Institutional Screening and the Sustainability of Conditional Cooperation

Ethan Holdahl[*]and Jiabin Wu[†]

October 4, 2023

Abstract

This paper studies a preference evolution model in which a population of agents are matched to play a sequential prisoner's dilemma in an incomplete information environment. An institution can design an incentive-compatible screening scheme, such as a special zone that requires an entry fee, or a costly label for purchase, to segregate the conditional cooperators from the non-cooperators. We show that institutional intervention of this sort can help the conditional cooperators to prevail when the psychological benefit of cooperating for them is sufficiently strong and the membership of the special zone or the label is inheritable with a sufficiently high probability.

Keywords: Conditional Cooperation, Incomplete Information, Screening, Evolutionary Game Theory

---
[*]Department of Economics, University of California, Irvine. Email: eholdahl@uci.edu.
[†]Department of Economics, University of Oregon. Email: jwu5@uoregon.edu.

1 Introduction

Conditional cooperation has interested researchers across various disciplines for decades [Ostrom, 2000]. Human-subject experiments [Fischbacher et al., 2001, Fischbacher and Gächter, 2010, and many others] have provided ample evidence for the existence of conditional cooperation. It refers to the phenomenon that people cooperate in dilemma situations only when they believe that others also cooperate. Conditional cooperation can be explained by people's other regarding preferences such as reciprocal preferences [Rabin, 1993, Dufwenberg and Kirchsteiger, 2004].

Initiated by Güth and Yaari [1992] and Güth [1995], the indirect evolutionary approach has become a workhorse model for studying the evolution of other-regarding social preferences [see Alger and Weibull, 2019, for a survey]. It assumes that evolution takes place at the preference level. Preferences determine behavior, which affects people's material payoffs. Material payoffs in turn determine the evolution of preferences. The indirect evolutionary approach thus offers a way to understand the evolutionary roots of various behavioral level phenomena, including conditional cooperation, from the preference level. In particular, Güth and Kliemt [1998] propose a model on the evolution of trustworthiness and show that its survival must be accompanied by conditional cooperation. When preferences are observable, the trustworthy type prevail in the long run because the trustworthy agents only cooperate when their opponents are trustworthy. In this case, the trustworthy agents act as conditional cooperators. However, when preferences are not observable, the share of trustworthy people is bound to decline because they can no longer behave differently accordingly to their opponents' types.

It is desirable for the conditional cooperators to differentiate themselves from the non-cooperators. This not only enables the conditional cooperators to behave differently when matching with different opponents, but also allows them to match only with other conditional cooperators. How can they achieve this when preferences are not observable? In this paper, we consider a potential solution: institutional screening. Institutional screening has been explored in the context of religion. For example, Iannaccone [1992] and Carvalho and Sacks [2021], among others, theoretically show how religious groups can use specific traits and rituals to screen out non-believers. The idea that institution can foster cooperation in dilemma situations can be traced back to Thomas Hobbes' Leviathan. We assume that a society consists of a population of people, including conditional cooperators and non-cooperators who are driven by different preferences, and an institution. Following Bisin and Verdier [2021], we conceptualize the institution as a mechanism through which social choices are delineated and implemented. This view on institutions is in line with [North, 1981, 1990, David, 1994, Grief, 2006]. If differentiating the conditional cooperators from

the non-cooperators serves the institutional goal, the institution may want to implement some screening mechanism to achieve it. For example, the institution can set up a special zone that requires entry fee or create an artificial label for people to purchase. The entry fee to the special zone or the cost of label should be designed in an incentive compatible way so that only the conditional cooperators are willing to pay for it. We show that such an entry fee/cost exists, which helps segregating the conditional cooperators from the non-cooperators. We show that whether the conditional cooperators can prevail in the long run depends on two critical features. First, the psychological benefit of cooperating for the conditional cooperators is sufficiently high, which guarantees that the institution always has an incentive to implement a screening scheme. Second, the membership of the special zone or the label is inheritable to a sufficiently high degree. That is to say that descendants of those with a membership face a lower membership cost than those whose parents were not members themselves. For example, assume the special zone is an island. Descendants of the special zone will be born on the island and wouldn't have to pay transportation costs to get to the special zone, thus reducing the effective membership cost. This discounted cost lowers the burden of future generations of conditional cooperators which is necessary to ensure their relative material payoffs remain competitive with the rest of the population.

This paper considers endogenous matching as opposed to the majority of the indirect evolutionary approach literature which considers exogenously random matching. A few papers consider matching being determined by institutions or democratic processes in the complete information environment [Rigos and Nax, 2016, Wu, 2016, 2022]. We instead consider the role of institutions in determining matching in the incomplete information environment. Since the institutions cannot directly manipulate the matching patterns in the incomplete information environment, we utilize the concept of stable matching to describe how agents are matched after the screening stage. Wu [2021], Hiller et al. [2022] and Wang and Wu [2022] are a few recent attempts to introduce stable matching to evolutionary models.

The paper is organized as follows. Section 2 lays out the setting of the model. Section 3 first re-illustrates the classic results without institutional intervention, then considers both a myopic and a forward-looking institution. Section 4 concludes.

2 The Model

2.1 Basic Setup

Consider a continuous population of agents of two preference types: θ and τ. The amount of θ-type agents is $x > 0$ and the amount of τ-type agent is $y > 0$. (x, y) denote the

population state. The agents are randomly matched in pairs to play a one-shot sequential prisoner's dilemma game as shown in Figure 1. In each pair, each agent takes the role of player 1 with probability 0.5 independent of their types.

Figure 1: A sequential prisoner's dilemma game with material payoffs
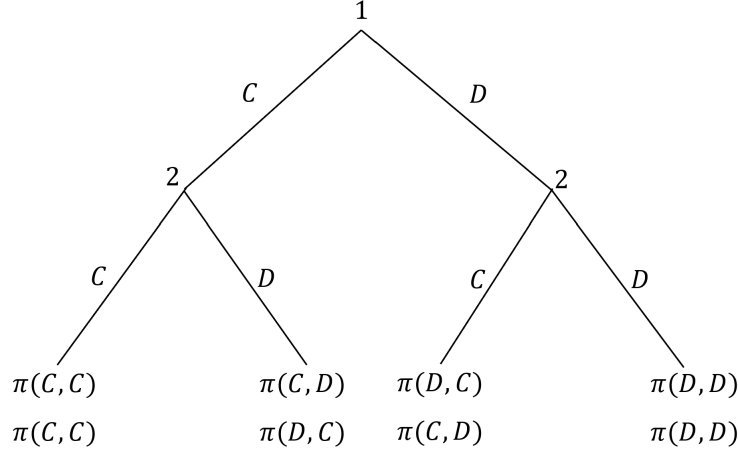
In the game, each agent has two actions: $\{C, D\}$. The game generates material payoffs for the agents. The material payoff function is given by $\pi(\cdot, \cdot) : \{C, D\}^2 \to \mathbb{R}$, which satisfies $\pi(D, C) > \pi(C, C) > \pi(D, D) > \pi(C, D)$, the standard criterion for a prisoner's dilemma. The τ-type agents' have materialistic preferences, that is, their preferences can be represented by the material payoff function. The τ-type agents always choose D as player 2, as it is their strictly dominant strategy. Hence, the τ-type agent can be regarded as non-cooperators. A θ-type agent's preferences are represented by the utility function $U_\theta(\cdot, \cdot) : \{C, D\}^2 \to \mathbb{R}$. The utility function satisfies $U_\theta(C, C) = \pi(C, C) + \alpha$, where $\alpha > \pi(D, C) - \pi(C, C)$; and $U_\theta(s, s') = \pi(s, s')$, for $(s, s') \neq (C, C)$. In words, a θ-type agent gains an extra psychological benefit of α only when choosing C against another agent choosing C. Given the condition imposed on α, a θ-type agent as player 2 would choose C upon observing that player 1 chooses C, and she would chooses D upon observing that player 1 chooses D. Hence, the θ-type agents can be regarded as conditional cooperators.

2.2 Evolution

We first add time index to the population state: (x_t, y_t) denotes the population state at time $t \in \mathbb{N}$ (assume that $x_0 > 0$ and $y_0 > 0$). Let $F_\theta(x_t, y_t)$ and $F_\tau(x_t, y_t)$ denote the average material payoffs of the θ-type agents and the τ-type agents at time t, respectively. We also define $\bar{U}_\theta(x_t, y_t)$ and $\bar{U}_\tau(x_t, y_t)$ as the average utility of the θ-type agents and the τ-type agents at time t, respectively.

Table 1: Subgame perfect equilibria under complete information

Player 1	Player 2	SPE
θ	θ	$(C, (C, D))$
θ	τ	$(D, (D, D))$
τ	θ	$(C, (C, D))$
τ	τ	$(D, (D, D))$

The population evolves across generations over time. The average material payoff of a type determines the reproductive rate of the agents of that type. Importantly, the evolution is governed by the comparison of material payoffs instead of utilities. When the average material payoff of the θ-type agents is higher than that of the τ-type agents in the current generation, the θ group grows faster than the τ group. The following model describes the evolutionary process:

$$x_{t+1} = x_t F_\theta(x_t, y_t), \tag{1}$$

$$y_{t+1} = y_t F_\tau(x_t, y_t). \tag{2}$$

Definition 1 *We say θ-type (respectively, τ-type) **dominates** the population as t approaches infinity. when we have $\lim_{t \to \infty} x_t/y_t = \infty$ (respectively, $\lim_{t \to \infty} y_t/x_t = \infty$.) We say the population reaches a **steady state** at time t if all agents in the population have the same reproduction rate from t on, this requires $F_\theta(x_{t+k}, y_{t+k}) = F_\tau(x_{t+k}, y_{t+k}) = c$, for some constant $c > 0$ and any integer $k \geq 0$.*

3   Analysis

3.1   Complete Information

In this section, we consider the benchmark case with complete information. That is, the agents' types are observable. In line with the literature on the indirect evolutionary approach, we first assume that agents in the population are uniformly randomly matched. Table 1 summarizes all the subgame perfect equilibria (SPEs). One can observe that the equilibrium predictions depend solely on player 2's type. When player 2 is a conditional cooperator (θ-type), both players would choose C in equilibrium; when player 2 is a non-cooperator (τ-type), both players would choose D in equilibrium.

The average material payoffs of the two types are given by:

$$F_\theta(x_t, y_t) = \frac{x_t}{x_t + y_t}\pi(C,C) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(D,D) + \frac{1}{2}\pi(C,C)), \tag{3}$$

$$F_\tau(x_t) = \frac{x_t}{x_t + y_t}(\frac{1}{2}\pi(C,C) + \frac{1}{2}\pi(D,D)) + \frac{y_t}{x_t + y_t}\pi(D,D). \tag{4}$$

Since $\pi(C,C) > \pi(D,D)$, we have $F_\theta(x_t, y_t) > F_\tau(x_t, y_t)$ for $x_t > 0$, implying that θ-type dominates the population as t approaches infinity for any initial condition (x_0, y_0). The rationale is that a conditional cooperator always cooperates with another conditional cooperator regardless of her role, but only cooperates with a non-cooperator when she is player 2. This makes sure that the conditional cooperators grab the benefit of cooperation among themselves and at the same time protect themselves from the non-cooperators' exploitation.

Since the types are observable, it is reasonable to assume that the agents can match voluntarily with those they prefer to match with. To capture this idea, we introduce the notion of stable matching.

A matching configuration is a vector $\mu(x_t, y_t) = (\mu_{a,b}(x_t, y_t))$, where $\mu_{a,b}(x_t, y_t)$ denotes the amount of type-a agents that are matched with type-b agents, with $a, b \in \{\theta, \tau\}$ that satisfies the following requirements.

$$\mu_{\theta,\theta}(x_t, y_t) + \mu_{\theta,\tau}(x_t, y_t) = x_t, \ \mu_{\tau,\theta}(x_t, y_t) + \mu_{\tau,\tau}(x_t, y_t) = y_t,$$
$$\text{and } \mu_{\theta,\tau}(x_t, y_t) = \mu_{\tau,\theta}(x_t, y_t). \tag{5}$$

We assume that the agents will reach a stable matching within each time period before they play the prisoner's dilemma. Since we are looking at a continuous population with finite types, we adopt the definition of stable matching for aggregate matching from Echenique et al. [2013] to our context, which serves as a natural generalization of the stability concept by Gale and Shapley [1962]. Let $K(a,b) : \{\theta,\tau\}^2 \to R$ denote the expected equilibrium utility of a type-a agent against a type-b agent.

Definition 2 *A pair of types (a,b), with $a, b \in \{\theta,\tau\}$, is a **blocking pair** for $\mu(x_t, y_t)$ if for $c \neq a$, $d \neq b$ and $c, d \in \{\theta,\tau\}$, we have $K(a,b) > K(a,d)$, $K(b,a) > K(b,c)$, $\mu_{a,d}(x_t, y_t) > 0$ and $\mu_{b,c}(x_t, y_t) > 0$. A matching configuration $\mu(x_t, y_t)$ is **stable** if there are no blocking pairs for $\mu(x_t, y_t)$.*

We will see that perfectly assortative matching (all θ-type agents are matched with one another, and all τ-type agents are matched with one another) is the unique stable matching as shown in Lemma 1:

Lemma 1 *The unique stable matching is $\mu_{\theta,\theta}(x_t, y_t) = x_t, \mu_{\tau,\tau}(x_t, y_t) = y_t$.*

Given that the agents are perfectly segregated through matching, the average material payoffs of the two types are given by $F_\theta(x_t) = \pi(C, C)$ and $F_\tau(x_t) = \pi(D, D)$. Since $\pi(C, C) > \pi(D, D)$, we have $F_\theta(x_t, y_t) > F_\tau(x_t, y_t)$ for $x_t > 0$, implying that θ-type dominates the population as t approaches infinity for any initial condition (x_0, y_0). Hence, through endogenous matching, conditional cooperators can completely shut themselves from matching with non-cooperators, ensuring their prevalence in the long run.

3.2 Incomplete Information

In this section, we consider the case with incomplete information. That is, the agents' types are not observable. In this case, since the agents cannot identify the types of others, they cannot discriminate between whom they match with. Hence, matching is uniformly random. We assume that the population state is common knowledge for the agents and the agents play a perfect Bayesian equilibrium (PBE) with one another after they are matched. First observe that in any PBE, a τ-type agent would always choose her strictly dominant strategy D as player 2, and a θ-type agent as player 2 would choose C upon observing that player 1 chooses C, and she would chooses D upon observing that player 1 chooses D. Lemma 2 summarizes the choices made by player 1 in PBEs. We do not consider the possibility that the agents play mixed strategies. Hence, to break ties, we assume that the agents choose C whenever they are indifferent between choosing C and D.

Lemma 2 *Under incomplete information,*

(1) *When $\frac{x_t}{x_t + y_t} \geq \Delta_\theta := \frac{\pi(D,D) - \pi(C,D)}{\pi(C,C) + \alpha - \pi(C,D)}$, a θ-type player 1 chooses C in PBE. Otherwise, she chooses D in PBE.*

(2) *When $\frac{x_t}{x_t + y_t} \geq \Delta_\tau := \frac{\pi(D,D) - \pi(C,D)}{\pi(C,C) - \pi(C,D)}$, a τ-type player 1 chooses C in PBE. Otherwise, she chooses D in PBE.*

Lemma 2 which characterizes player 1's decision is illustrated by Figure 2.

Given Lemma 2, we can characterize the long-run prediction of the evolutionary process of preferences:

Proposition 1 *Under incomplete information, a steady state population is characterized by $\frac{x_t}{x_t + y_t} \in (0, \Delta_\theta)$.*

The relationship between the average realized material payoffs by each type and the evolution of the population composition are depicted in graph (a) and (b) respectively in Figure 3.

Figure 2: Player 1's expected utility for each strategy
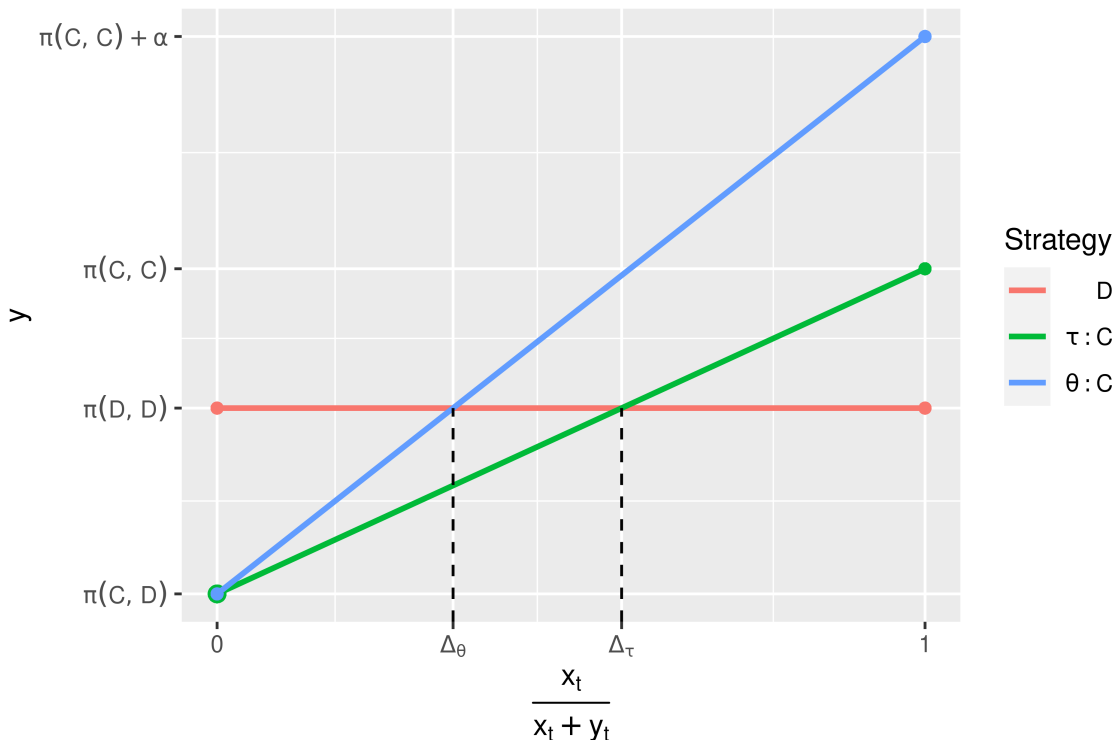
Note that it is always the case that $0 < \Delta_\theta < \Delta_\tau < 1$. Where Δ_X is the minimum proportion of conditional cooperators in the population required to make the expected utility of playing C as the first player at least as high as the utility received from playing D for player type X.

Proposition 1 shows that under incomplete information, the conditional cooperators can no longer dominate the entire society in the long run because they are not able to protect themselves from non-cooperators' exploitation. Nevertheless, they can still co-exist with the non-cooperators when their group size is sufficiently small so that they choose to defect as the non-cooperators. Note that the set $[0, \Delta_\theta)$ shrinks as α increases. In other words, a stronger psychological benefit of conditional cooperation makes the conditional cooperators more vulnerable in evolution. Note importantly, the results obtained in this section are merely the reminiscence of the results of Güth and Kliemt [1998]'s analysis on trust games.

3.3 Institutional Screening

In this section, we introduce an institution to the society. The institution is characterized by the **utilitarian** social welfare function:

$$W^{util}(x_t, y_t) = x_t \bar{U}_\theta(x_t, y_t) + y_t \bar{U}_\tau(x_t, y_t), \tag{6}$$

Figure 3: Evolution of Types under Incomplete Information
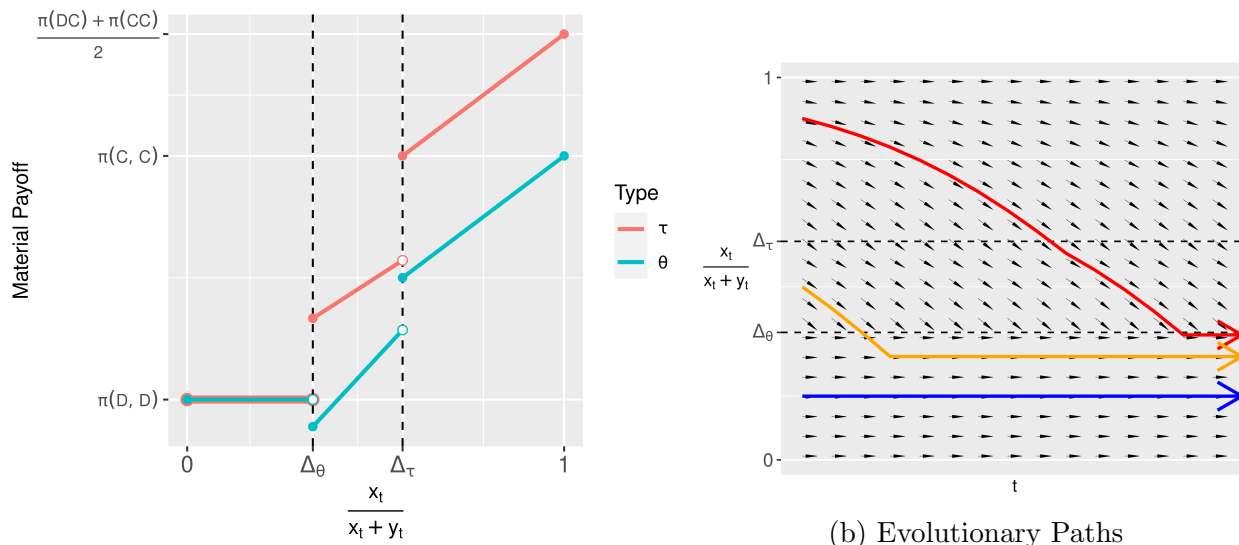
(a) Material Payoffs Realized for Each Type

(b) Evolutionary Paths

When $\frac{x_t}{x_t+y_t} \geq \Delta_\theta$ the average material payoffs (a) are always greater for τ types than θ types. This results in the proportion of θ types in the population ($\frac{x_t}{x_t+y_t}$) decreasing until the level is below Δ_θ. This dynamic can be seen in plot (b). When $\frac{x_t}{x_t+y_t} < \Delta_\theta$ both types get the same average material payoff and so a steady state is realized.

which is the total utility of the population in the current period. The institution aims to maximize the social welfare function by affecting the matching outcome. Since agents' types are unobservable, the institution cannot directly match agents according to their types. Hence, the institution may want to implement certain screening scheme to isolate different types of agents if differentiating them increases social welfare. Note that we assume the institution is short sighted as it only cares about the social welfare of the population in the current period. We believe that this assumption is reasonable because policy cycles often have a much shorter duration than a single generation.

To differentiate different types of agents, the institution can set up a special zone that requires an entry fee. The entry fee should be set in way that only the conditional cooperators would enter. Alternatively, the institution can create a costly label such that only the conditional cooperators are willing to pay for having the label. If the screening scheme works, it would be common knowledge that all agents in the special zone or those who have the labels are conditional cooperators. Assume that the entry fee or the cost of label is collected as tax and redistributed to everyone evenly, so that it would not directly affect the social welfare. Observe that the unique stable matching given that all conditional cooperators are in the special zone or having the costly label is characterized by perfectly assortative

9

matching since we are back to the complete information scenario.

Let the institution decision to implement a screening scheme be denoted by $z_t \in \{0, 1\}$ at time t. z_t can be regarded as the policy variable of the institution. Rewrite $W^{util}(x_t, y_t) = W^{util}(x_t, y_t, z_t)$ to reflect that it is a function of z_t. We first ask the question whether the institution wants to implement a screening mechanism to segregate the conditional cooperators from the defectors. Then we examine how to set the entry fee for the special zone or the cost for the label such that it is incentive compatible that only the conditional cooperators would enter the special zone or purchase the label.

Lemma 3 *When $\frac{x_t}{x_t+y_t} < \Delta_\tau$, $z_t^* = 1$. When $\frac{x_t}{x_t+y_t} \geq \Delta_\tau$, 1) if $\frac{x_t}{x_t+y_t} > \Delta_W := \frac{\pi(C,D)+\pi(D,C)-2\pi(D,D)}{\alpha}$, $z_t^* = 1$; 2) if $\frac{x_t}{x_t+y_t} < \Delta_W$, $z_t^* = 0$; 3) if $\frac{x_t}{x_t+y_t} = \Delta_W$, $z_t^* = 0$ or 1.*

The institution's screening decision is depicted in Figure 4.

Figure 4: Average Utility in the Population Under Varying Levels of Δ_W
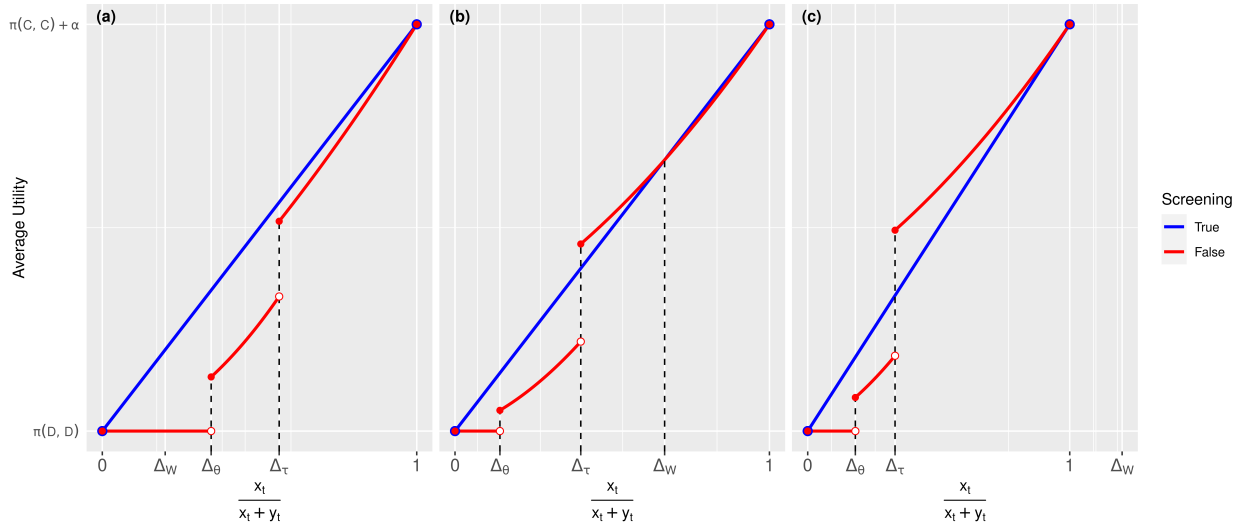
Presented are the three general cases for the position of Δ_W. In (a), $\Delta_W \leq \Delta_\tau$ so screening yields a higher average utility for all type distributions than not screening. In (b), $\Delta_\tau < \Delta_W < 1$. In this case screening yields a higher average utility than not screening only when $\frac{x_t}{x_t+y_t} \in (0, \Delta_\tau) \cup (\Delta_W, 1)$. In (c), the case where $\Delta_W \geq 1$, screening yields a higher average utility only when $\frac{x_t}{x_t+y_t} < \Delta_\tau$.

Let δ denote the entry fee for the special zone or the cost of label. Lemma 4 characterizes the range of δ that would make the institution's screening mechanism incentive compatible.

Lemma 4 *For any $\delta \in [\pi(D,C) - \pi(D,D), \pi(C,C) + \alpha - \pi(D,D)]$, $z_t^* = x_t$ can be implemented.*

Lemma 4 shows that the institution is able to differentiate the conditional cooperators from the non-cooperators according to its goal stated in Lemma 3.

10

Now, we can look at how institutional intervention would affect the evolution of preference types. We make two additional assumptions before we proceed with the analysis. First, since the end goal of the institution is to segregate the two types of agents, but not to maximize revenue (fees collected from the agents) we believe it is reasonable to assume that the institution would choose the minimum value of δ, $\frac{1}{2}\pi(C,C) + \frac{1}{2}\pi(D,C) - \pi(D,D)$, if it wants to implement a screening mechanism. Second, we allow the price of the membership to the special zone or label to be discounted by parameter $p \in [0,1]$ for descendants of those already in the special zone or label holders. Hence, the cost for descendants of the special zone or label is only $\delta * (1-p)$. When $p = 1$, there is no cost for descendants, they all inherit the label or membership to the special zone. This is analogous to most cases of citizenship where it may be costly for parents to travel to a country and gain citizenship but once there, their progeny gain citizenship for free. When $p = 0$, each generation has to pay the full cost of membership. $p \in (0,1)$ means that descendants of the special zone or label have to pay some cost to maintain their status but not the full amount. Perhaps there are two costs for membership: an initiation fee or transaction cost and a maintenance fee. Such a setup can be seen from country club memberships to real estate. The descendants inherit the membership from their parents and only have to pay the ongoing maintenance fees. Those born without membership have to also pay the initiation fee. In this scenario, the value of p is equal to the proportion of the initiation fee to the total cost of membership. Note that this is equivalent to a model where p descendants inherit membership to the special zone or label. In this case, to maintain the screening scheme at generational $t > 0$, the institution only needs to collect entry fee to the special zone or the cost of label from $x_t^*(1-p)$ amount of θ-type agents.

Proposition 2 *Given the institutional intervention, τ-type dominates the population as t approaches infinity for any initial state (x_0, y_0) if p is sufficiently small. On the other hand, θ-type dominates the population as t approaches infinity for any initial state (x_0, y_0) if p and α are sufficiently large.*

Recall that we assume that $\delta = \frac{1}{2}\pi(C,C) + \frac{1}{2}\pi(D,C) - \pi(D,D)$. Therefore, which type realizes higher material payoffs under screening depends on the relationship between p and $\Delta_p := \frac{\pi(D,C) - \pi(C,C)}{\pi(C,C) + \pi(D,C) - 2\pi(D,D)}$.

Illustrations of the average material payoffs for each type in the case where $\Delta_W \in (\Delta_\tau, 1)$ can be found in Figure 5. Figures of the average material payoffs for each type in the other cases can be found in the appendix. A representation of the resulting evolutionary paths in each case can be found in Figures 6, 7, and 8.

Proposition 2 shows that whether conditional cooperators can prevail through evolution in the long run depends on two critical factors: 1) a sufficiently large psychological benefit

Figure 5: Material Payoff by Type under Institutional Screening when $\Delta_W \in (\Delta_\tau, 1)$
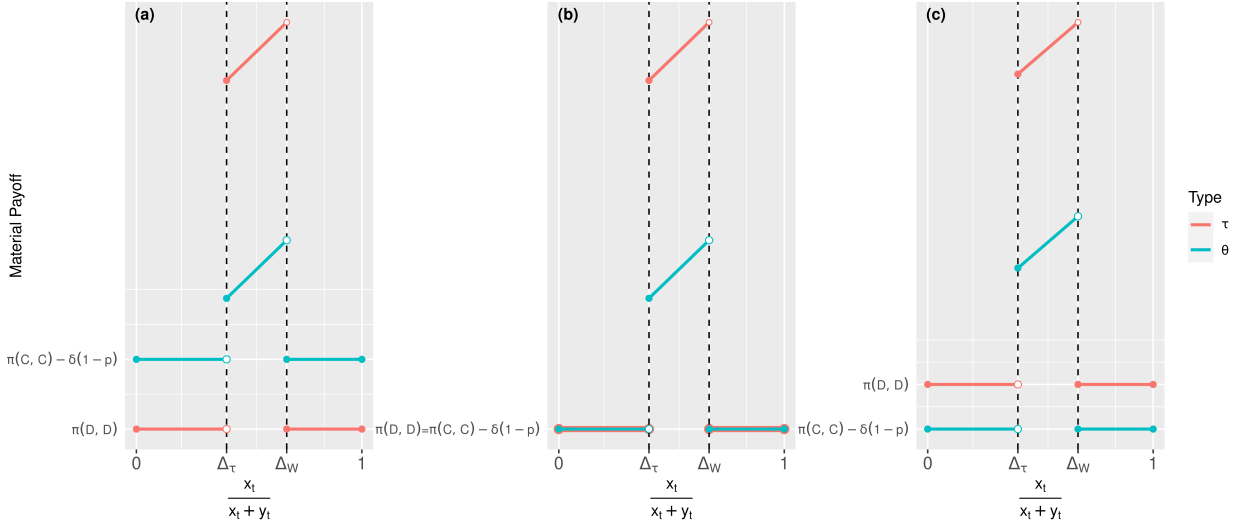

In all cases, between Δ_τ and Δ_W there is no screening and the τ types receive a higher payoff. In (a), $p > \Delta_p$ so the θ types receive a higher material payoff than τ types under screening. In (b), $p = \Delta_p$. In this case, material payoffs are the same for both types under screening. In (c), the case where $p < \Delta_p$, τ types receive higher material payoffs than θ types under screening.

Figure 6: Evolutionary Paths under Institutional Screening when $\Delta_W \in (\Delta_\tau, 1)$
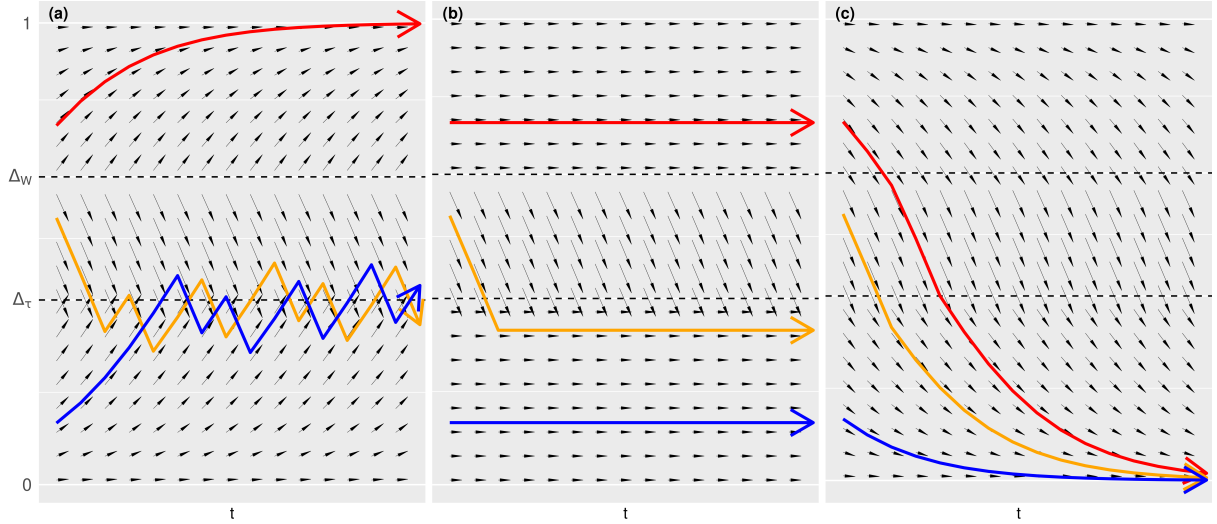
In (a), $p > \Delta_p$. If the proportion of θ types in the population is greater than Δ_W then θ types will evolve to dominate the population. Otherwise, the proportion of θ types in the population will oscillate around Δ_τ. In (b), since $p = \Delta_p$ any population proportion under screening is a steady state. If $\frac{x_t}{x_t + y_t} \in [\Delta_W, \Delta_\tau]$ then the institution will not institute screening until the proportion of θ types falls below Δ_τ. In (c), the case where $p < \Delta_p$, τ types receive higher material payoffs than θ types under both screening and no screening and will evolve to dominate the population.

Figure 7: Evolutionary Paths under Institutional Screening when $\Delta_W \leq \Delta_\tau$
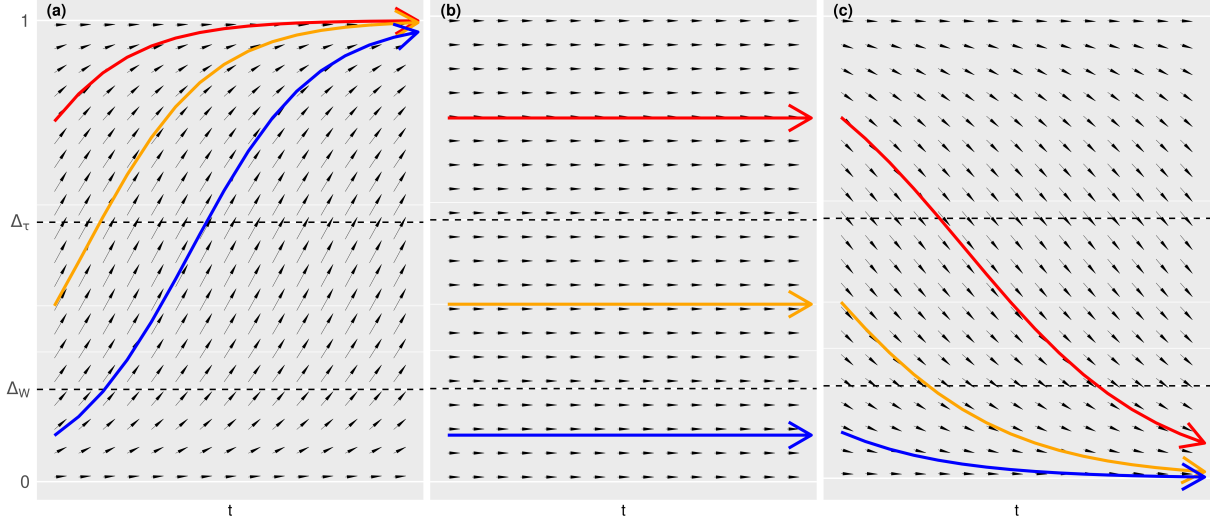
In (a), $p > \Delta_p$ and the θ types will evolve to dominate. In (b), since $p = \Delta_p$ any population proportion is a steady state. In (c), the case where $p < \Delta_p$, τ types will evolve to dominate the population.

Figure 8: Evolutionary Paths under Institutional Screening when $\Delta_W \geq 1$
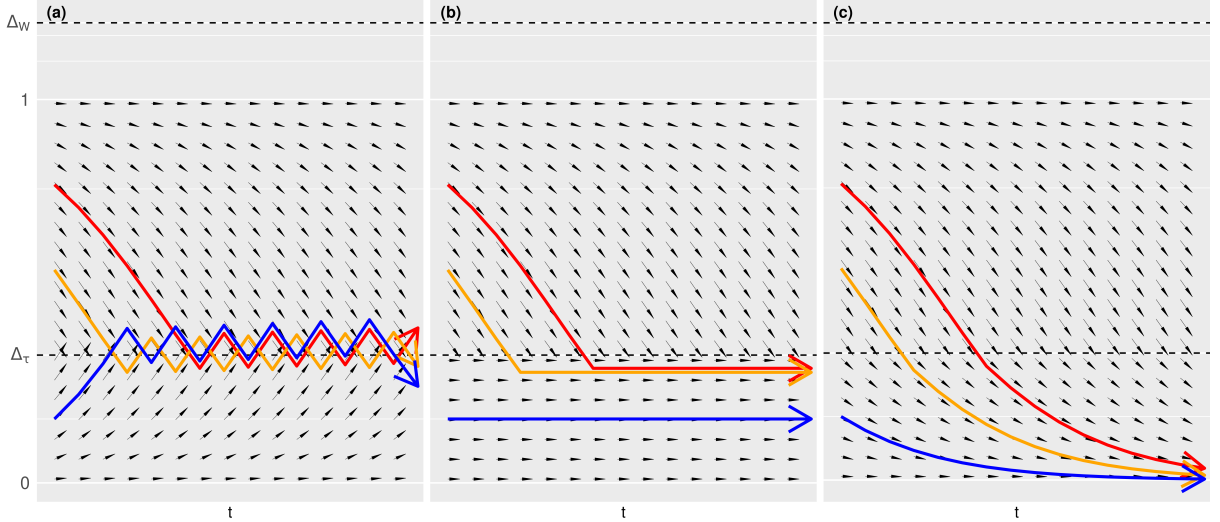
In (a), $p > \Delta_p$. There is no steady state and the proportion of θ types in the population will oscillate around Δ_τ. In (b), since $p = \Delta_p$ any population proportion under screening is a steady state. If $\frac{x_t}{x_t+y_t} \in [\Delta_\tau, 1]$ then the institution will not institute screening until the proportion of θ types falls below Δ_τ. In (c), the case where $p < \Delta_p$, τ types receive higher material payoffs than θ types under both screening and no screening and will evolve to dominate the population.

of conditional cooperation α to make sure that the institution has a strong incentive to differentiate the conditional cooperators from the defectors, 2) a sufficient high probability of inheritance, which lowers the burden of future generations of conditional cooperators on

paying the entry fee for the special zone or the label cost.

4 Conclusion

We study the evolution of conditional cooperation under incomplete information by using the indirect evolutionary approach. We introduce an institution to the society, who can implement a screening scheme to differentiate the conditional cooperators from the non-cooperators. A screening scheme may benefit the conditional cooperators because it allows the conditional cooperators to match with and cooperate with other conditional cooperators. It is also desirable from the societal perspective in the short run as it can increase social welfare. In the long run, whether such a screening scheme can help to sustain cooperation through populating the society with conditional cooperators depends on several factors including the psychological benefit of conditional cooperation for the conditional cooperators and the probability of inheritance of the special zone membership or the label.

In this paper, we consider institutional screening as a way to differentiate different agents in the population, which can be viewed as a centralized approach. On the other hand, costly signaling by Zahavi [1975] and Spence [1973] can serve as an alternative way from a decentralized perspective.[1] How signaling affects the evolution of conditional cooperation in the framework of the indirect evolutionary approach would be an interesting research question.

References

I. Alger and J. W. Weibull. Evolutionary models of preference formation. *Annual Review of Economics*, 11:329–354, August 2019.

A. Bisin and T. Verdier. On the joint evolution of culture and institutions: Elites and civil society. 2021.

J.-P. Carvalho and M. Sacks. The economics of religious communities. *Journal of Public Economics*, 201:104481, 2021.

---

[1]See Grafen [1990], Maynard Smith and Harper [1995], Johnstone [1997], Zahavi and Zahavi [1997], Gintis et al. [2001] Maynard Smith and Harper [2003], Searcy and Nowicki [2005], Getty [2006], Grose [2011] and Számadó [2012], Hopkins [2014], Przepiorka and Diekmann [2021], Holdahl and Wu [2023], among many others, for works that apply the idea to explaining phenomena ranging from life sciences to social sciences and further discussions on the subject.

P. A. David. Whyare institutions the 'carriers of history'?: Path dependence and the evolution of conventions, organizations and institutions. *Structural Change and Economic Dynamics*, 5:205–220, 1994.

M. Dufwenberg and G. Kirchsteiger. A theory of sequential reciprocity. *Games and Economic Behavior*, 47:268–298, 2004.

F. Echenique, S. Lee, M. Shum, and B. M. Yenmez. The revealed preference theory of stable and extremal stable mathcings. *Econometrica*, 81(1):153–171, 2013.

U. Fischbacher and S. Gächter. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review*, 100(1):541–556, 2010.

U. Fischbacher, S. Gächter, and E. Fehr. Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404, 2001.

D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, January 1962.

T. Getty. Sexually selected signals are not similar to sports handicaps. *Trends in ecology and evolution*, 21:83–88, 2006.

H. Gintis, E. Alden Smith, and S. Bowles. Costly signaling and cooperation. *Journal of Theoretical Biology*, 213:103–119, 2001.

A. Grafen. Biological signals as handicaps. *Journal of Theoretical Biology*, 144:517–546, 1990.

A. Grief. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade.* Cambridge University Press, New York, 2006.

J. Grose. Modelling and the fall and rise of the handicap principle. *Biology & Philosophy*, 26:677–696, 2011.

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

W. Güth and H. Kliemt. Indirect evolutionary approach: Bridging the gap between rationality and adaptation. *Rationality and Society*, 10(3):377–399, 1998.

W. Güth and M. Yaari. An evolutionary approach to explaining reciprocal behavior in a simple strategic game. In U. Witt, editor, *Approaches to Evolutionary Economics.* University of Michigan Press, 1992.

V. Hiller, J. Wu, and H. Zhang. Heterophily, stable matching, and intergenerational transmission in cultural evolution. 2022.

E. Holdahl and J. Wu. Conflicts, assortative matching, and the evolution of signaling norms. *Journal of Economic Interaction and Coordination*, pages 936–941, 2023.

E. Hopkins. Competitive altruism, mentalizing and signaling. *American Economic Journal: Microeconomics*, 2656(4):272–292, 2014.

L. R. Iannaccone. Sacrifice and stigma: Reducing free-riding in cults, communes, and other collectives. *Journal of Political Economy*, 100(2):271–291, 1992.

R. Johnstone. The evolution of animal signals. In J. Krebs and N. Davies, editors, *Behavioral ecology, an evolutionary approach*, pages 465–485. Blackwell Scientific Publications, Oxford, 4 edition, 1997.

J. Maynard Smith and D. Harper. Animal signals: models and terminology. *Journal of Theoretical Biology*, 177:305–311, 1995.

J. Maynard Smith and D. Harper. *Animal Signals*. Oxford University Press, Oxford, 2003.

D. North. *Structure and Change in Economic History*. Norton, New York, 1981.

D. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, New York, 1990.

E. Ostrom. Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3):137–158, 2000.

W. Przepiorka and A. Diekmann. Parochial cooperation and the emergence of signalling norms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376: 20200294, 2021.

M. Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83:1281–1302, 1993.

A. Rigos and H. Nax. Assortativity evolving from social dilemmas. *Journal of Theoretical Biology*, 395:194–203, 2016.

W. Searcy and S. Nowicki. *The evolution of animal communication. Reliability and deception in signalling systems*. Princeton University Press, Princeton, 2005.

A. Spence. Job market signaling. *Quarterly Journal of Economics*, 87:355–374, 1973.

S. Számadó. The rise and fall of the handicap principle: a commentary on the "modelling and the fall and rise of the handicap principle. *Biology & Philosophy*, 27:279–286, 2012.

Z. Wang and J. Wu. Preference evolution under stable matching. 2022.

J. Wu. Evolving assortativity and social conventions. *Economics Bulletin*, 36:936–941, 2016.

J. Wu. Matching markets and cultural selection. *Review of Economic Design*, 25:267–288, 2021.

J. Wu. Institutions, assortative matching, and cultural evolution. 2022.

A. Zahavi. Mate selection: a selection for a handicap. *Journal of Theoretical Biology*, 53:205–214, 1975.

A. Zahavi and A. Zahavi. *The handicap principle*. Oxford University Press, Oxford, 1997.

A   Additional Figures

Figure 9: Material Payoff by Type under Institutional Screening when $\Delta_W \leq \Delta_\tau$
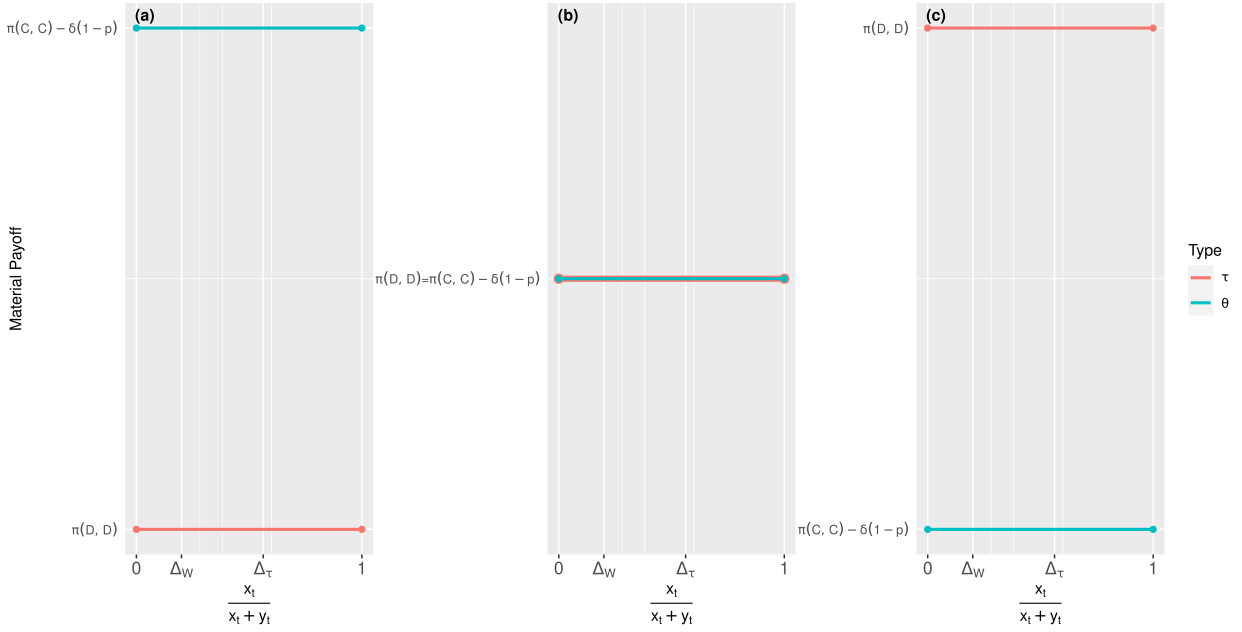
In this case there is always screening implemented. In (a), $p > \Delta_p$ so the θ types receive a higher material payoff than τ types under screening. In (b), $p = \Delta_p$. In this case, material payoffs are the same for both types under screening. In (c), the case where $p < \Delta_p$, τ types receive higher material payoffs than θ types under screening.

Figure 10: Material Payoff by Type under Institutional Screening when $\Delta_W \geq 1$
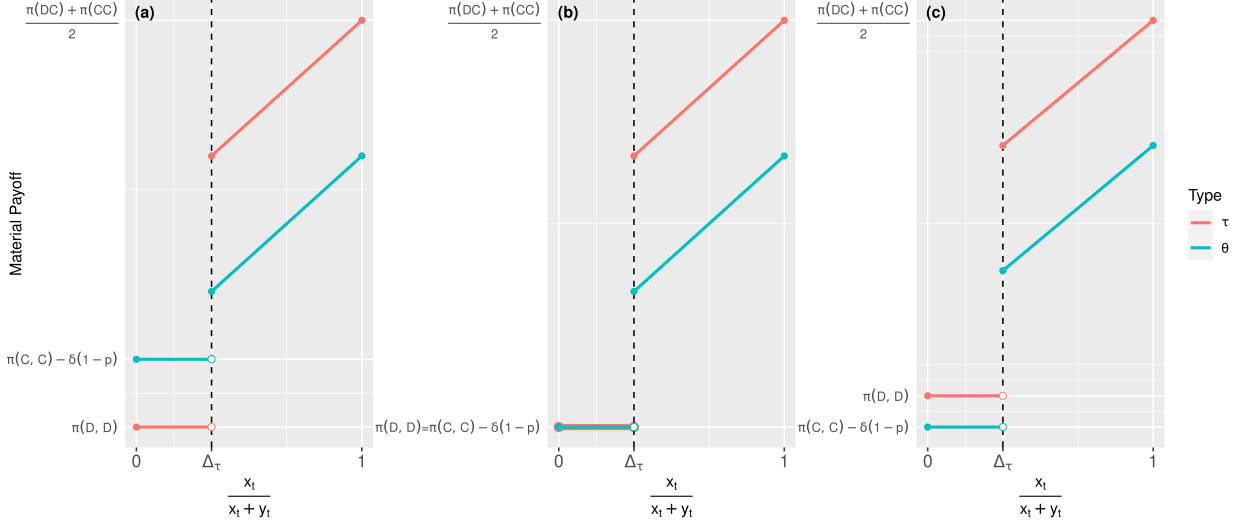
In this case screening implemented only when $\frac{x_t}{x_t+y_t} < \Delta_\tau$. Under no screening the τ types always recieve a higher material payoff than the θ types. In (a), $p > \Delta_p$ so the θ types receive a higher material payoff than τ types under screening. In (b), $p = \Delta_p$. In this case, material payoffs are the same for both types under screening. In (c), the case where $p < \Delta_p$, τ types receive higher material payoffs than θ types under screening.

B   Proofs

Proof of Lemma 1: First, we prove that it is a stable matching. The only potential blocking pair is (θ, τ). However, since $K(\theta, \tau) = \frac{1}{2}\pi(D, D) + \frac{1}{2}(\pi(C, C) + \alpha) < K(\theta, \theta) = \pi(C, C) + \alpha$, it is not a blocking pair. Second, we prove that it is unique. Suppose there exists another stable matching such that $\mu_{\theta,\tau}(x_t, y_t) > 0$, then (θ, θ) is a blocking pair because $K(\theta, \tau) < K(\theta, \theta)$, which is a contradiction. *Q.E.D.*

Proof of Lemma 2: Given what player 2 would choose in the second stage of the game, a θ-type player 1's expected utility of choosing C is $\frac{x_t}{x_t+y_t}(\pi(C, C) + \alpha) + \frac{y_t}{x_t+y_t}\pi(C, D)$; and her expected utility of choosing D is $\pi(D, D)$. Hence, she chooses C if and only if $\frac{x_t}{x_t+y_t} \geq \Delta_\theta = \frac{\pi(D,D)-\pi(C,D)}{\pi(C,C)+\alpha-\pi(C,D)}$. A τ-type player 1's expected utility of choosing C is $\frac{x_t}{x_t+y_t}\pi(C, C) + \frac{y_t}{x_t+y_t}\pi(C, D)$; and her expected utility of choosing D is $\pi(D, D)$. Hence, she chooses C if and only if $\frac{x_t}{x_t+y_t} \geq \Delta_\tau = \frac{\pi(D,D)-\pi(C,D)}{\pi(C,C)-\pi(C,D)}$. *Q.E.D*

Proof of Proposition 1: When $\frac{x_t}{x_t+y_t} \geq \Delta_\tau$, the average material payoffs of the two types

are given by

$$F_\theta(x_t, y_t) = \frac{x_t}{x_t + y_t}\pi(C, C) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(C, D) + \frac{1}{2}\pi(C, C)), \tag{7}$$

$$F_\tau(x_t, y_t) = \frac{x_t}{x_t + y_t}(\frac{1}{2}\pi(C, C) + \frac{1}{2}\pi(D, C)) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(C, D) + \frac{1}{2}\pi(D, C)). \tag{8}$$

$F_\theta(x_t, y_t) - F_\tau(x_t, y_t) = -\frac{1}{2}(\pi(D, C) - \pi(C, C)) < 0$, which tells us that $\frac{x_t}{x_t + y_t}$ is decreasing in t.

When $\Delta_\theta \leq \frac{x_t}{x_t + y_t} < \Delta_\tau$, the average material payoffs of the two types are given by

$$F_\theta(x_t, y_t) = \frac{x_t}{x_t + y_t}\pi(C, C) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(C, D) + \frac{1}{2}\pi(D, D)), \tag{9}$$

$$F_\tau(x_t, y_t) = \frac{x_t}{x_t + y_t}(\frac{1}{2}\pi(D, D) + \frac{1}{2}\pi(D, C)) + \frac{y_t}{x_t + y_t}\pi(D, D). \tag{10}$$

$F_\theta(x_t, y_t) > F_\tau(x_t, y_t)$ is equivalent to $\frac{x_t}{x_t + y_t} > \frac{\pi(D,D)-\pi(C,D)}{2\pi(C,C)-\pi(C,D)-\pi(D,C)}$. However, $\frac{\pi(D,D)-\pi(C,D)}{2\pi(C,C)-\pi(C,D)-\pi(D,C)} > \Delta_\tau$. Hence, $F_\theta(x_t, y_t) < F_\tau(x_t, y_t)$ when $\frac{x_t}{x_t + y_t} \in [\Delta_\theta, \Delta_\tau)$, which tells us that $\frac{x_t}{x_t + y_t}$ is decreasing in t.

When $\frac{x_t}{x_t + y_t} < \Delta_\theta$, we must have $F_\theta(x_t, y_t) = F_\tau(x_t, y_t) = \pi(D, D)$. Hence, $\frac{x_t}{x_t + y_t}$ remains a constant and the population reaches a steady state. *Q.E.D.*

Proof of Lemma 3: If the institution implements a screening scheme, we must have $W^{util}(x_t, y_t, 1) = x_t(\pi(C, C) + \alpha) + y_t\pi(D, D)$.

If the institution does not implement a screening mechanism, then we calculate the social welfare according to Lemma 2. First, suppose $\frac{x_t}{x_t + y_t} < \Delta_\theta$. In this case, we have $W^{util}(x_t, y_t, 0) = (x_t + y_t)\pi(D, D) < W^{util}(x_t, y_t, 1)$. Hence, $z_t^* = 1$.

Second, suppose $\frac{x_t}{x_t + y_t} \in [\Delta_\theta, \Delta_\tau)$. In this case, we have

$$\begin{aligned}
W^{util}(x_t, y_t, 0) &= x_t\Big(\frac{x_t}{x_t + y_t}(\pi(C, C) + \alpha) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(C, D) + \frac{1}{2}\pi(D, D)) \\
&\quad + y_t\Big(\frac{x_t}{x_t + y_t}(\frac{1}{2}\pi(D, D) + \frac{1}{2}\pi(D, C)) + \frac{y_t}{x_t + y_t}\pi(D, D)\Big).
\end{aligned} \tag{11}$$

Simple calculation yields:

$$W^{util}(x_t, y, 1) - W^{util}(x_t, y_t, 0) = \frac{x_t y_t}{x_t + y_t}\Big((\pi(C, C) + \alpha) - \frac{1}{2}\pi(C, D) - \frac{1}{2}\pi(D, C)\Big) > 0. \tag{12}$$

Hence, $z_t^* = 1$.

Suppose $\frac{x_t}{x_t+y_t} \geq \Delta_\tau$, In this case, we have

$$W^{util}(x_t, y_t, 0) = x_t\Big(\frac{x_t}{x_t + y_t}(\pi(C,C) + \alpha) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(C,D) + \frac{1}{2}(\pi(C,C) + \alpha)))$$
$$+y_t\Big(\frac{x_t}{x_t + y_t}(\frac{1}{2}\pi(C,C) + \frac{1}{2}\pi(D,C)) + \frac{y_t}{x_t + y_t}(\frac{1}{2}\pi(C,D) + \frac{1}{2}\pi(D,C)))). \qquad (13)$$

Simple calculation yields:

$$W^{util}(x_t, y, 1) - W^{util}(x_t, y_t, 0) = y_t\Big(\frac{x_t}{x_t + y_t}\frac{1}{2}\alpha + \pi(D,D) - \frac{1}{2}(\pi(C,D) + \pi(D,C)). \quad (14)$$

which can be positive or negative depending on the comparison between $\frac{x_t}{x_t+y_t}$ and $\frac{\pi(C,D)+\pi(D,C)-2\pi(D,D)}{\alpha} = \Delta_W$. In sum, when $\frac{x_t}{x_t+y_t} < \Delta_\tau$, the institution always wants to implement a screening scheme. When $\frac{x_t}{x_t+y_t} \geq \Delta_\tau$, whether the institution prefers a screening mechanism depends on the value of Δ_W. *Q.E.D.*

Proof of Lemma 4: From an ex-ante perspective, given that the agents expect that all conditional cooperators are in the special zone or with the label, while all non-cooperators are outside the special zone or without the label and the unique stable matching is perfectly assortative, we need to ensure that the conditional cooperators have the incentive to pay the entry fee or the cost of label, while the non-cooperators do not. Hence, two conditions need to be satisfied at the same time:

$$\pi(C,C) + \alpha - \delta \geq \pi(D,D); \qquad (15)$$
$$\frac{1}{2}\pi(C,C) + \frac{1}{2}\pi(D,C) - \delta \leq \pi(D,D). \qquad (16)$$

They imply that $\delta \in [\frac{1}{2}\pi(C,C) + \frac{1}{2}\pi(D,C) - \pi(D,D), \pi(C,C) + \alpha - \pi(D,D)]$. Since $\pi(C,C) + \alpha > \pi(D,C)$ by assumption, the set is non-empty. *Q.E.D.*

Proof of Proposition 2:
i) Suppose $\alpha \geq \frac{1}{\Delta_\tau}(\pi(C,D) + \pi(D,C) - 2\pi(D,D))$. Thus, $\Delta_W \leq \Delta_\tau$ which means $z_t^* = 1$ for any (x_t, y_t). Hence,

$$F_\theta(x_t, y_t) - F_\tau(x_t, y_t) = \pi(C,C) - \delta(1 - p) - \pi(D,D). \qquad (17)$$

When $p > \Delta_p$, we must have $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) > 0$ for any (x_t, y_t), implying that θ-type dominates the population as t approaches infinity for any initial condition (x_0, y_0). When $p < \Delta_p$, $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) < 0$ for any (x_t, y_t), implying that τ-type dominates

the population as t approaches infinity for any initial condition (x_0, y_0). When $p = \Delta_p$, $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) = 0$ for any (x_t, y_t), implying that any state is a steady state.

ii) Suppose $\alpha \leq \pi(C, D) + \pi(D, C) - 2\pi(D, D)$. Then $\Delta_W \geq 1$ and $z_t^* = 1$ for any $\frac{x_t}{x_t + y_t} < \Delta_\tau$, and $z_t^* = 0$ for any $\frac{x_t}{x_t + y_t} \geq \Delta_\tau$. This implies that $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) < 0$ for any $\frac{x_t}{x_t + y_t} \geq \Delta_\tau$. When $p > \Delta_p$, we must have $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) > 0$ for any $\frac{x_t}{x_t + y_t} < \Delta_\tau$. Hence, there is no steady state in this case. When $p < \Delta_p$, $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) < 0$ for any (x_t, y_t), implying that τ-type dominates the population as t approaches infinity for any initial condition (x_0, y_0). When $p = \Delta_p$, $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) = 0$ for any $\frac{x_t}{x_t + y_t} \in (0, \Delta_\tau)$, implying that any state such that $\frac{x_t}{x_t + y_t} \in (0, \Delta_\tau)$ is a steady state.

iii) Suppose $\alpha \in (\pi(C, D) + \pi(D, C) - 2\pi(D, D), \frac{1}{\Delta_\tau}(\pi(C, D) + \pi(D, C) - 2\pi(D, D)))$. Thus, $\Delta_W \in (\Delta_\tau, 1)$ such that for any $\frac{x_t}{x_t + y_t} \in (\Delta_W, 1)$, $z_t^* = 1$, for any $\frac{x_t}{x_t + y_t} \in (\Delta_\tau, \Delta_W)$, $z_t^* = 0$, and for $\frac{x_t}{x_t + y_t} = \Delta_W$, $z_t^* = 0$ or 1. Also, we still have $z_t^* = 1$ for any $\frac{x_t}{x_t + y_t} < \Delta_\tau$. When $p > \Delta_p$, we must have $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) > 0$ for any $\frac{x_t}{x_t + y_t} \in (0, \Delta_\tau) \cup (\Delta_W, 1)$, but $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) < 0$ for any $\frac{x_t}{x_t + y_t} \in (\Delta_\tau, \Delta_W)$ and either $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) < 0$ or $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) > 0$ when $\frac{x_t}{x_t + y_t} = \Delta_W$. Hence, θ-type dominates the population as t approaches infinity for any initial condition (x_0, y_0) such that $\frac{x_0}{x_0 + y_0} \in (\Delta_W, 1)$. When $p < \Delta_p$, $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) < 0$ for any (x_t, y_t), implying that τ-type dominates the population as t approaches infinity for any initial condition (x_0, y_0). When $p = \Delta_p$, $F_\theta(x_t, y_t) - F_\tau(x_t, y_t) = 0$ for any $\frac{x_t}{x_t + y_t} \in (0, \Delta_\tau) \cup (\Delta_W, 1)$, implying that any state such that $\frac{x_t}{x_t + y_t} \in (0, \Delta_\tau) \cup (\Delta_W, 1)$ is a steady state. If $z_t = 1$ when $\frac{x_t}{x_t + y_t} = \Delta_W$ for all t then $\frac{x_t}{x_t + y_t} = \Delta_W$ is also a steady state when $p = \Delta_p$.

Q.E.D.