},s_y)-\eta_{y|D}(s_x+\frac{\epsilon}{\beta},s_y)))-c>0, \tag{10}$$

which implies that mutants obtain a higher payoff than incumbents, in contrast with (s_x, s_y) being evolutionarily stable.

A.4 Proof of Proposition 4

The validity of Proposition 4 can be shown along the lines of the proof of Proposition 3. In the following we limit ourselves to highlight the differences and provide brief comments.

In place of (2) and (3), we have:

$$\pi(C,y|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(D,y|(s_x-\epsilon_x,s_y-\epsilon_y))=p\left[b+(1-q)\left(-\frac{\beta-\epsilon_x}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_x}{\epsilon_x+\epsilon_y}d\right)\right]-c; \tag{11}$$

$$\pi(C,x|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(D,x|(s_x-\epsilon_x,s_y-\epsilon_y))=p\left[b+(1-q)\left(-\frac{1-\beta-\epsilon_y}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_y}{\epsilon_x+\epsilon_y}d\right)\right]-c; \tag{12}$$

from which we can derive the necessary and sufficient condition for evolutionarily stability, $p \geq \frac{c}{b-(1-q)\beta d}$, which shows the validity of the first bullet in the statement of the proposition.

In place of (4) and (5), we have:

$$\pi(D,y|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,y|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left[-b+(1-q)\left(-\frac{\beta-\epsilon_x}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_x}{\epsilon_x+\epsilon_y}d\right)\right]+c; \tag{13}$$

$$\pi(D,x|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,x|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left[-b+(1-q)\left(-\frac{1-\beta-\epsilon_y}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_y}{\epsilon_x+\epsilon_y}d\right)\right]+c; \tag{14}$$

from which we can derive the necessary and sufficient condition for evolutionarily stability, $p \leq \frac{c}{b+(1-q)\beta d}$, which shows the validity of the second bullet in the statement of the proposition.

In place of (6) and (7), we have:

$$\pi(D,y|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,y|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left[-b+(1-q)\left(+\frac{1-\beta-\epsilon_x}{1-\beta-\epsilon_x+\epsilon_y}d-\frac{\epsilon_x}{\beta+\epsilon_x-\epsilon_y}d\right)\right]+c; \tag{15}$$

$$\pi(C,x|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(D,x|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left[b+(1-q)\left(\frac{1-\beta-\epsilon_y}{1-\beta-\epsilon_y+\epsilon_x}d-\frac{\epsilon_y}{\beta-\epsilon_x+\epsilon_y}d\right)\right]-c; \tag{16}$$

from which we can derive the necessary and sufficient condition for evolutionarily stability, $\frac{c}{b-(1-q)d}>p>\frac{c}{b+(1-q)d}$. In place of (8) and (9), we can write analogous expressions, which however coincide with (15) and (16), since here we are considering the case of symmetric cultural aversion. Hence, we have shown the validity of the third bullet in the statement of the proposition.

Finally, the same argument used at the end of the proof of Proposition 3 can be used to show that no other state can ever be evolutionarily stable, with the only difference that: $\pi(C,x|(s_x,s_y))=\pi(D,x|(s_x,s_y))$ implies that $p[b-(1-q)d(\eta_{y|C}(s_x,s_y)-\eta_{y|D}(s_x,s_y))]-c=0$, and $\pi(C,x|(s_x+\frac{\epsilon}{\beta},s_y))-\pi(D,x|(s_x+\frac{\epsilon}{\beta},s_y))=p[(b-(1-q)d(\eta_{y|C}(s_x+\frac{\epsilon}{\beta},s_y)-\eta_{y|D}(s_x+\frac{\epsilon}{\beta},s_y))]-c\geq 0$.

A.5 Proof of Proposition 5

The validity of Proposition 5 can be shown along the lines of the proof of Proposition 3. In the following we limit ourselves to highlight the differences and provide brief comments.

In place of (2) and (3), we have:

$$\pi(C,y|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(D,y|s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))=p\left(b-\frac{\beta-\epsilon_x}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_x}{\epsilon_x+\epsilon_y}d-rf\right)+rf-c; \tag{17}$$

$$\pi(C,x|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))-\pi(D,x|(s_x-\frac{\epsilon_x}{\beta},s_y-\frac{\epsilon_y}{1-\beta}))=p\left(b-\frac{1-\beta-\epsilon_y}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_y}{\epsilon_x+\epsilon_y}d-rf\right)+rf-c; \tag{18}$$

from which we can derive the necessary and sufficient condition for evolutionarily stability: since $rf<c$, (17) and (18) can both be positive for all ϵ_x and ϵ_y when $\epsilon_x+\epsilon_y$ is close to zero only if $b-\beta d-rf>0$ and they are actually so if and only if, in addition, $p\geq\frac{c-rf}{b-\beta d-rf}$. This shows the validity of the first bullet in the statement of the proposition.

In place of (4) and (5), we have:

$$\pi(D,y|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,y|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(-b-\frac{\beta-\epsilon_x}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_x}{\epsilon_x+\epsilon_y}d+rf\right)-rf+c; \tag{19}$$

$$\pi(D,x|s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,x|(s_x+\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(-b-\frac{1-\beta-\epsilon_y}{1-\epsilon_x-\epsilon_y}d+\frac{\epsilon_y}{\epsilon_x+\epsilon_y}d+rf\right)-rf+c; \tag{20}$$

from which we can derive the necessary and sufficient condition for evolutionarily stability: since $rf<c$, (19) and (20) are both positive for all ϵ_x and ϵ_y when $\epsilon_x+\epsilon_y$ is close to zero if and only if $p\leq\frac{c-rf}{b+\beta d-rf}$. This shows the validity of the second bullet in the statement of the proposition.

In place of (6) and (7), we have:

$$\pi(D,y|(s_x-\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(C,y|(\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(-b+\frac{1-\beta-\epsilon_x}{1-\beta-\epsilon_x+\epsilon_y}d-\frac{\epsilon_x}{\beta+\epsilon_x-\epsilon_y}d+rf\right)-rf+c; \tag{21}$$

$$\pi(C,x|(\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))-\pi(D,x|(\frac{\epsilon_x}{\beta},s_y+\frac{\epsilon_y}{1-\beta}))=p\left(b+\frac{1-\beta-\epsilon_y}{1-\beta-\epsilon_y+\epsilon_x}d-\frac{\epsilon_y}{\beta-\epsilon_x+\epsilon_y}d-rf\right)+rf-c; \tag{22}$$

from which we can derive the necessary and sufficient condition for evolutionarily stability, as follows. Since $rf<c$, the quantity in (22) is positive for all ϵ_x and ϵ_y when $\epsilon_x+\epsilon_y$ is close to zero if and only if $p\geq\frac{c-rf}{b+d-rf}$. Moreover, because $rf<c$, (21) can be positive for all ϵ_x and ϵ_y when $\epsilon_x+\epsilon_y$ is close to zero only if $b-d-rf>0$, and it will actually be so if and only if, in addition, $p\leq\frac{c-rf}{b-d-rf}$. Altogether, these conditions show the validity of the third bullet in the statement of the proposition. In place of (8) and (9), we can write analogous expressions, which however coincide with (21) and (22), since here we are considering the case of symmetric cultural aversion. Hence, we have shown the validity of the third bullet in the statement of the proposition.

Finally, the same argument used at the end of the proof of Proposition 3 can be used to show that no other state can ever be evolutionarily stable, with the only difference that: $\pi(C,x|(s_x,s_y))=\pi(D,x|(s_x,s_y))$ implies that $p(b-d(\eta_{y|C}(s_x,s_y)-\eta_{y|D}(s_x,s_y))-rf)+rf-c=0$, and $\pi(C,x|(s_x+\frac{\epsilon}{\beta},s_y))-\pi(D,x|(s_x+\frac{\epsilon}{\beta},s_y))=p(b-d(\eta_{y|C}(s_x+\frac{\epsilon}{\beta},s_y)-\eta_{y|D}(s_x+\frac{\epsilon}{\beta},s_y))-rf)+rf-c\geq 0$.

A.6 Proof of Proposition 6

The validity of Proposition 6 can be shown by checking, for $\epsilon_x+\epsilon_y$ that tends to zero, the signs of expressions from (17) to (22) when $rf>c$, first for $b-d\geq c$ (first bullet of the proposition) and then for $b-d<c$ (all subsequent bullets), and then applying the arguments developed in Proposition 5.

A.7 Proof of Proposition 7

The validity of Proposition 7 can be shown by considering expressions from (2) to (7) re-written for the special case $d_x=d_y=d$ and substituting p with $a(s')$, where s' is the new state resulting from the considered invasion made of ϵ_x mutants of type x and ϵ_y mutants of type y. In particular, to establish the truth of the three bullets of Proposition 7, it is enough to follow the same arguments applied in the proof of Proposition 3 with the only difference that the claims about the signs of (2)-(7) have to be read as proving the sufficiency of the conditions involved (instead of both necessity and sufficiency).

Moreover, to establish that no other state can ever be evolutionarily stable, we can adjust the argument used in the last paragraph of the proof of Proposition 3, for which the following observation allows to apply the same reasoning. Let s be the original state supposed, ad absurdum, to be evolutionarily stable and let s' be the state resulting from an invasion of $\epsilon_x+\epsilon_y$ mutants. We observe that one can always consider the case where x-types are cooperators and y-types are defectors (or vice versa) where ϵ_x/ϵ_y is such that $a(s)=a(s')$, i.e., the fraction of cooperators in the whole population does not change. This allows to treat the index of assortativity in actions as fixed for the purpose of establishing the success of the invasion.

A.8 Proof of Proposition 8

By using equation 1, we can compute:

$$a(s_x = 0, s_y = 0) = 0 \tag{23}$$

$$a(s_x = 1, s_y = 0) = \frac{\beta(1-\beta)(n^2 - m^2)}{\beta(1-\beta)(n-m)^2 + nm} \tag{24}$$

$$a(s_x = 0, s_y = 1) = \frac{\beta(1-\beta)(n^2 - m^2)}{\beta(1-\beta)(n-m)^2 + nm} \tag{25}$$

$$a(s_x = 1, s_y = 1) = 0 \tag{26}$$

We can then verify whether the conditions of Proposition 7 are satisfied, thus obtaining the statement of the proposition.