Long-Term Relationships as Safeguards

Rafael Rob[*] and Huanxing Yang[†]

December 15, 2006

Abstract

We analyze a repeated prisoners' dilemma game played in a community setting with heterogeneous types. The setting is such that individuals choose whether to continue interacting with their present partner, or separate and seek a new partner. Players' types are not directly observed, but may be imperfectly inferred from observed behavior. We focus on a class of equilibria that satisfy zero tolerance (an individual separates immediately if her partner defects), and fresh start (behavior in a new relationship does not depend on experience in previous relationships). We find that the punishment for defecting and the reward for cooperating are driven by the formation and the dissolution of long-term, high-paying relationships: An individual that defects, aborts a long-term relationships that he is in, or that he might have entered into, is thrown into short-term interactions with individuals who are likely to defect and, consequently, receives low payoffs. On the flip side, an individual that cooperates, enters into or prolongs a long-term interaction with a partner who cooperates and, consequently, receives high payoffs.

JEL Classification numbers: C73, C78, D82.

Keywords: community games, information flows, heterogeneity of types, long term relationships, endogenous determination of types.

[*] Corresponding author. Phone number: 215-898-6775. Fax number: 215-573-2057. Email: rrob@ssc.upenn.edu. Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104. Acknowledges NSF support under grant number 01-36922.

[†] Department of Economics, The Ohio State University, 1945 N. High Street, Columbus, OH 43210. Email: yang.1041@osu.edu.

1 Introduction

Overview and Results. In this paper we study a prisoners' dilemma game in which individuals interact with varying opponents over time. The environment is such that individuals have partial control over who to interact with. Namely, individuals choose - based on their past experience - whether to prolong the interaction with their present partner for another period, or seek a new partner. In addition, the population of individuals is heterogeneous in that some individuals are strategic types, i.e., they choose actions based on incentives, while others are behavioral types, i.e., are programmed to take a fixed action. Our aim is to study the structure of incentives for a certain class of equilibria in this environment and, in particular, to determine the impact of the "demographics" of the population, i.e., the effect that the type-distribution has on equilibrium behavior.

The equilibria we identify have the feature that behavior and, hence, payoffs depend on whether a player is in one of two states. One state is that a player is in an ongoing relationship, i.e., she has interacted with her current partner at least once. The other state is that a player is in a new relationship. Since players are able to condition their behavior on state, being in an ongoing relationship might deliver higher payoffs than being in a new relationship because players cooperate in the former, but defect in the latter. This possibility dictates the structure of equilibrium incentives. In particular, a player in an ongoing relationship, who is "scheduled" to cooperate, has an incentive to do so because, otherwise, the relationship he is in will be terminated and he will be forced to interact with players who are likely to defect and, thereby, inflict low payoffs on him. In addition, a player who is in a new relationship might, nonetheless, cooperate because if he is matched to another player who cooperates, he will enter into a long-term, high-paying relationship. Therefore, cooperation in this setting is a form of investment aimed at creating or maintaining a high-paying status. Our goal is to explore this logic and, more specifically, to pin down the conditions under which this force is sufficient to guarantee that the equilibrium in which strategic types cooperate - called the good equilibrium - exists and, analogously, to pin down the conditions under which other (pure and mixed strategy) equilibria exist.

We report two sets of results. In the first set, and as suggested above, we determine when pure and mixed strategy equilibria exist as a function of model parameters. One of our main findings is that the fraction of bad types (behavioral types that chronically defect) has to be in an intermediate range to sustain the good equilibrium. The reason for this is that if a strategic player causes a long-term relationship to terminate by defecting and if the fraction of bad types is sufficiently small, the defector is likely to meet another strategic type fairly quickly and, hence, bounce back to another long-term, high-paying relationship, so he is not made to pay for the infraction. On the other hand, if the fraction of bad types is sufficiently large, the time it takes to form a long-term relationship is excessively long, and the cost incurred in the process is excessively high, so strategic types do not try to form such relationship, i.e., they simply defect. A similar principle applies to bad equilibria, which are shown *not* to exist, if the fraction of good types is in some intermediate range. These results indicate that the type distribution has indeed an impact on the equilibrium behavior. Another result we report is that the problem that a good equilibrium does not exist because the fraction of bad types is too low can be rectified if we allow mixed strategies. Namely, additional bad types can be "created" endogenously if some of the strategic types defect (while other strategic types cooperate in accordance with the good equilibrium). In the second set of results we extend the model so that the distribution of types is endogenized based on equilibrium payoff-differentials. Specifically, we show that since strategic types collect higher equilibrium payoffs than bad types, the latter may be willing to incur an expense to have a larger set of actions available to them, "converting" them thereby to strategic types. In this way (or similarly because of evolutionary pressures), the type distribution is endogenously determined. In this extension of the model we also execute welfare analysis to identify externalities that cause discrepancies between the equilibrium and the optimum number of strategic types. As it turn out, there are both positive and negative externalities, so there may be either too many or too few strategic types.

Although this paper is intended as a theoretical exploration, anecdotal evidence suggests that the forces we identify here are of some relevance in the real world. One anecdote suggesting this comes from the banking industry and, in particular, the practice of "customer relationships." Roughly speaking, this practice is such that an established customer, who pays back his loans on time, gets to sustain his long-term relationship with the bank and, thereby, borrow at a lower interest rate, or borrow larger amounts. On the other hand, a new customer may have to pay a higher interest rate or borrow a smaller amount, and a customer who is not current on his loans, is denied credit altogether and may have to turn to other institutions for future business, and pay a higher interest rate. In this way, payoffs to borrowers depend on their status, and sanctions are invoked by changing one's status. A similar arrangement applies to depositors, who receive extra services (e.g., investment advisory) or higher interest rates on their deposits, if they maintain a minimum balance or keep their deposits for a sufficiently long time. Looking at the industry not from the angle of an individual customer, but from the perspective of the market as a whole, our results also offer a reason for capital-market imperfections. Indeed, if corruption is rampant in the society, banks may be leery of new borrowers and, in the extreme, refuse to extend credit to them, or insist on very high interest rates. Such state of affairs is consistent with our result that the good equilibrium is unsustainable when the fraction of bad types is sufficiently high. Other institutions that feature a similar structure of incentives are seniority in employment relationships, or securing long-term contracts in procurement and buyer-supplier relationships. A more extensive discussion of real-world institutions of this type that operate in various contexts may be found in papers by Johnson *et al.* (2002), Kali (1999), Kranton (1996), and Taylor (2000).

Brief literature review This paper relates to several strands of literature. The first strand is repeated games in a community setting, pioneering papers in this literature being Kandori (1992) and Ellison (1994). Our point of departure from that literature is that we incorporate the decision whether to keep interacting with the same partner, and the heterogeneity of types. More relevant to our setting are the papers by Datta (1993) and Ghosh and Ray (1996), who develop and analyze the "building trust" mechanism. We depart from that literature in that we analyze a wider class of equilibria, fully characterize them, and elucidate on the role of the heterogeneity of types in sustaining good equilibria or eliminating bad equilibria (by contrast, heterogeneity in Ghosh and Ray is used to select an equilibrium, using the criterion that bi-lateral deviation is not profitable). A third strand of literature is the (Kreps-Wilson/Milgrom-Roberts) reputation literature, which in a context similar to ours is found in papers by Watson (1999, 2002). These papers study a fixed relationship between two players and, as such, do not consider the possibility of endogenously forming long-term relationship and the impact of the demographics through the endogenous composition of types in the pool of players that seek new partners. Another relevant paper is Sobel (2006). He focuses, however, on the role of labor market aspects, and does not consider the heterogeneity of types. Another paper that focuses on labor market issues, relating to racial discrimination is Eeckhout (2006). He does not study, however, the disciplinary role of the heterogeneity of types.

Other papers include Tirole (1996) and Dixit (2003), who study the role of information intermediaries that make information available to players. They, again, do not study the disciplinary role of endogenizing relationships. A recent paper is Okuno and Fujiwara (2006) that studies a similar formulation to ours, but from an evolutionary perspective and without heterogeneity of types. The endogenous determination of the type distribution and its welfare implications are not studied in any of the above papers.

More broadly, our paper relates to three major themes in economic theory. One theme is that one may sustain cooperation under long-run competition via promises and threats, see Fudenberg and Maskin (1986). Here we offer a different enforcement mechanism, namely, where no player is specifically called upon to inflict a punishment. Rather, punishment is inherent in the fact that it takes time to re-establish a relationship, during which time costs are incurred. This idea brings us to the second theme, which is the efficiency-wage literature, see Shapiro and Stiglitz (1984). In that literature a shirking worker is punished by being unemployed, which is similar to the idea that a player who defects is forced into interactions with players who chronically defect. That literature, however, is couched in a competitive (as opposed to a game-theoretic) setting, and does not consider the heterogeneity of types, and what bearing it has on the type of equilibria that may be sustained. The third theme is the search and matching literature à la Diamond (1982) and Mortensen (1982). In that literature, like here, it takes time to be matched with an acceptable type. On the other hand, that literature does not study the strategic interaction between agents once they are matched.

Preview. The rest of the paper is organized as follows. The next section introduces our framework. In Section 3 we determine when the pure-strategy good equilibrium, in which strategic types always cooperate, exists and how it depends on parameters. In Section 4 we do the same thing with respect to the pure-strategy bad equilibrium in which strategic types defect. In Section 5 we study mixed-strategy equilibria. In Section 6 we classify all equilibria, and relate them to parameter values. In Section 7 we relate social welfare to the heterogeneity of types. And, in section 8, we extend the model to study investment in expanding the range of actions, how it interacts with cooperative behavior in the community game, and what departures may exist between equilibrium investments and socially-optimal investments. Most proofs are found in a technical appendix.

2 Model Formulation

The Environment We consider a community of individuals (or players or agents), modeled as a continuum of measure 1. Time is discrete and the horizon is infinite. Each individual is infinitely lived.

At the beginning of each period, the community is divided into partnerships (or relationships). Then, the following sequence of events occurs. First, each pair of partners plays a prisoners' dilemma game, and each partner chooses either C, which stands for "cooperate," or D, which stands for "defect." The payoff matrix of this game is specified momentarily. Second, after playing this game, each partnership persists with probability ρ, and breaks with probability $1-\rho$. Third, if a partnership persists, the two partners go into a simultaneous-move game, in which each partner makes a stay-or-separate decision. If both partners choose to stay, the current partnership continues into the next period. If at least one partner chooses to separate, or if the partnership (exogenously) breaks, both partners go into a pool of unmatched individuals. No direct payoffs are associated with the stay-or-separate game; its only role is to endogenize the decision whether to interact with the same individual in the next period. Finally, all individuals in the pool of unmatched individuals are randomly matched, so that all individuals are matched at the beginning of the next period.

Since there is a countable number of time periods and a continuum of players, we assume that no player is ever matched with one of his ex-partners. The timing convention we just described is shown in Figure 1.

There are three types of players in the population. There is a measure α of opportunistic types that we denote by O, a measure β of bad types that we denote by B, and a measure γ $(= 1-\alpha-\beta)$ of good types that we denote by G. A G-type player always chooses C in the prisoners' dilemma game, and a B-type player always chooses D. An O-type player chooses either C or D, depending on which gives her a higher payoff (which depends on the equilibrium play). The payoff matrix of an O-type, considered as a row player, is shown in Table 1. The payoff matrix of a G-type is the C row of Table 1, and the payoff of a B-type is the D row.

	C	D
C	a	$-l$
D	b	0

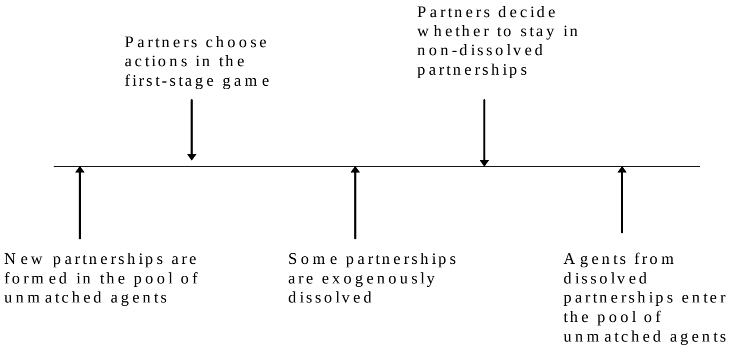
Partners choose actions in the first-stage game
Partners decide whether to stay in non-dissolved partnerships
New partnerships are formed in the pool of unmatched agents
Some partnerships are exogenously dissolved
Agents from dissolved partnerships enter the pool of unmatched agents

Figure 1: Time Line

Table 1: Payoff matrix of an O-type

We assume $0 < a < b$, $0 < l$, and $b - l < 2a$. The first two restrictions say that this game, when played by two O-types, is a prisoners' dilemma game. The third restriction says that the action profile (C, C) maximizes the sum of players' payoffs when the game is played between two O-types. The objective of all players is to maximize the discounted sum of payoffs. The discount factor is common to all players and is denoted by δ, where $\delta \in (0, 1)$.

We assume that monitoring is perfect inside each partnership: A player observes his partner's actions - beginning with the date at which this partnership is commenced. However, when a player is matched to a new partner he knows nothing about the partner's past history of actions with other partners. That is, there are no information flows across matches. Also, a player's type is private information. However, players make statistical inferences about types (of other players), based on the actions they observe. In particular, a player observed to choose C is known not to be a B-type, and a player observed to choose D is known not to be a G-type. Finally, we assume that the configuration of types, (α, β, γ), is common knowledge.

Steady-state equilibria We focus in this paper on a particular class of equilibria, delineated by three properties. The first property is "fresh start": a player's behavior in a new relationship is independent of his past history. The second property is “zero tolerance”: when a player encounters D, he immediately separates from his partner. The third property is “quick familiarity”: a player’s behavior within a relationship depends only on whether the partnership has just started, or whether it is an ongoing relationship. There are other equilibria that do not satisfy one or more of these properties, and we later comment on them. Our focus in this paper, however, is on this class of equilibria because the properties that define them seem to represent realistic features of social interaction, and the model analysis under these properties is quite tractable.[1] Given this, our aim is to provide a complete characterization of equilibria that satisfy these three properties. To simplify the analysis, we further assume that the distribution of types in each phase (defined more precisely below) has settled to a steady-state at $t = 0$. For brevity we call this class of equilibria steady-state equilibria.

Objective of Analysis Having delineated the game and the class of equilibria we focus on, we proceed to analyze them. Specifically, for any configuration of parameter values (i.e., some $(a, b, l, \delta, \rho, \alpha, \beta, \gamma)$-tuple) we determine whether an equilibrium exists, what type of behavior it manifests, and whether it is unique. To this end, we note that some aspects of agents’ behavior are already “hard-wired” into our setting. In particular, G and B-types are hard-wired to play C and D, respectively, in the prisoners’ dilemma game. In addition, we already specified that all player types separate in the stay-or-separate game if they encounter D (and this behavior is optimal because it gives them a chance to interact with players who play C, which generates higher payoffs).[2] Given this, the only aspect of behavior that remains to be endogenously determined is the behavior of O-types in the prisoners’ dilemma game.

It is convenient to analyze and discuss equilibria, using the following terminology: If two partners are about to interact for the first time, we say they are in the stranger phase, denoted S, whereas if they have previously interacted, we say they are in the friendly phase, denoted F.[3] Also, we call a mapping from phases to actions (the object to be determined in equilibrium) a behavior pattern.

[1] Note that the equilibria we derive are such that if all agents adopt strategies that satisfy these properties, the remaining agent’s *unconstrained* best response is to adopt a strategy that satisfies them as well.

[2] G-types always choose C, yet they separate if they encounter D, because they get a higher payoff if they play against an opponent that chooses C as opposed to an opponent that chooses D. In separating, therefore, G-types are seeking future partners that bestow higher payoffs on them, which is driven by discounted payoff maximization.

[3] Terminology borrowed from Ghosh and Ray (1996).

3 The Good Equilibrium

In this section we analyze a pure-strategy equilibrium, referred to as the good equilibrium, in which the behavior pattern of O-types is to play C in both phase S and phase F. That is, O-types behave exactly like G-types.

Steady State This behavior pattern, along with the zero tolerance property, induce a steady-state over the measure of agents in phase S, and its composition. To determine this steady-state, we note that all B-types are always in phase S. In addition, the fact that agents are sometimes exogenously separated implies that a certain measure of G and O-types, henceforth called non-bad types, are also in phase S. We let $x \in [0, 1-\beta]$ be the measure of non-bad types in phase S. Then, the overall measure of agents in phase S is $x+\beta$, and the overall measure of agents in phase F is $1-x-\beta$. In the steady-state of the good equilibrium x satisfies

$$(1-\rho)(1-x-\beta) = x\rho\frac{x}{x+\beta}. \tag{1}$$

To understand (1), note that its left hand side is the measure of agents flowing from phase F into phase S each period. This "inflow" is simply the probability of exogenous dissolutions, $1-\rho$, times the measure of agents in phase F, $1-x-\beta$. The right hand side of (1) is the measure of agents flowing from phase S to phase F each period. This "outflow" is the product of x, which is the measure of agents that could possibly depart phase S, the probability that one of these agents is matched with another non-bad agent, which is $\frac{x}{x+\beta}$, and the probability, ρ, that such a match is not exogenously dissolved after the first interaction. In a steady-state the inflow equals the outflow, which is satisfied for any $x \in [0, 1-\beta]$ that solves (1). There is exactly one such x which, as stated earlier, is the measure of non-bad types in phase S.

As (1) shows, this x depends on β and ρ, but, since the ensuing analysis focuses mostly on the role of β, we consider x as a function of β only, writing it as $x = X(\beta)$. Given $X(\beta)$ and β we define the variable $y = Y(\beta) \equiv \beta/X(\beta)$, which reflects the composition of bad versus non-bad types in phase S. Given the behavior pattern we focus on, y also reflects the composition of *behavior* in phase S, i.e., the ratio of the measure of agents choosing D to the measure of those choosing C. We next state a simple and useful property of $Y(\beta)$.

Lemma 1 *$Y(\beta)$ is increasing in β, ranging from zero to infinity, as β ranges from* 0 *to* 1*.*

Proof. See the Appendix. ■

Value functions Given the behavior pattern prescribed by the good equilibrium and given the steady-state corresponding to it, we define beginning-of-period value functions for O-types. Let V_F and V_S be the discounted payoffs in phases F and S, respectively. Let V_F^d be the discounted payoff when in phase F, deviating to D, and returning to prescribed behavior (i.e., C) thereafter, a one-shot deviation. And let V_S^d be the discounted payoff to a one-shot deviation when in phase S. The equations defining these values are:

$$\begin{aligned} V_F &= a + \delta[\rho V_F + (1-\rho)V_S] && (2) \\ V_S &= \frac{x}{x+\beta}\{a + \delta[\rho V_F + (1-\rho)V_S]\} + \frac{\beta}{x+\beta}(-l + \delta V_S) && (3) \\ V_F^d &= b + \delta V_S && (4) \\ V_S^d &= \frac{x}{x+\beta}(b + \delta V_S) + \frac{\beta}{x+\beta}(0 + \delta V_S). && (5) \end{aligned}$$

To understand how these equations are formed, consider the RHS of (2), which is the discounted payoff of an O-type in state F. This payoff is the sum of two terms: The period payoff a (all agents in phase F are non-bad types, play C and, consequently, receive a), and the continuation payoff: With probability ρ the partnership continues and an O-type gets δV_F; with probability $1-\rho$ the partnership dissolves and an O-type gets δV_S. The remaining three equations are based on a similar logic.

Equations (2) and (3) represent a system of two linear equations in two unknowns, V_F and V_S, so one can explicitly solve them. Doing so we get

$$\begin{aligned} V_F &= \frac{(x+\beta-\delta\beta)a - \beta\delta(1-\rho)l}{(1-\delta)[x+\beta(1-\delta\rho)]} && (6) \\ V_S &= \frac{xa - \beta(1-\delta\rho)l}{(1-\delta)[x+\beta(1-\delta\rho)]}. && (7) \end{aligned}$$

Incentive Constraints Above we considered the "mechanics" of the good equilibrium, computing the steady-state, and O-types' discounted payoffs - *assuming* O-types follow the hypothesized behavior pattern. Now we determine the conditions under which O-types have the incentive to carry out this behavior pattern, i.e., the conditions under which this behavior pattern is part of an equilibrium. For that, the following two incentive constraints must be satisfied:

$$\text{No deviation in phase } F \quad : \quad 0 \leq V_F - V_F^d. \tag{8}$$

$$\text{No deviation in phase } S \quad : \quad 0 \leq V_S - V_S^d. \tag{9}$$

Analysis of these incentive constraints gives the first result.

Lemma 2 *(i) (8) is redundant if (9) is satisfied. (ii) The good equilibrium exists if, and only if,*

$$b - a \leq \frac{\beta}{x+\beta}\delta\rho b - \frac{\beta}{x}(1-\delta\rho)l. \tag{10}$$

Proof. (i) From (4) and (5), we have

$$V_S^d = \frac{x}{x+\beta}V_F^d + \frac{\beta}{x+\beta}\delta V_S.$$

Subtracting this last equation from (3), we get

$$0 \leq V_S - V_S^d \Leftrightarrow 0 \leq \frac{x}{x+\beta}(V_F - V_F^d) - \frac{\beta}{x+\beta}l.$$

Since $0 < l$, this last equivalency shows that (9) implies (8).

(ii) Subtracting (5) from (3), we get

$$V_S - V_S^d = \frac{x}{x+\beta}(-b + V_F - \delta V_S) - \frac{\beta}{x+\beta}l.$$

From (2) we have

$$V_F - \delta V_S = a + \delta\rho(V_F - V_S).$$

Substituting the last equation into the one just before it, we get

$$0 \leq V_S - V_S^d \Leftrightarrow \frac{b-a}{\delta\rho} + \frac{\beta l}{x\delta\rho} \leq V_F - V_S.$$

Solving for $V_F - V_S$ from (6) and (7) and substituting the result into the last inequality, we obtain (10). ■

In words, Lemma 2 tells us two things. The first thing is that it is "safer" to play C in phase F than in phase S. Indeed, in phase F an O-type is sure to encounter C from her partner, resulting in a payoff of a, while in phase S she may encounter D, resulting in a payoff of $-l$. Therefore, if it pays to play C in phase S, it certainly pays to play C in phase F. The second thing that Lemma

2 gives is a reduced-form expression, (10), telling us when O-types optimally choose C, so that the good equilibrium exists.

To elaborate on what (10) entails, let us note that the choice between C and D in phase S is governed by three forces. First, there is the long-term gain of switching from phase S to phase F, which is $V_F - V_S$. Second, there is the probability that this gain is realized, $\frac{x}{x+\beta}$. Third, there is the short-term cost from playing C instead of D: An opportunist gets $-l$ instead of 0 when paired with a bad type, and she gets a instead of b when paired with a non-bad type. Condition (10) summarizes the interplay between these three forces, giving us a reduced-form criterion to determine whether the good equilibrium exists.

Existence of the good equilibrium Inspection of condition (10) reveals that it depends on all parameter values. As stated earlier, however, we wish to isolate the role that the heterogeneity of types plays, i.e., the role that (α, β, γ) plays as regards the existence of the good equilibrium. To this end, we use the definition $y \equiv \frac{\beta}{x}$ to re-write (10) as

$$b - a \leq \frac{y}{1+y}\delta\rho b - y(1-\delta\rho)l \equiv f(y). \tag{11}$$

We give the RHS of (11) a name, $f(y)$, since it will be used frequently in the analysis. Figure 2 shows one possibility for what the graph of f looks like.

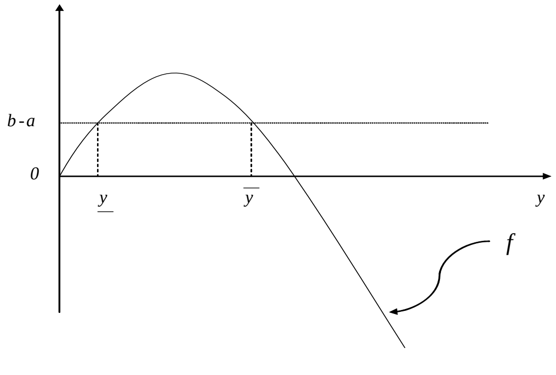

Figure 2: The graph of f

Inspecting (11) we see that its LHS, $b - a$, is positive and independent of y. On the other hand, its RHS is strictly concave in y, goes to 0 as y goes to 0, and goes to $-\infty$ as y goes to ∞ (see

Figure 2). Also, f is uniquely maximized at

$$y^* = \sqrt{\frac{\delta\rho b}{(1-\delta\rho)l}} - 1.$$

Consequently, for the good equilibrium to exist, two conditions must hold: $0 < y^*$, and $b - a \leq f(y^*)$. The first condition is necessary because, if $y^* \leq 0$, then f is strictly decreasing and $f(y) \leq 0$ for all $0 \leq y$, so obviously there is no $0 \leq y$ for which $0 < b - a \leq f(y)$. The second condition is necessary because, if the inequality were reversed, $f(y^*) < b - a$, there would again not be a y for which $b - a \leq f(y)$. After some manipulations, we eliminate the endogenous variable y, and write the two conditions in terms of model primitives only:

$$(1-\delta\rho)l \leq \delta\rho b \text{ and } 4\delta\rho b(1-\delta\rho)l \leq [a + (1-\delta\rho)(l-b)]^2. \tag{12}$$

This analysis shows that (12) is a necessary condition for the existence of the good equilibrium. Condition (12) is also sufficient. Indeed, if (12) is satisfied, then, as shown in Figure 2, there is an interval of y's (a single point "interval" is possible), call it $[\underline{y}, \overline{y}]$, where (11) holds and, thus, where the good equilibrium exists. $\underline{y}$ and $\overline{y}$ are the small and the large roots of the equation $f(y) = b - a$, which are independent of (α, β, γ) (because f is). Since, as per Lemma 1, y is strictly increasing in β, $y \in [\underline{y}, \overline{y}]$ is equivalent to $\beta \in [\underline{\beta}, \overline{\beta}]$, where $\underline{\beta}$ is defined by $\underline{y} = Y(\underline{\beta})$, and $\overline{\beta}$ is defined by $\overline{y} = Y(\overline{\beta})$. Moreover, $[\underline{\beta}, \overline{\beta}]$ does *not* include 0 or 1. This is because when $\beta = 0$, $y = 0$, and $f(0) = 0$. And, when $\beta = 1$, $y = \infty$, and $f(\infty) = -\infty$. Either way, (11) does not hold. Therefore, the interval of β's that satisfy (11) is interior to $(0, 1)$. Finally, observe that criterion (11) is independent of γ, the proportion of good types.

We have now shown how the existence of the good equilibrium depends on the configuration of types. Summarizing our analysis, we have the following result.

Proposition 1 *Hold all parameter values other than (α, β, γ) constant. Then: (i) The existence of the good equilibrium does not hinge on γ, the measure of good types. (ii) If (12) is not satisfied, then there is no β for which the good equilibrium exists. (iii) If (12) is satisfied, then the good equilibrium exists if, and only if, $\beta \in [\underline{\beta}, \overline{\beta}]$, where $0 < \underline{\beta} < \overline{\beta} < 1$, $\underline{\beta}$ and $\overline{\beta}$ being the roots of $f(Y(\beta)) = b - a$.*

The main insight from Proposition 1 is that for the good equilibrium to exist the measure, β, of B-types must not be too small or too large. If β is too small, say $\beta = 0$, the fraction of B-types in phase S is zero, which implies that behavior (under the hypothesized equilibrium strategy) in phase S is the same as behavior in phase F. But, then, there is no punishment for playing D, and no reward for playing C. If an O-type chooses D in phase F, he goes into phase S, where he encounters the same behavior he encountered in phase F, and receives the same payoff, which means he is not being punished. Conversely, if an O-type chooses C in phase S he goes into phase F, where he again encounters the same behavior and receives the same payoff, which means he is not being rewarded. Therefore, if $\beta = 0$, $V_F = V_S$ and the good equilibrium unravels. At the other end of the spectrum, if the measure of B-types is too large, the probability of being matched with a non-bad type in phase S, $\frac{x}{x+\beta}$, is next to nil, which destroys the incentive to play C, and the good equilibrium unravels again. Only if the proportion of B-types is in some intermediate range, not too small to reduce the effectiveness of punishment in phase F, and not too large to discourage cooperation in phase S, does the good equilibrium exist.

Another way to think about the structure of incentives in the good equilibrium is as follows. The proportion of bad types in the community as a whole is β. However, as a result of the equilibrium play, the proportion of bad types in phase S, $\frac{\beta}{\beta+x}$, is bigger than β ($\beta < \frac{\beta}{\beta+x}$ because $\beta + x < 1$). Intuitively, phase S is "contaminated" by a disproportionately large measure of bad types because bad types never leave this phase. But this induces O-types to choose C, because choosing D means going to (or staying at) phase S, interacting with bad types with a non-negligible probability, and receiving low payoffs. Without (a critical mass of) bad types this threat does not exist, and neither does the good equilibrium.

Note also that the measure of G-types has no bearing on the existence of the good equilibrium. The reason for this is that the incentive of an O-type to play C hinges only on the composition of *behavior* in phase S. But, since G-types and O-types behave alike in the good equilibrium, the breakdown between the measures of these types makes no difference. Only the overall measure of non-B-types (or, equivalently, the measure of B-types[4]) makes a difference.

Observe, finally, that the good equilibrium may not exist at all - no matter what β is. This possibility is due to the values that other parameters assume. Most notably, if $b - a$ is large enough, so is the temptation to play D, which destroys the good equilibrium.

[4] Recall that the measures of B and non-B types add up to 1, so any condition on the measure of non-B types is equivalent to a condition on the measure of B-types.

Stability Having commented on the structure of incentives at the good equilibrium, let us now comment on its stability, and on how the good equilibrium compares in this regard to the contagious equilibrium à la Kandori (1992). (The contagious equilibrium is one where a player defects forever, following a defection by himself or by one of his partners).

To this point we assumed that monitoring is perfect within a relationship. Consider now the possibility of observational errors: A player observes her partner to play D (C) with probability $\varepsilon > 0$, even though the partner actually chose C (D). Then, no matter how small ε is, an observational error eventually occurs, i.e., some player is erroneously observed to play D. Once that happens, a contagious process is set in motion under the contagious equilibrium, whereby more and more players defect, so cooperation in the community breaks down. By contrast, consider the good equilibrium in our setting. This equilibrium continues to exist under the presence of observational errors - for conditions analogous to (12), and as long as ε is small enough (one has to appropriately modify the steady-state condition and the incentive constraints to account for the observational errors). More importantly, cooperation does *not* break down in this equilibrium. Intuitively, in the good equilibrium an agent that mis-observes his partner's action separates from the partner, and both get a fresh start in a new relationship next period. In this new relationship, each partner ignores the past and expects (rationally) that playing C bears a chance of being rewarded in the future. Thus, the effect of an observational error is local; it does not trigger the spread of uncooperative behavior, and has no effect on global behavior in the community. This difference between the good equilibrium and the contagious equilibrium comes from the fact that we endogenize separations and re-start of relationships, which is exactly what 'contains' the impact of observational errors.

Let us mention at this juncture that Ellison (1994) proposed - within the context of the contagious equilibrium - a different way to contain the spread of defection. In Ellison's framework the contagious equilibrium is made resilient if players have access to a public randomization device. Such device allows the severity of punishments to be adjusted and coordinated based on the outcome of a device that everyone in the community can perfectly observe. By contrast, such device is not necessary in our framework. Instead, the threat of terminating a relationship and the consequent interaction with bad types are sufficient to enforce cooperative behavior.

Comparative Statistics Since Proposition 1 provides a closed-form criterion (namely, (12)) - written in terms of model primitives - to determine when the good equilibrium exists, one can readily use it to derive comparative statics results. One comparative statics result, which is just a re-statement of Proposition 1, is that the effect of a change in β on the existence of the good equilibrium is non-monotonic: When β is small the effect is positive (an increase in β widens the set of other parameter values under which the good equilibrium exists), but when β is large the effect is negative.

Now we perform a similar comparative statics exercise on other parameters of the model. That is, we show how changes in a, b, l, δ and ρ affect the existence of a good equilibrium. To this end we use condition (10) which determines a region in parameter space where a good equilibrium exists. By studying this condition one determines whether an increase in the value of some parameter expands or contracts this region. Let us re-write condition (10) as follows

$$F(a,b,l,\delta,\rho) \equiv a - (1 - \frac{\beta}{x+\beta}\delta\rho)b - \frac{\beta}{x}(1-\delta\rho)l \geq 0. \tag{13}$$

Then if the derivative of F with respect to some variable, say a, is positive a higher value of this variable expands the region where the good equilibrium exists; otherwise an increase in the value of this variable contracts this region. After some calculations, the result of this exercise is

1. $F_a > 0$.
2. $F_b < 0$.
3. $F_l < 0$.
4. $F_\delta > 0$.

These results are in conformity with the theory of repeated games. For example a higher discount factor makes punishment more severe and thereby expands the region where a good equilibrium exists. Similar interpretations apply to the effect of the component game parameters, a, b and l.

However, the effect of the persistence probability, ρ, is not so conventional, and is, in fact, non-monotonic. In one sense, an increase in ρ, "should be" equivalent to an increase in δ because it prolongs the longevity of relationships and, as such, should always have a positive effect. What we find, instead (under a mild extra restriction), is that the effect is non-monotonic. We first state the result, then explain the intuition.

Proposition 2 *Assume* $(1-\delta)(a+l) < b < \frac{a+l}{1-\delta}$,[5] *and a good equilibrium exists for some value of* ρ*. Then, there exist a* $\underline{\rho}$ *and a* $\overline{\rho} \in (0,1)$*, where* $\underline{\rho} < \overline{\rho}$*, so that the good equilibrium exists if, and only if,* $\rho \in [\underline{\rho}, \overline{\rho}]$*.*

Proof. See the Appendix. ■

The intuition is that an increase in ρ has two effects. The first effect is what we mentioned earlier: An increase in ρ prolongs the expected amount of time spent in phase F and, thus, makes it more rewarding to play C in that phase. The second effect is that an increase in ρ reduces the measure of non-bad types in phase S**.** As a result, an O-type is less likely to be matched with a non-bad type in phase S, which makes it less rewarding to play C in that phase. These two effects work in opposite directions. It turns out that when ρ is small the first effect dominates, whereas when ρ is large the second effect dominates. Thus, in a community setting, a small possibility of exogenous turnover $(1-\rho)$ may help, rather than hinder, cooperation. Another way to look at this is that turnover introduces "fluidity" into the system,[6] enabling movements from phase S to phase F and, thereby, generating incentives to play C in phase S.

Other comparative statics results, namely, with respect to parameters of the constituent game, a, b and l, are derived straightforwardly and conform with expected intuitions; consequently, we do not spell them out here (they may be found in the working paper version).

4 The Bad Equilibrium

In this and the next section we expand our results to other steady-state equilibria. Our analysis here expands the analysis in Section 3 in the sense that we explore the structure of incentives at these other equilibria, and pin down the conditions under which they exist. More broadly, our analysis makes three points. The first point is that good types play a "dual" role vis-à-vis bad types: While an intermediate measure of bad types ensures that the good equilibrium exists, an intermediate measure of good types ensures that the bad equilibrium (namely, the equilibrium in which all O-types play D) does *not* exist. The second point is that when the good equilibrium

[5] This assumption is satisfied if δ is large enough or if $b = a + l$.

[6] When $\rho = 1$ agents are "stuck" in phase S, so there is no long-term reward for playing C. This can be seen from equation (1), which shows that $x = 0$, if $\rho = 1$.

fails to exist for some configuration of parameter values, another steady-state equilibrium may exist. More than that, we show that some steady-state equilibrium exists for *any* configuration of parameter values. The third point is that for some configurations of parameter values, there may exist more than one steady-state equilibrium.

Steady state To start with, we study a pure-strategy equilibrium, that we call the bad equilibrium, in which O-types play D in phase S. Given zero tolerance, B-types and O-types, henceforth called non-good types, are always in phase S. On top of those there is a certain measure of G-types in phase S - because of exogenous dissolutions. Let $x \in [0, \gamma]$ be the measure of G-types in phase S. Then, the steady-state condition corresponding to the bad equilibrium is

$$(1-\rho)(\gamma - x) = x\rho \frac{x}{x+1-\gamma}. \tag{14}$$

Analogous to (1), the solution to (14) determines x as a function of γ, which we continue to call $X(\gamma)$. Likewise, we let the ratio of non-good types to good type in phase S be $y = Y(\gamma) \equiv \frac{1-\gamma}{X(\gamma)}$, which, as before, is also the ratio of the measure of agents choosing D to the measure of those choosing C in phase S. Similar to the good equilibrium, one shows that Y is strictly decreasing in γ, approaches 0 as γ goes to 1, and approaches ∞ as γ goes to 0.

Value Functions and Incentive Constraints Since the hypothesized behavior pattern of O-types here is such that they play D in phase S, they are never in phase F. Nevertheless, to check whether this strategy is part of an equilibrium, the choice in phase F has to be specified. Obviously, there are two possible specifications: either play D, or play C in phase F. We analyze these two possibilities in turn.

- **O-types play D in phase F**

We first define value functions. The notation is similar to that of the previous section, except that the hypothesized behavior pattern in the bad equilibrium is different. This generates a different steady-state and different period payoffs. Making the requisite adjustments, the new value functions are:

$$
\begin{aligned}
V_F &= b+\delta V_S && (15)\\
V_S &= \frac{x}{x+1-\gamma}b+\delta V_S && (16)\\
V_F^d &= a+\delta[\rho V_F+(1-\rho)V_S] && (17)\\
V_S^d &= \frac{x}{x+1-\gamma}\{a+\delta[\rho V_F+(1-\rho)V_S]\}+\frac{1-\gamma}{x+1-\gamma}(-l+\delta V_S). && (18)
\end{aligned}
$$

Given these value functions, the incentive constraints are:

$$
\begin{aligned}
\text{No deviation in phase } F &: \quad 0 \leq V_F - V_F^d. && (19)\\
\text{No deviation in phase } S &: \quad 0 \leq V_S - V_S^d. && (20)
\end{aligned}
$$

Analyzing these constraints, we have the following result.

Lemma 3 *(i) (20) is redundant if (19) is satisfied. (ii) A bad equilibrium in which O-types defect in phase F exists if, and only if,*

$$\frac{1-\gamma}{x+1-\gamma}\delta\rho b \leq b-a. \tag{21}$$

Proof. See the Appendix. ■

Although Lemma 3 is the analogue of Lemma 2, two differences should be noted. First, the binding incentive constraint here is in phase F, not in phase S. Second, $b-a$ has to be bigger, not smaller, than some threshold value. This is due to the fact that in the bad equilibrium opportunists are supposed to defect, not cooperate.

- **O-types play C in phase F**

We carry out similar analysis as in the last case. For brevity, we just report the end result (a proof is found in the appendix).

Lemma 4 *A bad equilibrium in which O-types play C in phase F exists if, and only if,*

$$\frac{1-\gamma}{x+1-\gamma}\delta\rho b-\frac{1-\gamma}{x}(1-\delta\rho)l \leq b-a \leq \frac{1-\gamma}{x+1-\gamma}\delta b. \tag{22}$$

Unlike in Lemmas 2 and 3, no deviation in phase F does not imply no deviation in phase S, and no deviation in phase S does not imply no deviation in phase F. That is why two inequalities (rather than one) have to be satisfied in condition (22).

Combining Lemma 3 and Lemma 4, we see that a bad equilibrium exists if, and only if,

$$\frac{1-\gamma}{x+1-\gamma}\delta\rho b - \frac{1-\gamma}{x}(1-\delta\rho)l \leq b-a. \tag{23}$$

Existence of the bad equilibrium As we did with the good equilibrium, we transform condition (23) to a condition that involves only the primitive data. To this end we re-write the RHS of (23) in terms of y, giving us:

$$\frac{y}{1+y}\delta\rho b - y(1-\delta\rho)l \leq b-a. \tag{24}$$

As can be readily seen, (24) is similar to (11), with $1-\gamma$ replacing β and reversing the inequality. Thus, following the analysis leading up to Proposition 1, we derive the following result.

Proposition 3 *Hold all parameter values other than* (α, β, γ) *constant. Then: (i) The existence of the bad equilibrium does not hinge on* β*, the proportion of bad types. (ii) If (12) is not satisfied, then the bad equilibrium exists for any* γ*. (iii) If (12) is satisfied, then the bad equilibrium exists if, and only if,* $\gamma \in [0, \underline{\gamma}] \cup [\overline{\gamma}, 1]$*, where* $\underline{\gamma}$ *and* $\overline{\gamma}$ *are found by solving* $f(Y(\gamma)) = b-a$*, and are such that* $0 < \underline{\gamma} < \overline{\gamma} < 1$*.*

Although Proposition 3 is analogous to Proposition 1, one feature of it merits discussion and comparison to the traditional theory of repeated games. Namely, Proposition 3 shows that the bad equilibrium does not exist for some parameter configurations. This contrasts with the theory of repeated games, where an indefinite repetition of a Nash equilibrium (the bad equilibrium in our context) is the simplest equilibrium to construct. This is still true in our context if we consider a community setting with good types, but *without endogenously formed* long-term relationships. Therefore, Proposition 3 shows that with endogenously formed relationships, a new force comes into play: An opportunist may cooperate in phase S in the hope of hooking up with a good type, entering into phase F, and enjoying high future payoffs. Therefore, having good types *and* the possibility of forming long-term relationships may destroy the bad equilibrium. Proposition 3 pins down the set of circumstances under which this force is sufficiently strong that the bad equilibrium does not exist.

To be more specific about this set of circumstances, Proposition 3 shows that a bad equilibrium does not exist if γ is in some intermediate range. If γ is small, all opportunists playing D in phase S is an equilibrium because the probability of meeting a good type is too small. If γ is big, all opportunists playing D in phase S is again an equilibrium, since the difference between the continuation payoffs in phase F and phase S is too small. Thus, in both cases the bad equilibrium exists. However, if γ is in some intermediate range, opportunists in phase S have a reasonable chance of meeting a good type, and opportunists in phase F enjoy a significantly higher continuation payoff than in phase S, so they cooperate. Thus, the bad equilibrium does not exist when γ is in this range.

A convenient feature of Propositions 1 and 3 that we are going to exploit later is that there is a duality between the existence of the good equilibrium and the non-existence of the bad equilibrium. The incentive of an opportunist to cooperate in phase S (which is what it means for the good equilibrium to exist, or for the bad equilibrium not to exist) depends on the proportion of agents cooperating in that phase. Since this proportion is strictly decreasing in β in the good equilibrium and strictly increasing in γ in the bad equilibrium, there is a duality between β and γ: If the good equilibrium exists for some β, then the bad equilibrium does not exist for $\gamma = 1 - \beta$, and if the bad equilibrium does not exist for some γ, then the good equilibrium exists for $\beta = 1 - \gamma$. One implication of this property is that $\overline{\beta} = 1 - \underline{\gamma}$, and $\underline{\beta} = 1 - \overline{\gamma}$.[7]

5 The Mixed Strategy Equilibrium

In this section we study mixed-strategy equilibria in which the behavior pattern of O-types is to mix instead of play a pure strategy (which is what they do in the good and the bad equilibria). Since opportunists may mix in either or both phases, there are possibly 5 types of mixed strategy equilibria, which are listed below according O-types' behavior pattern.

1. Randomize in phase S, play C in phase F.
2. Play D in phase S, randomize in phase F.
3. Play C in phase S, randomize in phase F.

[7] Another feature of this duality is that the presence of bad types gives rise to the good equilibrium, while the presence of good types does not. Analogously, the presence of good types eliminates the bad equilibrium, while the presence of bad types does not.

4. Randomize in phase S, play D in phase F.
5. Randomize in phase S, randomize in phase F.

The following lemma shows that some types of mixed strategy equilibria are not possible.

Lemma 5 *For any configuration of parameter values, an equilibrium of type 1 or 2 might exist but equilibria of type 3, 4 or 5 cannot exist.*

Proof. See the Appendix. ■

The intuition for Lemma 5 is related to the result of Lemma 2: if an O-type has a preference for C over D in phase S, then he has a stronger preference for C in phase F. Or, technically stated, equilibrium behavior patterns are monotonically increasing. This rules out equilibria in which the behavior pattern is decreasing, which is the case with equilibria of type 3, 4 and 5. Equilibria of type 2 are payoff- and equilibrium-behavior equivalent to the bad equilibrium that is already analyzed in Section 4. Consequently from this point on we focus on equilibria of type 1, investigating the circumstances under which it gives rise to an equilibrium. As a matter of notation, we let λ be O-types' probability of playing D in phase S and, as a matter of focus, we analyze completely mixed-strategy equilibria, i.e., where $\lambda \in (0, 1)$.

Steady state and value functions In a mixed-strategy equilibrium good types, bad types and opportunistic types all behave differently. This requires the introduction of additional notation. Let x_α be the measure of O-types, and let x_γ be the measure G-types in phase S. The steady-state of a mixed-strategy equilibrium is characterized by a pair $(x_\alpha, x_\gamma) \in [0, \alpha] \times [0, \gamma]$, which satisfies

$$(1-\rho)(\alpha - x_\alpha) = (1-\lambda)x_\alpha \rho \frac{(1-\lambda)x_\alpha + x_\gamma}{x_\alpha + x_\gamma + \beta} \qquad (25)$$

$$(1-\rho)(\gamma - x_\gamma) = x_\gamma \rho \frac{(1-\lambda)x_\alpha + x_\gamma}{x_\alpha + x_\gamma + \beta}. \qquad (26)$$

Let $z \equiv x_\alpha + x_\gamma$ be the measure of non-bad types in phase S, and $x \equiv (1-\lambda)x_\alpha + x_\gamma$ be the measure of non-bad types that play C in phase S. Then, $\beta + z$ is the overall measure of types in phase S, and $\frac{\beta+z-x}{x}$ is the ratio of the measure of agents playing D to the measure of agents playing C in phase S.[8]

[8] $\beta + z$ is the analogue of $\beta + x$ in the good equilibrium and $1-\gamma+x$ in the bad equilibrium; $\frac{\beta+z-x}{x}$ is the analogue of $\frac{\beta}{x}$ in the good equilibrium and $\frac{1-\gamma}{x}$ in the bad equilibrium.

The value functions of O-types, defined under this mixed behavior pattern, are:

$$\begin{aligned} V_F &= a + \delta[\rho V_F + (1-\rho)V_S^C] && (27)\\ V_S^C &= \frac{x}{z+\beta}V_F + \frac{z+\beta-x}{z+\beta}(-l+\delta V_S^C) && (28)\\ V_F^d &= b + \delta V_S^C \\ V_S^D &= \frac{x}{z+\beta}(b+\delta V_S^C) + \frac{z+\beta-x}{z+\beta}(0+\delta V_S^C), \end{aligned}$$

where the superscripts (C or D) on V_S refer now to (candidate) equilibrium behavior, rather than to deviation from such behavior, while the superscript (d) on V_F continues to refer to deviation.

Incentive constraints This mixed behavior pattern is an equilibrium if, and only if, analogous incentive constraints are satisfied. After some manipulations, we simplify these constraints as follows.

$$\text{No-deviation in phase } F\text{: } 0 \leq V_F - V_F^d \Leftrightarrow \frac{b-a}{\delta\rho} \leq V_F - V_S. \quad (29)$$

$$\text{Indifference in phase } S\text{: } V_S^D = V_S^C \Leftrightarrow V_F - V_S = \frac{b-a}{\delta\rho} + \frac{(z+\beta-x)l}{\delta\rho x}. \quad (30)$$

Since the RHS of (30) exceeds the RHS of (29), it suffices to require (30), which we re-write - after solving for V_F and V_S - as:

$$\frac{xa - (1-\delta\rho)(z+\beta-x)l}{(z+\beta)(1-\delta\rho)+\delta\rho x} = \frac{xb}{z+\beta}. \quad (31)$$

As before, letting $y \equiv \frac{\beta+z-x}{x}$, equation (31) is re-written as

$$b - a = \frac{y}{1+y}\delta\rho b - y(1-\delta\rho)l \equiv f(y). \quad (32)$$

Existence of mixed-strategy equilibria We note that (32) is the same as (11), except that an equality is in place of the inequality. This narrows down the set of y's that can be associated with a mixed-strategy equilibrium to at most two values, $\underline{y}$ and $\overline{y}$, which are the small and the large roots of (32). From the discussion in Section 3 we know that if (12) is not satisfied, there are no roots to equation (32) and, hence, no mixed-strategy equilibria. Therefore, to proceed, we assume that (12) is satisfied.

Analogous to previous notation, the dependence of y on λ is denoted as $y = Y(\lambda)$. Observe now that when $\lambda = 0$, $y = \frac{\beta}{x_G}$, where x_G satisfies the steady-state condition of the good equilibrium (under β), (1), and that when $\lambda = 1$, $y = \frac{1-\gamma}{x_B}$, where x_B satisfies the steady-state condition of the bad equilibrium, (14). Furthermore, straightforward calculations show that for any (α, β, γ), $\frac{\beta}{x_G} < \frac{1-\gamma}{x_B}$, and that $Y(\lambda)$ is strictly increasing in λ.[9] Therefore, as λ varies over $[0, 1]$, the value of y varies over $[\frac{\beta}{x_G}, \frac{1-\gamma}{x_B}]$. Combining this with the fact that the y associated with any mixed strategy equilibrium is either $\underline{y}$ and $\overline{y}$, we conclude that a mixed-strategy equilibrium exists if, and only if, at least one of $\underline{y}$ or $\overline{y}$ is in $(\frac{\beta}{x_G}, \frac{1-\gamma}{x_B})$, and that a mixed-strategy equilibrium is unique if exactly one of $\underline{y}$ or $\overline{y}$ is in $(\frac{\beta}{x_G}, \frac{1-\gamma}{x_B})$.

To be more precise about the set of circumstances under which a mixed strategy equilibrium exists, consider the condition $\frac{\beta}{x_G} < \underline{y} < \frac{1-\gamma}{x_B}$. The LHS of this condition is equivalent to $\beta < \underline{\beta}$ and the RHS is equivalent to $\gamma < \overline{\gamma}$; this follows from the monotonicity of $\frac{\beta}{x_G}$ in β, and $\frac{1-\gamma}{x_B}$ in γ, and from the definitions of $\underline{\beta}$ and $\overline{\gamma}$. If this condition is satisfied, i.e., if $(\beta, \gamma) \in [0, \underline{\beta}) \times [0, \overline{\gamma})$, a $\lambda \in (0, 1)$ can be found which gives rise to a mixed-strategy equilibrium in which the ratio of the measures of agents choosing D to agents choosing C is $\underline{y}$. Likewise, the condition $\frac{\beta}{x_G} < \overline{y} < \frac{1-\gamma}{x_B}$ is equivalent to $\beta < \overline{\beta}$ and $\gamma < \underline{\gamma}$, and when this condition is satisfied, a $\lambda \in (0, 1)$ can be found which gives rise to a mixed-strategy in which the ratio of the measures of agents choosing D to agents choosing C is $\overline{y}$. This gives us a complete characterization of when mixed-strategy equilibria exist as a function of underlying parameters. We summarize this analysis as follows.

Proposition 4 *Hold all parameter values other than (α, β, γ) constant. Then, if (12) is violated, there are no mixed-strategy equilibria. If (12) holds, then: (i) A mixed-strategy equilibrium exists if, and only if, there is a $\lambda \in (0, 1)$ so that (25), (26) and (32) are satisfied. (ii) This holds if, and only if, $(\beta, \gamma) \in [0, \underline{\beta}) \times [0, \overline{\gamma}) \cup [0, \overline{\beta}) \times [0, \underline{\gamma})$. (iii) A mixed-strategy equilibrium is unique if, and only if, $(\beta, \gamma) \in [0, \underline{\beta}) \times [0, \overline{\gamma})$ or $(\beta, \gamma) \in [0, \overline{\beta}) \times [0, \underline{\gamma})$, but not both. (iv) In any mixed-strategy equilibrium the ratio of the measure of agents playing D to the measure of agents playing C in phase S is either $\underline{y}$ or $\overline{y}$, where $\underline{y}$ and $\overline{y}$ are the small and the large roots of $f(y) = b - a$.*

Having shown the set of circumstances under which a mixed-strategy equilibrium can be constructed and how to compute it, let us comment now on how this mixed-strategy equilibrium relates to the procedure for constructing mixed-strategy equilibria in general, and how it relates to the pure-strategy equilibria we studied in Sections 3 and 4. To be concrete we make these comments

[9] This is parallel to the property that y is increasing in β for the good equilibrium, and in $1 - \gamma$ for the bad equilibrium.

for parameter configurations in the domain $(\beta,\gamma) \in [0,\underline{\beta}) \times (\underline{\gamma},\overline{\gamma})$. We know - from Propositions 1 and 3 - that a pure-strategy equilibrium does not exist for such parameter values, and we also know - from Proposition 4 - that a mixed strategy equilibrium does.

1. Let $(\beta,\gamma) \in [0,\underline{\beta}) \times (\underline{\gamma},\overline{\gamma})$. Then, if all opportunists play C (which is what they do in the good equilibrium), $y < \underline{y}$ (because $\beta < \underline{\beta}$), which implies that an opportunist is better off playing D. On the other hand, if all opportunists play D, $\underline{y} < y < \overline{y}$ (because $\underline{\gamma} < \gamma < \overline{\gamma}$), which implies that an opportunist is better off playing C. As usual, the existence of such "cycle" suggests that a mixed strategy equilibrium may be found by letting *some* opportunists play C and others play D, or, more precisely, by finding an intermediate value of $\lambda \in (0,1)$, so that when a measure λ of opportunists play D and a measure $1-\lambda$ play C, each opportunist's choice is a best response to others' choices.

2. One way to think about the mixed strategy equilibrium is that it endogenizes the measure of bad types. Indeed, there is a measure β of bad types to begin with, but the measure of agents that play D (which is the behavior manifested by bad types) is actually $\underline{\beta}$, where $\beta < \underline{\beta} = \beta + z - x$. This, in effect, means that the measure of bad types is endogenously increased via uncooperative behavior of opportunists. Alternatively, one may think of the mixed-strategy equilibrium as endogenously increasing the measure of good types from γ to $\overline{\gamma}$.

3. Once the measures of behavioral types is endogenously increased in this way, we can think of the mixed strategy equilibrium as replicating the good equilibrium in a fictional community with $\underline{\beta}$ bad types or, equivalently, as replicating the bad equilibrium in a fictional community with $\overline{\gamma}$ good types. Either way, the measure of agents in phase S is $\beta + z$ and the ratio of the measure of agents playing D to the measure of agents playing C in phase S is $\frac{\beta+z-x}{x} = \frac{\underline{\beta}}{X(\underline{\beta})}$. These two variables, $\beta + z$ and $\frac{\beta+z-x}{x}$, are independent of the particular value that (β,γ) assumes, as long as $(\beta,\gamma) \in [0,\underline{\beta}) \times (\underline{\gamma},\overline{\gamma})$. Therefore, if we define aggregate behavior as this pair of variables, we see that aggregate behavior in the community, at this mixed-strategy equilibrium, is the same for all $(\beta,\gamma) \in [0,\underline{\beta}) \times (\underline{\gamma},\overline{\gamma})$.

Likewise, mixed-strategy equilibria over other regions in the parameter space are equivalent to pure-strategy equilibria (good or bad) in fictional communities with $\underline{\beta}$ or $\overline{\beta}$ bad types, or $\underline{\gamma}$ or $\overline{\gamma}$ good types. As stated earlier, what mixed strategies do is to (endogenously) increase the measure of bad types to $\underline{\beta}$ or $\overline{\beta}$ and the measure of good types to $\underline{\gamma}$ or $\overline{\gamma}$, enabling thereby the construction of a pure-strategy equilibrium. This trick works whenever there are sufficiently many opportunists to increase the measure of behavioral types to the requisite critical values. Obviously, this trick does not work to *decrease* the measures of bad or good types (and it, obviously, does not work to transform the behavior of behavioral types).

6 Summary of Steady-State Equilibria and Discussion of Other Equilibria

6.1 Classification of Equilibria

Propositions 1, 3, and 4 give a complete picture of how parameter configurations relate to different types of steady-state equilibria. In particular, taking some configuration of parameter values, we are now able to tell whether some steady-state equilibrium exists for this configuration and, if so, whether it is unique and of which type(s) it is. To graphically illustrate the result, we fix the values of all parameters other than (α, β, γ), and show how the equilibrium depends on (α, β, γ) only. Since $\alpha + \beta + \gamma = 1$, it is convenient to represent the various (α, β, γ)-triples in the simplex $\beta + \gamma \leq 1$, which is shown in Figure 3.

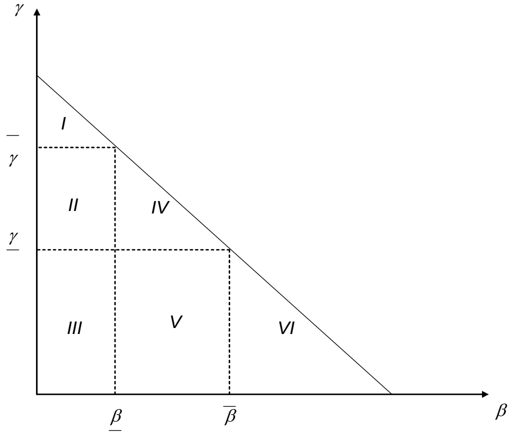

Figure 3: Classification of Equilibrium Outcomes

To elaborate on what Figure 3 shows, let us first consider the existence of pure-strategy equilibria. We know from Propositions 1 and 3 that the good equilibrium exists if and only if $\beta \in [\underline{\beta}, \overline{\beta}]$, and that the bad equilibrium exists if and only if $\gamma \notin (\underline{\gamma}, \overline{\gamma})$. Also, due to duality, $\underline{\gamma} = 1 - \overline{\beta}$ and $\overline{\gamma} = 1 - \underline{\beta}$. Because of this, the simplex $\beta + \gamma \leq 1$ is partitioned into six regions.[10] In regions I, IV, and VI, the bad equilibrium exists, while the good equilibrium does not exist. In region III, the good equilibrium exists, while the bad equilibrium does not exist. In region V, both the good and the bad equilibria exist. In region II, neither the good nor the bad equilibrium exists.

Let us turn now to mixed-strategy equilibria, determining whether they exist in each of the above six regions, whether they are unique, and what type of behavior they manifest. Proposition 4 and the discussion following it provide complete answers to these questions and, re-stating these answers in terms of the geometry of Figure 3, we have the following summary. There are two mixed strategy equilibria in region IV because we can either increase β to $\underline{\beta}$ or increase γ to $\overline{\gamma}$. On the other hand, there are no mixed strategy equilibria in regions I, III, or VI because neither β nor γ can be increased to bring them into a region in which a pure strategy equilibrium exists. Finally, a unique mixed strategy equilibrium exists in region II because, although both β or γ may be increased, the two increases lead to equivalent equilibria (corresponding either way to a ratio $\underline{y}$ of defectors to cooperators in phase S). And, likewise, a unique mixed strategy equilibrium exists in region V (with a ratio $\overline{y}$ of defectors to cooperators).

We summarize the existence of pure and mixed-strategy equilibria in Table 2.

Table 2: Characterization of Equilibria

Regions	Pure-strategy equilibria	Mixed-strategy equilibria
I	Bad equilibrium	None
II	None	One replicating $\underline{y}$
III	Good equilibrium	None
IV	Bad equilibrium	Two
V	Both equilibria	One replicating $\overline{y}$
VI	Bad equilibrium	None

[10] In most statements below a region is understood as the interior of a region.

In conclusion, our analysis, as summarized in Table 2, shows that a steady-state equilibrium exists for *each* configuration of parameter values, and that the equilibrium is sometimes, but not always, unique. The analysis also shows, for each of the six regions whether zero, one, or two pure-strategy equilibria exist, and whether zero, one, or two mixed-strategy equilibria exist.

A numerical example We illustrate this characterization by means of a numerical example. Let us specify parameter values, other than the configuration of types, as follows:

$$a = 4, b = 6, l = 2, \delta = 0.9, \rho = 0.9.$$

Then, it is readily verified that (12) is satisfied for these parameter values, which, as per Proposition 1, means that the good equilibrium exists for a range of β values. Indeed, the good equilibrium exists if, and only if, $f(Y(\beta)) \leq 2 = b - a$. The two roots of $f(Y(\beta)) = 2$ are $\underline{\beta} = 0.143$ and $\overline{\beta} = 0.702$. Therefore, the good equilibrium exists if, and only if, $\beta \in [0.143, 0.702]$. By duality, the bad equilibrium does not exist if and only if $\gamma \in (0.298, 0.857)$. Table 3 specializes Table 2 to these numerical results, and provides examples of mixed-strategy equilibria.

Table 3: Numerical Example

Regions	Parameter Values	Pure equilibria	Mixed equilibria
I	$\beta \in [0, 0.143); \gamma \in [0.857, 1]$	Bad	None
II	$\beta \in [0, 0.143); \gamma \in (0.298, 0.857)$	None	$\beta = 0.1, \gamma = 0.5$; $\underline{\lambda} = 0.406$
III	$\beta \in [0.143, 0.702]; \gamma \in (0.298, 0.857)$	Good	None
IV	$\beta \in [0, 0.143); \gamma \in [0, 0.298]$	Bad	$\beta = 0.1, \gamma = 0.2$; $\underline{\lambda} = 0.271, \overline{\lambda} = 0.936$
V	$\beta \in [0.143, 0.702]; \gamma \in [0, 0.298]$	Both	$\beta = \gamma = 0.2$; $\overline{\lambda} = 0.914$
VI	$\beta \in (0.702, 1]; \gamma \in [0, 0.298]$	Bad	None

6.2 Other Equilibria

To show that equilibria other than the ones we have analyzed exist, we discuss now an equilibrium that violates quick familiarity. Specifically, this equilibrium is such that any relationship starts with a "getting acquainted" phase consisting of T periods of playing (D, D) (without endogenous separation) and, then, if the relationship has not been exogenously dissolved, opportunists revert to

(C, C). If such reversion occurs at $T+1$ (which is the analogue of the stranger phase), the partners enter into a friendly phase and keep playing (C, C) until exogenously separated. Otherwise, i.e., if reversion to (C, C) did not occur at $T+1$, the partners endogenously separate. The incentive to play C in the friendly phase of this equilibrium is stronger than in the good equilibrium because the getting acquainted period is longer and, hence, the punishment for defecting is greater. On the other hand, the reward for playing C at the stranger phase $(T+1)$ is smaller than in the good equilibrium - because the average amount of time spent in the friendly phase is the same, but this is followed by a longer stretch of time in which one's opponents are playing D. Further, the structure of incentives in this equilibrium is the same as in the good equilibrium, namely, the binding incentive constraint is in the stranger phase. It follows, then, that this equilibrium exists under a smaller set of parameter values compared to the good equilibrium. Also, players in this equilibrium spend a larger fraction of their time playing D, so payoffs are lower, making this equilibrium is less efficient. This equilibrium is also less efficient than the mixed-strategy equilibrium with the lower λ (over the set of parameter values where both equilibria exist).

7 Welfare

In this section we construct measures of social welfare at certain steady-state equilibria, and show how they relate to the configuration of types, (α, β, γ). We already know from the analysis in Section 6 that some (α, β, γ) configurations give rise to multiple equilibria, so numerous welfare measures may be calculated. To limit the number of cases to report and to prepare for the analysis in the next section, we focus on two calculations. In the first calculation we fix the measure of good types at zero, $\gamma = 0$, and compute welfare as a function of β at the best equilibrium corresponding to this β. Then, in the second calculation, we fix the measure of bad types at zero, $\beta = 0$, and compute welfare as a function of γ at the worst equilibrium.[11] Our measure of welfare is the total per-period payoff to the whole community at the equilibrium in question. Since the overall measure of agents is one, this is the same as the average per-period payoff.

[11] These two calculations relate to our previous results that the presence of B-type can support the good equilibrium and the presence of G-type can upset the bad equilibrium.

Welfare as a function of β Suppose $\gamma = 0$. Then, specializing the analysis in Section 6, we have a tripartite partition. When $\beta < \underline{\beta}$ (region IV), three equilibria exist and the best equilibrium is the mixed-strategy equilibrium replicating $\underline{y}$. When $\underline{\beta} \leq \beta \leq \overline{\beta}$ (region V), two equilibria exist and the best equilibrium is the good equilibrium. When $\overline{\beta} < \beta$ (region VI), the unique steady-state equilibrium is the bad equilibrium.

Altogether, social welfare takes the following form.

$$W(\beta) = \begin{cases} (1-z-\beta)a + (\beta+z-x)\frac{x}{z+\beta}b + x[\frac{x}{z+\beta}a - \frac{\beta+z-x}{z+\beta}l] & \text{if} \quad \beta < \underline{\beta} \\ (1-x-\beta)a + \beta\frac{x}{x+\beta}b + x[\frac{x}{x+\beta}a - \frac{\beta}{x+\beta}l] & \text{if} \quad \underline{\beta} \leq \beta \leq \overline{\beta} \\ 0 & \text{if} \quad \overline{\beta} < \beta \end{cases}, \tag{33}$$

where x in the second line comes from the solution to (1), and x and z in the first line come from the solution to (25) and (26).

To elaborate on how (33) is arrived at, consider the middle term, which applies to the range $\underline{\beta} \leq \beta \leq \overline{\beta}$. Then, as stated above, welfare is evaluated at the good equilibrium. Opportunists in this equilibrium get a period payoff of a in phase F, and get either a or $-l$ in phase S, depending on whom they meet. Bad types get either b or 0, depending again on whom they meet. Using the measures of agents at each phase (which come from the solution to the steady-state equation), we take the average over these payoffs, and get the reported expression.

Analyzing equation (33) we derive the following result, which is graphically illustrated in the left panel of Figure 4.

Lemma 6 *(i) When $\beta < \underline{\beta}$, $W(\beta)$ is constant; (ii) when $\underline{\beta} \leq \beta \leq \overline{\beta}$, $W(\beta)$ is strictly decreasing and is, hence, maximized at $\underline{\beta}$; (iii) when $\overline{\beta} < \beta$, $W(\beta)$ is zero.*

Proof. See the Appendix. ■

The reason W is zero for $\overline{\beta} < \beta$ is that welfare is evaluated at the bad equilibrium, where all agents play D and collect zero. The reason W decreases for $\underline{\beta} \leq \beta \leq \overline{\beta}$ is that welfare is evaluated at the good equilibrium at which having more bad types is not necessary to induce opportunists to play C. As Proposition 1 shows, β is already in the range that induces all opportunists to play C, so having more bad types only reduces the average level of cooperation and, hence, the average payoff in the community. Finally, the reason welfare is constant for $\beta \leq \underline{\beta}$ is that welfare (for each

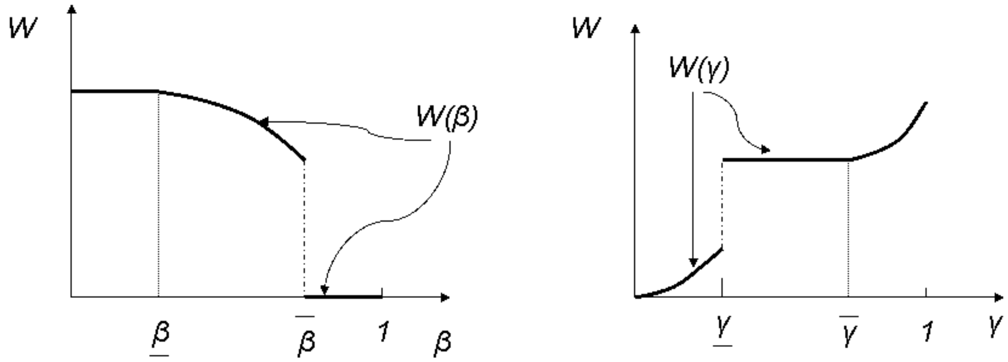

Figure 4: Welfare Measures

β in this range) is measured at the mixed-strategy equilibrium replicating $\underline{y}$. As commented earlier (see comment 3 after Proposition 4), the aggregate behavior in the community at each of these mixed-strategy equilibria is the same and, thus, the aggregate payoff is also the same and is, thus, constant.

An interesting feature of Figure 4 is that welfare decreases discontinuously at $\beta = \overline{\beta}$. The reason for this is that an equilibrium sustaining some cooperation can be achieved for $\beta < \overline{\beta}$ and for $\beta = \overline{\beta}$, but not for β slightly above $\overline{\beta}$ (for $\overline{\beta} < \beta$, the only equilibrium is the bad one). Therefore, as β crosses $\overline{\beta}$, an infinitesimal increase in β has a quantum effect on the degree of cooperation in the community and on welfare.

Welfare as a function of γ Let us turn now to the case where there are no bad types, $\beta = 0$. As γ varies over $[0, 1]$, the worst equilibrium varies as follows: When $\gamma \in [0, \underline{\gamma}]$ or $\gamma \in [\overline{\gamma}, 1]$, the worst equilibrium is the bad equilibrium; and, when $\gamma \in (\underline{\gamma}, \overline{\gamma})$, the unique equilibrium is the mixed-strategy equilibrium replicating $\underline{y}$. Evaluating welfare at these equilibria, we get

$$W(\gamma) = \begin{cases} x(-l) + (\gamma - x)a + (1-\gamma)\frac{x}{x+1-\gamma}b & \text{if} \quad \gamma \leq \underline{\gamma} \text{ or } \overline{\gamma} \leq \gamma \\ (1-z)a + (z-x)\frac{x}{z}b + x[\frac{x}{z}a - \frac{z-x}{z}l] & \text{if} \quad \underline{\gamma} < \gamma < \overline{\gamma} \end{cases} . \tag{34}$$

Analyzing this welfare function we derive the following result, which is analogous to Lemma 6, and is proven in the appendix. A graphical representation of the result is found in the right panel of Figure 4.

Lemma 7 *(i)When $\gamma \in [0, \underline{\gamma}] \cup [\overline{\gamma}, 1]$, $W(\gamma)$ is increasing in γ; (ii) when $\gamma \in (\underline{\gamma}, \overline{\gamma})$, $W(\gamma)$ is constant in γ.*

Intuitively, as γ increases the average cooperation level in the bad equilibrium increases and, thus, social welfare increases. In the mixed-strategy equilibrium replicating $\underline{y}$, aggregate behavior is constant (i.e., independent of γ) and, thus, the social welfare in that equilibrium is constant too.

The relationship between social welfare at the worst equilibrium and γ is plotted in the right panel of Figure 4. Analogous to the best equilibrium, social welfare has an upward jump at $\underline{\gamma}$. This is because the bad equilibrium no longer exists when γ is infinitesimally bigger than $\underline{\gamma}$.

8 Endogenous Determination of the Type-Distribution

In the equilibria we constructed, an opportunist in a long term relationship can get the high payoff a by playing C. He can also get the payoff b by playing D, but that behavior terminates the relationship he is in, so the b payoff is short-lived. Since opportunistic types can choose either C or D, while bad types can only choose D, the equilibrium payoff of opportunistic types is no lower than the equilibrium payoff of bad types (in *any* equilibrium). This suggests that if bad types can somehow expand the range of actions available to them to include C, they might do so even if such expansion is costly. Further, if the decision whether to expand one's set of actions is incorporated into the model, then the distribution of types, which hitherto has been taken as a datum, is itself endogenously determined. The aim of this section is to explore these ideas.[12]

A simple way to analyze the endogenous determination of the type distribution is to assume that initially all individuals are bad types,[13] and that each individual has the option of becoming an opportunistic type by investing $c > 0$.[14] Investment decisions are made independently and

[12] An alternative extension along these lines is a setting in which the measure of bad types declines, while the measure of opportunistic types rises, over time due to evolutionary pressures.

[13] We briefly comment on the effect of having good types at the end of this section.

[14] One interpretation is that cooperation needs skills. By investing in skill acquisition, a B-type is able to play C, hence is transformed to an O-type.

simultaneously. Once these decisions are implemented, the distribution of types in the community is determined, and becomes common knowledge. Then, the infinitely repeated community game is played under this distribution. To limit the number of cases to consider, we assume that players coordinate on the best equilibrium in this community game. We also assume that a steady-state is reached immediately, and that individuals who invest are randomly assigned (at $t = 0$) to phase F or S according to the steady-state probabilities.

We analyze the overall game, using backwards induction. As usual, the equilibrium outcome in the community (sub-)game is what dictates the incentives to invest. In particular, whether and how many individuals invest depends on the level of cooperation in the community game, which in turn depends on how many individuals invested in the first place. Consequently, one goal of the analysis is to reveal the interplay between investments and cooperation in the community.

Before we proceed we note the existence of a degenerate equilibrium in which no one invests. This equilibrium arises because of a coordination problem: It takes a critical mass of agents to invest to make it worthwhile for anyone to invest. In the sequel we focus (naturally) on other equilibria.

Gross Return to Investment To determine equilibrium investments, we first derive the gross return to investment, using the following notation. Let $\pi^O(\beta)$ ($\pi^B(\beta)$) be an O-type's (B-type's) discounted payoff at the best equilibrium under β in the community game. These payoffs are derived from the value functions that correspond to this equilibrium. Specifically,

If $\beta \in (\overline{\beta}, 1]$, payoffs are evaluated at the bad equilibrium, so that

$$\pi^O(\beta) = \pi^B(\beta) = 0.$$

If $\beta \in [\underline{\beta}, \overline{\beta}]$, payoffs are evaluated at the good equilibrium, so that

$$\begin{aligned} \pi^O(\beta) &= \frac{x}{1-\beta} V_S(\beta) + \frac{1-\beta-x}{1-\beta} V_F(\beta) \\ \pi^B(\beta) &= \frac{1}{1-\delta} \frac{x}{x+\beta} b, \end{aligned}$$

where x is the solution to (1) under β, and $V_F(\beta)$ and $V_S(\beta)$ are given by (6) and (7).

Finally, if $\beta \in [0, \underline{\beta})$, payoffs are evaluated at the mixed-strategy equilibrium, so that

$$\begin{aligned} \pi^O(\beta) &= \frac{z-\beta}{1-\beta} V_S(\beta) + \frac{1-z}{1-\beta} V_F(\beta) \\ \pi^B(\beta) &= \frac{1}{1-\delta} \frac{x}{z+\beta} b, \end{aligned}$$

where x and z are derived from the solution to (25) and (26), and $V_F(\beta)$ and $V_S(\beta)$ are derived from the solution to (27) and (28).

Let $\Delta(\beta)$ be the gross return to investment, which is the (discounted) equilibrium payoff difference between being an O-type and a B-type at the best equilibrium in the community game,

$$\Delta(\beta) \equiv \pi^O(\beta) - \pi^B(\beta).$$

Then, we have the following result.

Lemma 8 *(i)* $0 \leq \Delta(\beta)$ *for all* $\beta \in [0,1]$*; (ii)* $0 < \Delta(0)$*, and* $\Delta(\beta)$ *is increasing in* β *for* $\beta \in [0, \underline{\beta}]$*; (iii)* $\Delta(\beta) = 0$ *for* $\beta \in (\overline{\beta}, 1]$*. (iv) Assume* $b \leq a + l$*. Then,* $\Delta(\beta)$ *increases at* $\underline{\beta}$*, and is either increasing throughout* $[\underline{\beta}, \overline{\beta}]$*, or is hump shaped, i.e., there exists a* $\widehat{\beta} \in (\underline{\beta}, \overline{\beta})$*, so that* $\Delta(\beta)$ *is increasing over* $\beta \in [\underline{\beta}, \widehat{\beta})$ *and decreasing over* $(\widehat{\beta}, \overline{\beta}]$*.*

Proof. See the Appendix. ■

The reason that $\Delta(\beta)$ is increasing in β over $[0, \underline{\beta}]$ is that payoffs are evaluated at the mixed-strategy equilibrium. Then, the aggregate behavior in the community is constant in β (see comment 3 after Proposition 4), which implies $\pi^B(\beta)$, $V_S(\beta)$, and $V_F(\beta)$ are constant as well. As a consequence, the only effect of an increase in β is that an O-type has a higher probability of being assigned to phase F (at $t = 0$), which makes $\pi^O(\beta)$ and, consequently, $\Delta(\beta)$ larger.

This effect is also present for $\beta \in [\underline{\beta}, \overline{\beta}]$ (where payoffs are evaluated at the good equilibrium). There is, however, a second effect for $\beta \in [\underline{\beta}, \overline{\beta}]$, which is that $V_F(\beta) - V_S(\beta)$ is increasing in β. These two effects work in opposite directions, resulting in a potentially hump-shaped Δ curve over the domain $[\underline{\beta}, \overline{\beta}]$.

Figure 5 illustrates the content of Lemma 8 (ignore for now the horizontal line with height c).

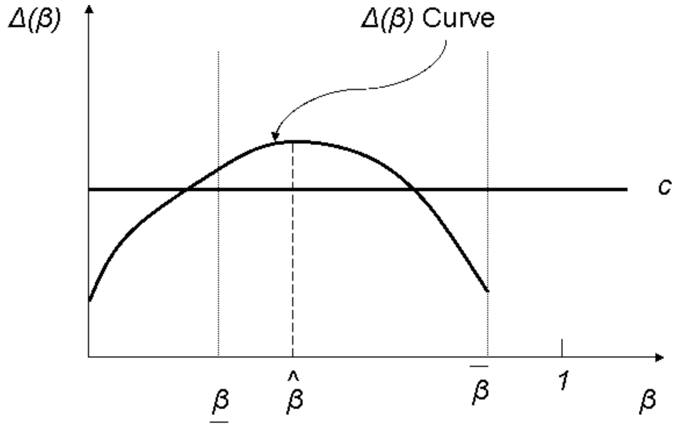

Figure 5: Endogenous Types

Equilibrium in the investment sub-game Given the shape of Δ, as shown in Figure 5, equilibrium investments may be interior (with some, but not all individuals investing), in which case the measure of individuals investing is characterized by indifference between investing and not investing. Alternatively, equilibrium investments may be such that all or none of the individuals invest, in which case the same measure is characterized by a weak preference for the unanimously chosen alternative. In symbols, the possibilities are:

$$\begin{aligned} \text{Everybody invests} &: \quad 0 \leq \Delta(0) - c \\ \text{Some but not all players invest} &: \quad \Delta(\beta) - c = 0 \text{ for some } \beta \in (0,1) \\ \text{Nobody invests} &: \quad \Delta(\beta) - c \leq 0 \text{ for all } \beta \in [0,1]. \end{aligned}$$

To determine which of these possibilities materializes, let us inspect Figure 5 that shows $\Delta(\beta)$, which is the gross return to investment, along with the horizontal line at height c, which is the cost of investment. This figure is drawn so that the c-line intersects the $\Delta(\beta)$-curve at two points. The other possibilities for drawing this figure are that the c-line lies entirely above the $\Delta(\beta)$-curve, or that it lies below it over the range $[0, \overline{\beta}]$. Which of these possibilities materializes depends on parameter values.

Let us consider the possibility shown in Figure 5. Since $\Delta(\beta)$ is hump-shaped, there are (potentially) two intersection points, giving rise to two equilibria. We rule out the equilibrium at the higher intersection point, because it is unstable. Indeed, suppose that β is increased a bit from this equilibrium value (i.e., that less individuals invest). Then, from Figure 5, at the perturbed point, $\Delta(\beta - \varepsilon) < c$, so less individuals invest, which further increases β, drifting the system away from the original equilibrium value. On the other hand, if we increase β at the equilibrium with the lower intersection point, we get $c < \Delta(\beta - \varepsilon)$, so more individuals invest and β drifts back towards its original equilibrium value. As a consequence, the interior equilibrium at the smaller β is stable, while the other is unstable. We concentrate from this point onwards on the stable equilibrium.

Turning to corner equilibria, Lemma 8 tells us that $0 < \Delta(0)$. Thus, everybody invests if $c \leq \Delta(0)$, and we have a corner equilibrium. At the other end of the spectrum, if the c-line lies entirely above the $\Delta(\beta)$-curve, then no investment is a dominant strategy, and we have the other type of corner equilibrium, with no one investing. Summarizing the analysis, we have the following proposition.

Proposition 5 *(i) If $c \leq \Delta(0)$, then everybody invests. (ii) If $\Delta(0) < c \leq \Delta(\widehat{\beta})$, then somebody but not everybody invests; moreover, the measure of players that invest in the stable equilibrium is decreasing in c. (iii) If $\Delta(\widehat{\beta}) < c$, then nobody invests.*

Proposition 5 shows that the measure of individuals investing and the measure of individuals cooperating in the community game are positively correlated in equilibrium. Indeed, let us consider a decrease in c. Then, the equilibrium measure of individuals investing either increases if this equilibrium is interior, or stays constant if the equilibrium is corner. At the same time, the measure of individuals cooperating increases if the equilibrium β is such that the community is at the good equilibrium, or remains constant if the community is at the mixed-strategy or the bad equilibrium. Whatever combination of these possibilities materializes, a decrease in c induces a non-negative correlation between the measure of individuals investing and the measure of individuals cooperating. Therefore, if we interpret "investing" as acquiring skills (that enable playing C), what this prediction says is that there is both a direct and an induced return to skill acquisition. The direct return is that individuals with skills are more productive. The induced return is that individuals tend to cooperate more when more of them have skills, which further increases the return to skill acquisition.

Contrasting the equilibrium with the Social Optimum We contrast now the equilibrium in the endogenous type distribution game to the planner's optimum.

Proposition 6 *(i) If $c < \Delta(\underline{\beta})$, then individuals over-invest in equilibrium. (ii) If $\Delta(\widehat{\beta}) < c < \frac{W(\underline{\beta})}{1-\underline{\beta}}$, individuals under-invest in equilibrium.*

Proof. (i) If $c < \Delta(\underline{\beta})$, the equilibrium measure of individuals that invest exceeds $1 - \underline{\beta}$. But Lemma 6 tells us that gross welfare, $W(\beta)$, is constant over $[0, \underline{\beta}]$, and we assumed a positive investment cost $c > 0$, so it does not pay - from a social planner's perspective - for more than $1 - \underline{\beta}$ individuals to invest.

(ii) The social planner maximizes $S(\beta) \equiv W(\beta) - c(1 - \beta)$ over β. Given the shape of W (see Lemma 8), if $c < \frac{W(\underline{\beta})}{1-\underline{\beta}}$, then $0 = S(1) < S(\underline{\beta})$, so no one investing cannot be socially optimal. On the other hand, since $\Delta(\widehat{\beta}) < c$, no one invests in equilibrium. ■

Proposition 6 shows two departures of the equilibrium from the social optimum. On the one hand, individuals may under-invest because some of the benefit accrues to others who interact with them in the community game. On the other hand, which might be more surprising, individuals may over-invest. This is because individuals first invest but then "undo" the investments by not cooperating.[15] It may seem bizarre that individuals, on their own volition, will choose to do so. The point, however, is that there is a discrepancy between ex-ante and ex-post incentives. Ex-ante some agents invest because this enables them to enter into long-term, high-paying relationships. Ex-post, when in transit between such relationships, an opportunist has a short-run incentive to defect (and in a mixed-strategy equilibrium some of them do defect). Because of that investments are not fully utilized, which means they had been wasted from a social point of view.

The impact of G-type on the investment game As Proposition 5 shows, an equilibrium with no one investing may occur, depending on parameter values. This was shown on the assumption that all agents are bad types to begin with, which implies the bad equilibrium in the community game is a possibility. Suppose, on the other hand, that the measure of good types satisfies $\gamma \in [\underline{\gamma}, \overline{\gamma}]$.

[15] Another way to think about this is that the maximum cooperation level in the community is reached when there are $0 < \underline{\beta}$ bad types. Further decrease in β cannot increase the cooperation level, since to sustain cooperation a certain fraction of agents has to defect in the stranger phase. Therefore, if more agents than $1-\underline{\beta}$ invest, some agents' investment are "reversed"and are, hence, wasted.

Then, as the analysis in Section 4 shows, the bad equilibrium in the community game is no longer a possibility. As a result, if $\Delta(\widehat{\beta}) < c < \widetilde{\Delta}(1-\gamma)$, where $\widetilde{\Delta}$ is the analogue of Δ in a community with good types, the no investment equilibrium that would have occurred without good types no longer occurs. From this we conclude that the presence of good types can have a good influence on the investment behavior of bad types, and help agents reach a more efficient outcome.

References

[1] Dixit, A. "On Modes of Economic Governance" *Econometrica*, 2003, 71(2), 449-481.

[2] Datta, "Building Trust", 1993, Mimeo. London School of Economics.

[3] Diamond, P. "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, 1982, 90(5), 881-894.

[4] Eeckhout, J. "Minorities and Endogenous Segregation", *Review of Economic Studies*, 2006, 73(1), 31-53.

[5] Ellison, G. "Cooperation in the Prisoners-Dilemma with Anonymous Random Matching", *Review of Economic Studies*, 1994, 61(3), 567-88.

[6] Fudenberg, D. and Maskin, E., "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information," *Econometrica*, 1986, 54(3), 533-554.

[7] Okuno-Fujiwara, M. and Fujiwara-Greve, T. "Voluntarily Separable Prisoners' dilemma," CIRJE Discussion paper, August 2006.

[8] Ghosh, P. and Ray, D. "Cooperation in Community Interaction without Information Flows", *Review of Economic Studies*, 1996, 63(3), 491-519.

[9] Johnson, S., McMillan J., and Woodruff C. "Courts and Relational Contracts", *Journal of Law Economics & Organization*, 2002, 18(1), 211-277.

[10] Kali, R. "Endogenous Business Networks", *Journal of Law Economics & Organization*, 1999, 15(3), 615-36.

[11] Kandori, M. "Social Norms and Community Enforcement", *Review of Economics Studies*, 1992, 59(1), 63-80.

[12] Kranton, R. "The Formation of Cooperative Relationships", *Journal of Law Economics & Organization*, 1996, 12(1), 214-33.

[13] Mortensen, D., "Property Rights and Efficiency in Mating, Racing, and Related Games," *American Economic Review*, 1982, 72(5), 968-979.

[14] Shapiro, C. and J. Stiglitz, “Equilibrium Unemployment as a Worker Disciplinary Device,” *American Economic Review*, 1984, 74(3), 433-444.

[15] Sobel, J. “For Better or Forever: Formal versus Informal Enforcement” *Journal of Labor Economics*, 2006, 24(2), 271-297.

[16] Taylor, C. “The Old-Boy Network and the Young-Gun Effect”, *International Economic Review*, 2000, 41(4), 871-91.

[17] Tirole, J. “A Theory of Collective Reputations”, *Review of Economic Studies*, 1996, 63(1), 1-22.

[18] Watson, J. “Starting Small and Renegotiations, *Journal of Economic Theory*, 1999, 85(1), 52-90.

[19] Watson, J. “Starting Small and Commitment”, *Games and Economic Behavior*, 2002, 38(1), 176-199.

9 Appendix

Proof of Lemma 1

Proof. The solution to (1) is

$$x = \frac{(1-\rho)(1-2\beta) + \sqrt{(1-\rho)^2 + 4\beta(1-\beta)\rho(1-\rho)}}{2}. \tag{35}$$

Dividing (35) by β we get

$$\frac{1}{y} = \frac{x}{\beta} = \frac{(1-\rho)(\frac{1}{\beta}-2) + \sqrt{\frac{(1-\rho)^2}{\beta^2} + 4(\frac{1}{\beta}-1)\rho(1-\rho)}}{2}. \tag{36}$$

Since all terms in (36) decrease in β, $y(\beta)$ increases in β. Moreover, when $\beta \to 0$, $x/\beta \to \infty$, and $y \to 0$. On the other hand, when $\beta \to 1$, $x \to 0$, and $y \to \infty$. ■

Proof of Proposition 2

Proof. In this proof we hold β constant and consider x and value functions as functions of ρ only, $x(\rho)$, $V_S(\rho)$, etc. We define

$$D(\rho) \equiv V_S(\rho) - V_S^d(\rho) + l.$$

The proof is executed in 5 steps.

In the first step we show that

$$D(\rho) = -\frac{x}{x+\beta}b + \frac{x}{x+\beta(1-\delta\rho)}(a+l). \tag{37}$$

Proof of step 1: Using (3) and (5)

$$\begin{aligned} V_S - V_S^d &= -\frac{\beta}{x+\beta}l - \frac{x}{x+\beta}b + \frac{x}{x+\beta}(V_F - \delta V_S) \\ &= -l + \frac{x}{x+\beta}[l - b + a + \delta\rho(V_F - V_S)], \end{aligned}$$

where the last equality follows from (2) and (3). Now substituting into the last term from (6) and (7) we get

$$\begin{aligned} V_S - V_S^d &= -l + \frac{x}{x+\beta}\{l - b + a + \delta\rho\frac{(\beta-\beta\delta)(a+l)}{(1-\delta)[x+\beta(1-\delta\rho)]}\} \\ &= -l + \frac{x}{x+\beta}[l - b + a + \delta\rho\frac{\beta(a+l)}{x+\beta(1-\delta\rho)}] \\ &= -l - \frac{x}{x+\beta}b + \frac{x}{x+\beta(1-\delta\rho)}(a+l), \end{aligned}$$

which is the desired equation.

In the second step, we show that: (i) $x(0) = 1-\beta$ and $x(1) = 0$. (ii) $\frac{\partial x}{\partial \rho} \mid_{\rho=0} = -(1-\beta)^2 \leq 0$ and $\frac{\partial x}{\partial \rho} \mid_{\rho=1} = -\infty$. (iii) $\frac{\partial x}{\partial \rho} < 0$ and $\frac{\partial^2 x}{\partial \rho^2} < 0$ for all ρ.

Proof of step 2: (i) This follows by substitution into (35).

(ii) Differentiating (35) with respect to ρ we get

$$\frac{\partial x}{\partial \rho} = \frac{2\beta - 1 + \frac{-2(1-\rho)+4\beta(1-\beta)(1-2\rho)}{2\sqrt{(1-\rho)^2+4\beta(1-\beta)\rho(1-\rho)}}}{2} = \frac{2\beta - 1 + \frac{-(1-\rho)+2\beta(1-\beta)(1-2\rho)}{\sqrt{(1-\rho)^2+4\beta(1-\beta)\rho(1-\rho)}}}{2}. \tag{38}$$

Evaluating (38) at $\rho = 0$ we get

$$\frac{\partial x}{\partial \rho} \mid_{\rho=0} = -(1-\beta)^2 \leq 0. \tag{39}$$

Likewise evaluating (38) at $\rho = 1$ we get

$$\frac{\partial x}{\partial \rho} \mid_{\rho=1} = \frac{2\beta - 1 + \frac{-2\beta(1-\beta)}{0}}{2} = -\infty. \tag{40}$$

Given (39), if we show x is strictly concave in ρ it would follow that x decreases in ρ for all $\rho \in [0,1]$.

(iii) Using (38) let's compute the second derivative of x.

$$2\frac{\partial^2 x}{\partial \rho^2} = \frac{[1-4\beta(1-\beta)]\sqrt{(1-\rho)^2+4\beta(1-\beta)\rho(1-\rho)} - \frac{2[-(1-\rho)+2\beta(1-\beta)(1-2\rho)]^2}{2\sqrt{(1-\rho)^2+4\beta(1-\beta)\rho(1-\rho)}}}{(1-\rho)^2+4\beta(1-\beta)\rho(1-\rho)}.$$

This is negative if and only if the numerator is negative and the latter holds if and only if

$$[1-4\beta(1-\beta)][(1-\rho)^2+4\beta(1-\beta)\rho(1-\rho)] < [-(1-\rho)+2\beta(1-\beta)(1-2\rho)]^2.$$

After some calculation, the above inequality is equivalent to

$$\begin{aligned} & 4\beta(1-\beta)\rho(1-\rho) - 4\beta(1-\beta)(1-\rho)^2 + 4\beta(1-\beta)(1-2\rho)(1-\rho) - 16\beta^2(1-\beta)^2\rho(1-\rho) \\ < \; & 4\beta^2(1-\beta)^2(1-2\rho)^2 \end{aligned}$$

and this is true because the first three terms on the LHS add up to zero.

In the third step we show that if ρ is sufficiently small or sufficiently large, $V_S - V_S^d < 0$ and thus that the good equilibrium does not exist for such ρ's.

Proof of step 3: Substitute $\rho = 0$ into (37), recalling that $x(0) = 1-\beta$ (see step 2). Then

$$(V_S - V_S^d)(0) = -l - (1-\beta)b + (1-\beta)(a+l) < -(1-\beta)(b-a) < 0,$$

since $b > a$. Likewise, substitute $\rho = 1$ into (37), recalling that $x(1) = 0$. Then

$$(V_S - V_S^d)(1) = -l < 0.$$

Since D is continuous this completes the proof of step 3.

In the fourth step we show that D is increasing in a neighborhood of $\rho = 0$ and decreasing in a neighborhood of $\rho = 1$ and thus that a $\rho^* \in (0,1)$ exists for which $D'(\rho^*) = 0$ (we show later that ρ^* is unique).

Proof of step 4: Differentiating D we get

$$D'(\rho) = -(\frac{x}{x+\beta})'b + (\frac{x}{x+\beta(1-\delta\rho)})'(a+l). \tag{41}$$

Now let's find the derivative of $\frac{x}{x+\beta}$

$$(\frac{x}{x+\beta})' = \frac{x'(x+\beta) - x'x}{(x+\beta)^2} = \frac{x'\beta}{(x+\beta)^2}. \tag{42}$$

Let's also find the derivative of $\frac{x}{x+\beta(1-\delta\rho)}$

$$(\frac{x}{x+\beta(1-\delta\rho)})' = \frac{x'[x+\beta(1-\delta\rho)] - (x'-\beta\delta)x}{[x+\beta(1-\delta\rho)]^2} = \frac{x'\beta(1-\delta\rho) + \beta\delta x}{[x+\beta(1-\delta\rho)]^2}. \tag{43}$$

Using step 2 and equations (41), (42) and (43), we evaluate D' at $\rho = 0$

$$\begin{aligned} D'(0) &= \beta[-x'b + (x' + \delta(1-\beta))(a+l)] \\ &= \beta(1-\beta)[(1-\beta)(b-a-l) + \delta(a+l)]. \end{aligned}$$

From the above expression, the necessary and sufficient condition for $0 < D'(0)$ (for all β) is that $(1-\delta)(a+l) < b$.

Turning to $\rho = 1$, we repeat the same steps

$$\begin{aligned} D'(1) &= \frac{-x'}{\beta}b + \frac{x'}{\beta(1-\delta)}(a+l) \\ &= \frac{x'}{\beta}[\frac{a+l}{1-\delta} - b]. \end{aligned}$$

Since, by step 2, $x'(1) = -\infty$, the necessary and sufficient condition for $D'(1) < 0$ is that $b < \frac{a+l}{1-\delta}$.

Altogether, in order to have $D'(1) < 0 < D'(0)$, it is necessary and sufficient to have

$$(1-\delta)(a+l) < b < \frac{a+l}{1-\delta},$$

which is a maintained assumption for this Lemma. This completes the proof of step 4.

Let ρ^* be such that $D'(\rho^*) = 0$, i.e., so that

$$\frac{(\frac{x}{x+\beta})'}{(\frac{x}{x+\beta(1-\delta\rho)})'}(\rho^*)b - (a+l) = 0. \tag{44}$$

Since $(\frac{x}{x+\beta})' < 0$ for all ρ (which follows from equation (42) and Lemma 1), (44) implies that $(\frac{x}{x+\beta(1-\delta\rho)})'(\rho^*) < 0$. So to prove that $D'(\rho) < 0$ for $\rho^* < \rho$, it would suffice to prove that $\frac{(\frac{x}{x+\beta})'}{(\frac{x}{x+\beta(1-\delta\rho)})'}$ is decreasing in ρ and that $(\frac{x}{x+\beta(1-\delta\rho)})'(\rho) < 0$ for $\rho^* < \rho$. This is the fifth and final step of the proof.

Proof of step 5: Using (42) and (43) we form and simplify the ratio $\frac{(\frac{x}{x+\beta})'}{(\frac{x}{x+\beta(1-\delta\rho)})'}$.

$$\begin{aligned}\frac{(\frac{x}{x+\beta})'}{(\frac{x}{x+\beta(1-\delta\rho)})'} &= \frac{x'\beta}{(x+\beta)^2}\frac{[x+\beta(1-\delta\rho)]^2}{x'\beta(1-\delta\rho)+\beta\delta x} = \frac{[x+\beta(1-\delta\rho)]^2}{(x+\beta)^2}\frac{x'}{x'(1-\delta\rho)+\delta x} \\ &= [1-\frac{\beta\delta\rho}{x+\beta}]^2\frac{1}{1-\delta\rho+\delta\frac{x}{x'}}.\end{aligned}$$

Since, by step 2, $x' < 0$, the first term is decreasing in ρ. To show that the second term is decreasing, too, we need to show that the derivative of the denominator in the second term is positive. So let's compute this derivative.

$$(1-\delta\rho+\delta\frac{x}{x'})' = -\delta+\delta\frac{x'x'-x''x}{(x')^2} = -\delta+\delta(1-\frac{x''x}{(x')^2}) = -\delta\frac{x''x}{(x')^2}.$$

For this to be positive we require x'' to be negative. But this is already shown in step 2.

To show that $(\frac{x}{x+\beta(1-\delta\rho)})'(\rho) < 0$ for all $\rho^* < \rho$ we use (43)

$$(\frac{x}{x+\beta(1-\delta\rho)})' = \frac{x'\beta(1-\delta\rho)+\beta\delta x}{[x+\beta(1-\delta\rho)]^2}.$$

This is negative if and only if the numerator is negative which, since $x' < 0 < \beta$, is equivalent to $0 < 1-\delta\rho+\delta\frac{x}{x'}$. Now, by the proof of step 4, this holds at $\rho = \rho^*$ and as we have just shown, $1-\delta\rho+\delta\frac{x}{x'}$ is increasing in ρ. So this means $1-\delta\rho+\delta\frac{x}{x'}$ is positive for all $\rho^* < \rho$. This also shows that ρ^* is unique. So the proof of step 5 is complete.

The proof that $0 < D'(\rho)$ for $\rho < \rho^*$ is analogous.

Considering the five steps together, it is seen that the effect of ρ is analogous to the effect of β. Either $D(\rho) - l$ is negative for all $\rho \in (0,1)$ or it is positive for some ρ. In the first instance the good equilibrium does not exist; in the second instance it exists for all ρ's in some interval $[\underline{\rho}, \overline{\rho}]$, where $0 < \underline{\rho} < \overline{\rho} < 1$. ■

Proof of Lemma 3

Proof. (i) Substituting (17) into (18) we get

$$V_S^d = \frac{x}{x+1-\gamma} V_F^d + \frac{1-\gamma}{x+1-\gamma}(-l + \delta V_S).$$

From (16) and the last equality

$$V_S - V_S^d = \frac{x}{x+1-\gamma} b + \delta V_S - \frac{x}{x+1-\gamma} V_F^d - \frac{1-\gamma}{x+1-\gamma}(-l + \delta V_S).$$

Replacing δV_S by $V_F - b$, which is valid by (15), in this last equality and re-arranging we get

$$\begin{aligned} V_S - V_S^d &= \frac{x}{x+1-\gamma} b + V_F - b - \frac{x}{x+1-\gamma} V_F^d - \frac{1-\gamma}{x+1-\gamma}(-l + V_F - b) \\ &= \frac{x}{x+1-\gamma}(V_F - V_F^d) + \frac{1-\gamma}{x+1-\gamma} l. \end{aligned}$$

Since $l > 0$, the last equality shows that (19) implies (20).

(ii) From (15) and (16) we have

$$\rho V_F + (1-\rho) V_S = \frac{x + \rho(1-\gamma)}{x+1-\gamma} b + \delta V_S = \frac{x + \rho(1-\gamma)}{x+1-\gamma} b + V_F - b,$$

where the last equality follows from (15). Therefore (19) is equivalent to

$$\begin{aligned} V_F &\geq a + \delta \left[\frac{x + \rho(1-\gamma) - x - (1-\gamma)}{x+1-\gamma} b + V_F \right] \Leftrightarrow \\ (1-\delta) V_F &\geq a - \delta \left[\frac{(1-\rho)(1-\gamma)}{x+1-\gamma} b \right] = \frac{(x+1-\gamma)a - \delta(1-\rho)(1-\gamma)b}{x+1-\gamma}. \end{aligned}$$

Now we substitute (49) into the LHS, which gives

$$V_F \geq V_F^d \Leftrightarrow \frac{x + (1-\delta)(1-\gamma)}{x+1-\gamma} b \geq \frac{(x+1-\gamma)a - \delta(1-\rho)(1-\gamma)b}{x+1-\gamma}.$$

After some re-arrangement we get (21). ■

Proof of Lemma 4

Proof. In this case, the value functions of O-types are as follows.

$$V_F = a + \delta[\rho V_F + (1-\rho) V_S], \tag{45}$$

$$V_S = \frac{x}{x+1-\gamma} b + \delta V_S, \tag{46}$$

$$V_F^d = b + \delta V_S, \tag{47}$$

$$V_S^d = \frac{x}{x+1-\gamma} V_F + \frac{1-\gamma}{x+1-\gamma}(-l + \delta V_S). \tag{48}$$

Compared to (15)-(18), V_S and V_S^d are the same, whereas V_F and V_F^d are interchanged.

Solving (45) and (46) for V_S and V_F, one gets

$$V_F = \frac{(1-\delta)[x+1-\gamma]a + \delta(1-\rho)xb}{(1-\delta)(1-\delta\rho)[x+1-\gamma]}, \tag{49}$$

$$V_S = \frac{x}{(1-\delta)[x+1-\gamma]}b. \tag{50}$$

Given this solution one solves for V_F^d and V_S^d, using equations (47) and (48).

$$V_F^d = \frac{x+(1-\delta)(1-\gamma)}{(1-\delta)[x+1-\gamma]}b, \tag{51}$$

$$V_S^d = \frac{(1-\delta)[x+1-\gamma][xa-(1-\gamma)(1-\delta\rho)l] + \delta xb[(1-\rho)x+(1-\delta\rho)(1-\gamma)]}{(1-\delta\rho)(1-\delta)[x+1-\gamma]^2}. \tag{52}$$

The behavior pattern described above constitutes an equilibrium if and only if the following incentive constraints are satisfied

$$\text{No deviation in phase } F \quad : \quad V_F - V_F^d \geq 0. \tag{53}$$

$$\text{No deviation in phase } S \quad : \quad V_S - V_S^d \geq 0. \tag{54}$$

Now we use (49), (50), (51) and (52) to verify when (53) and (54) are satisfied. After some calculations we get

$$V_F - V_F^d \geq 0 \Leftrightarrow \tag{55}$$

$$b - a \leq \frac{1-\gamma}{x+1-\gamma}\delta b$$

and

$$V_S - V_S^d \geq 0 \Leftrightarrow \tag{56}$$

$$b - a \geq \frac{1-\gamma}{x+1-\gamma}\delta\rho b - \frac{1-\gamma}{x}(1-\delta\rho)l.$$

Combining (55) and (56), we get the desired inequality (22). ■

Proof of Lemma 5: We prove the statement for type 2-5 equilibria. The case of type 1 equilibria is treated in the text.

Proof. (ii) Suppose that O-types play D with probability 1 in phase S and play D with some positive probability in phase F . Let μ_F be the implied probability that an O-type bumps into another O-type that plays D in phase F. Then the value functions are

$$\begin{aligned}
V_F^C &= (1-\mu_F)\{a+\delta[\rho V_F+(1-\rho)V_S]\}+\mu_F(-l+\delta V_S),\\
V_F^D &= (1-\mu_F)b+\delta V_S,\\
V_S &= \frac{x}{x+1-\gamma}b+\delta V_S,\\
V_S^d &= \frac{x}{x+1-\gamma}V_F+\frac{1-\gamma}{x+1-\gamma}(-l+\delta V_S),
\end{aligned}$$

where x is the steady state proportion of agents playing C in phase S. If this is an equilibrium behavior pattern, the following must hold

$$\begin{aligned}
V_F^C &= V_F^D,\\
V_S &\geq V_S^d.
\end{aligned}$$

But

$$V_S \geq V_S^d \Longleftrightarrow \frac{x}{x+1-\gamma}b+\delta V_S \geq \frac{x(1-\mu_F)b-(1-\gamma)l}{x+1-\gamma}+\delta V_S,$$

which is always satisfied because of our parameter restrictions. Therefore, this behavior pattern is an equilibrium if and only if there is a $\mu_F \in (0,1)$ such that

$$V_F^C = V_F^H \Leftrightarrow -\mu_F l+(1-\mu_F)[a-b+\delta\rho(1-\frac{x}{x+1-\gamma})b]=0. \tag{57}$$

The LHS of (57) strictly decreasing in μ_F and hence maximized at $\mu_F=0$, where it equals

$$a-b+\frac{1-\gamma}{x+1-\gamma}\delta\rho b. \tag{58}$$

Also at $\mu_F=1$ the LHS of (57) is negative. Thus if (58) is positive there must be a μ_F so that (57) is satisfied. Therefore if $b-a<\frac{1-\gamma}{x+1-\gamma}\delta\rho b$, there is a mixed strategy behavior pattern that supports type 2 equilibrium.

(iii) Consider the mixed strategy equilibrium in which O-types randomize in both phases. Consider an O-type and let μ_F (μ_S) be the probability that his partner plays D in phase F (S). The value functions are

$$\begin{aligned}
V_F^H &= (1-\mu_F)\{a+\delta[\rho V_F+(1-\rho)V_S]\}+\mu_F(-l+\delta V_S), && (59)\\
V_F^C &= (1-\mu_F)b+\delta V_S,\\
V_S^H &= (1-\mu_S)\{a+\delta[\rho V_F+(1-\rho)V_S]\}+\mu_S(-l+\delta V_S),\\
V_S^C &= (1-\mu_S)b+\delta V_S.
\end{aligned}$$

The relevant incentive constraints are:

$$\begin{aligned} V_F^H &= V_F^C \Leftrightarrow V_F - V_S = \frac{b-a}{\delta\rho} + \frac{\mu_F}{1-\mu_F}\frac{l}{\delta\rho} > 0, \\ V_S &= V_S^C \Leftrightarrow V_F - V_S = \frac{b-a}{\delta\rho} + \frac{\mu_S}{1-\mu_S}\frac{l}{\delta\rho} > 0. \end{aligned}$$

The above two equations implies that $\mu_F = \mu_S$. But then $V_F = V_S$, which contradicts $V_F - V_S > 0$. Therefore, there is no mixed strategy equilibrium in which O-types randomize in both phases.(iv) Next consider a mixed strategy behavior pattern in which O-types randomize in phase F and play C in phase S. The value functions are:

$$\begin{aligned} V_F^C &= (1-\mu_F)\{a + \delta[\rho V_F + (1-\rho)V_S]\} + \mu_F(-l + \delta V_S), \qquad (60) \\ V_F^D &= (1-\mu_F)b + \delta V_S, \\ V_S &= (1-\mu_S)\{a + \delta[\rho V_F + (1-\rho)V_S]\} + \mu_S(-l + \delta V_S), \\ V_S^d &= (1-\mu_S)b + \delta V_S. \end{aligned}$$

The relevant incentive constraints are:

$$\begin{aligned} V_F^H &= V_F^C \Leftrightarrow V_F - V_S = \frac{b-a}{\delta\rho} + \frac{\mu_F}{1-\mu_F}\frac{l}{\delta\rho} > 0, \\ V_S &> V_S^d \Leftrightarrow V_F - V_S \geq \frac{b-a}{\delta\rho} + \frac{\mu_S}{1-\mu_S}\frac{l}{\delta\rho}. \end{aligned}$$

The above two equations implies that $\mu_F \geq \mu_S$. But then $V_S \geq V_F$, which contradicts $V_F - V_S > 0$. Therefore, this type of equilibrium does not exist.

(v) Finally consider a behavior pattern in which O-types randomize in phase S and play D in phase F. Consider an O-type and let μ_F (μ_S) be the probability that his partner plays D in phase F (S). Then the value functions corresponding to this behavior pattern are

$$\begin{aligned} V_F &= (1-\mu_F)b + \delta V_S, \qquad (61) \\ V_F^d &= (1-\mu_F)\{a + \delta[\rho V_F + (1-\rho)V_S]\} + \mu_F(-l + \delta V_S), \\ V_S^H &= (1-\mu_S)\{a + \delta[\rho V_F + (1-\rho)V_S]\} + \mu_S(-l + \delta V_S), \\ V_S^C &= (1-\mu_S)b + \delta V_S. \end{aligned}$$

The relevant incentive constraints are:

$$\begin{aligned} V_F &\geq V_F^d \Leftrightarrow V_F - V_S \leq \frac{b-a}{\delta\rho} + \frac{\mu_F}{1-\mu_F}\frac{l}{\delta\rho}, \\ V_S^H &= V_S^C \Leftrightarrow V_F - V_S = \frac{b-a}{\delta\rho} + \frac{\mu_S}{1-\mu_S}\frac{l}{\delta\rho} > 0. \end{aligned}$$

From the value functions we also get $V_F - V_S = (\mu_S - \mu_F)b$. So a necessary condition for this behavior pattern to be an equilibrium is that $\mu_S > \mu_F$. But then $\frac{\mu_S}{1-\mu_S} > \frac{\mu_F}{1-\mu_F}$, which implies that

$$V_F - V_S = \frac{b-a}{\delta\rho} + \frac{\mu_S}{1-\mu_S}\frac{l}{\delta\rho} > \frac{b-a}{\delta\rho} + \frac{\mu_F}{1-\mu_F}\frac{l}{\delta\rho},$$

which contradicts $V_F \geq V_F^d$. ■

Proof of Lemma 6

Proof. (i) In stationary state, by abusing notation (both z and x are functions of β),

$$\begin{aligned} (1-\rho)(1-\beta-z) = &\quad x\rho\frac{x}{z+\beta} \\ \Leftrightarrow &\quad \frac{1-\rho}{\rho}(\frac{1}{x} - \frac{\beta+z}{x}) = \frac{x}{z+\beta} \end{aligned} \tag{62}$$

But we know that, when $\beta \leq \underline{\beta}$, in the mixed-strategy equilibrium $\frac{\beta+z-x}{x} = \underline{y}$ is independent of β. Therefore, from (62) x is also independent of β. As a result, $\beta + z$ is also independent of β. Since $W(\beta)$ is only a function of x and $z + \beta$ (see equation (33), we reach the conclusion that $W(\beta)$ is constant when $\beta \leq \underline{\beta}$.

(ii) First we show that $1-x-\beta$, which is the measure of agents in phase F is strictly decreasing in β. Suppose not, that is, suppose there exist a β' and a β'' in $[\underline{\beta}, \overline{\beta}]$ so that $\beta' < \beta''$ and yet $1 - x' - \beta' \leq 1 - x'' - \beta''$, where x' (x'') is the steady state x under β' (β''). Then from (1)

$$x'\frac{x'}{x'+\beta'} \leq x''\frac{x''}{x''+\beta''},$$

which is equivalent to

$$x'\frac{1}{1+\beta'/x'} \leq x''\frac{1}{1+\beta''/x''}.$$

But $\beta'/x' < \beta''/x''$ since y is increasing in β. Therefore, we must have $x' < x''$, which implies

$$1 - x'' - \beta'' < 1 - x' - \beta',$$

a contradiction.

Lemma **??** shows that $\frac{x}{x+\beta}$, the probability of being matched with a non-bad type in phase S, is decreasing in β. From expression (33) we see that by increasing β, the average payoff in phase S decreases and the weight placed on this payoff increases. Hence the total social welfare in the community must decrease. ∎

Proof of Lemma 7

Proof. (i) From (33) and (34), we can see that when $\gamma \leq \underline{\gamma}$ or $\overline{\gamma} \leq \gamma$, $W(\gamma) = W(1-\beta)$, $\underline{\beta} \leq \beta \leq \overline{\beta}$. Directly applying part (ii) of Lemma (6), we reach the conclusion that $W(\gamma)$ is increasing in γ.

(ii) Similar to the proof of part (i) of Lemma (6), it can be shown that in the mixed strategy equilibrium replicating $\underline{y}$, both z and x are independent of γ. Thus by (34) $W(\gamma)$ is constant in γ if $\underline{\gamma} < \gamma < \overline{\gamma}$. ∎

Proof of parts (i)-(iii) of Lemma 8

Proof. (i) An O-type has the option of playing D independent of her personal history, in which case she realizes the same payoff as a B-type. Hence, $0 \leq \pi^O(\beta) - \pi^B(\beta) = \Delta(\beta)$.

(ii) If $\beta \in (\overline{\beta}, 1]$, the unique steady-state equilibrium is the bad one. Therefore, $\Delta(\beta) = 0$.

(iii) If $\beta \in [0, \underline{\beta}]$, the mixed-strategy equilibrium replicating $\underline{y}$ features

$$\pi^B(\beta) = \frac{1}{1-\delta}\frac{x(\beta)}{z(\beta)+\beta}b = \frac{1}{1-\delta}\frac{b}{1+\underline{y}},$$

which is independent of β. In addition,

$$\pi^O(\beta) = \frac{z-\beta}{1-\beta}V_S(\beta) + \frac{1-z}{1-\beta}V_F(\beta).$$

From the analysis in Section 5 we know that both $V_S(\beta) = \pi^B(\beta)$ and $V_F(\beta)$ are independent of β (which follows from the fact that aggregate behavior is independent of β), and $V_S(\beta) < V_F(\beta)$. From the same analysis, we also know that $\beta + z$ is constant in β and, thus, that z is decreasing in β. But then $\frac{z-\beta}{1-\beta}$ is decreasing in β and $\frac{1-z}{1-\beta}$ is increasing in β. Putting these facts together, we conclude that the weighted average $\frac{z-\beta}{1-\beta}V_S(\beta) + \frac{1-z}{1-\beta}V_F(\beta)$ is increasing in β, which implies $\Delta(\beta) = \pi^O(\beta) - \pi^B(\beta)$ is increasing too. Finally, when $\beta = 0$, $V_S(0) = V^B(0)$. So, since $0 = V_S(0) < V_F(0)$ and $\frac{1-z}{1-\beta}$ is positive, we have $0 < \Delta(0)$. ∎

Proof of part (iv) of Lemma 8

Proof. We first show (a) the hump shapedness of Δ, then we show (b) it increases at $\underline{\beta}$.

(a) Since Δ is evaluated at the good equilibrium, we have

$$\begin{aligned}\Delta(\beta) &= \frac{x}{1-\beta}V_S(\beta) + (1-\frac{x}{1-\beta})V_F(\beta) - \pi^B(\beta) \qquad (63)\\ &= V_S(\beta) - \pi^B(\beta) + (1-\frac{x}{1-\beta})[V_F(\beta) - V_S(\beta)]\\ &= \frac{1}{1-\delta}[\frac{xa - \beta(1-\delta\rho)l}{x+\beta(1-\delta\rho)} - \frac{x}{x+\beta}b] + (1-\frac{x}{1-\beta})\frac{\beta(a+l)}{x+\beta(1-\delta\rho)},\end{aligned}$$

where x comes from (1) and V_S and V_F are given by (6) and (7). Using the variable $y \equiv \beta/x(\beta)$, (1) tells us that

$$\frac{x}{1-\beta} = \frac{(1+y)(1-\rho)}{\rho+(1+y)(1-\rho)}.$$

Substituting this into (63), we get

$$\Delta(y) = \frac{1}{1-\delta}[\frac{a - y(1-\delta\rho)l}{1+y(1-\delta\rho)} - \frac{1}{1+y}b] + \frac{\rho}{\rho+(1+y)(1-\rho)}\frac{y(a+l)}{1+y(1-\delta\rho)}. \qquad (64)$$

Differentiating (64) and doing some algebra, we get:

$$\begin{aligned}\Delta'(y) &= \frac{1}{1-\delta}[\frac{-(1-\delta\rho)(a+l)}{(1+y(1-\delta\rho))^2} + \frac{b}{(1+y)^2}] + \frac{\rho}{\rho+(1+y)(1-\rho)}\frac{(a+l)}{(1+y(1-\delta\rho))^2}\\ &\quad -\frac{\rho(1-\rho)}{(\rho+(1+y)(1-\rho))^2}\frac{y(a+l)}{1+y(1-\delta\rho)}\\ &= \frac{1}{(1-\delta)[1+y(1-\delta\rho)]^2} \times\\ &\quad \times \left\{ \frac{[b-(1-\delta\rho)(a+l)] + 2y(1-\delta\rho)[b-(a+l)] + (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y^2}{(1+y)^2} \right.\\ &\quad \left. + \frac{\rho[1-(1-\rho)(1-\delta\rho)y^2]}{(\rho+(1+y)(1-\rho))^2}(a+l) \right\}. \qquad (65)\end{aligned}$$

We are going to show now that there is a $0 \leq \widehat{y}$ so that $\Delta'(y)$ is positive for $0 < y < \widehat{y}$ and negative for $\widehat{y} < y$, which implies that Δ has the desired hump shape property (if $\widehat{y} = 0$, Δ is increasing throughout). Since $0 < \frac{1}{(1-\delta)[1+y(1-\delta\rho)]^2}$, it suffices to show this for the term inside the braces, which we abbreviate as

$$\varphi(y) = \frac{f_1(y)}{g_1(y)} + \frac{f_2(y)}{g_2(y)}.$$

Inspecting the two terms of φ we see that: (1) The denominator of each term is positive and increasing in y. (2) Each numerator is quadratic and, because $a+l \leq b$, it decreases in y and tends to $-\infty$ as $y \to \infty$. From these observations we infer that there are two points $0 \leq y_1$ and $0 < y_2$

so that the first term is positive for $y < y_1$ and negative for $y_1 < y$, and similarly for the second term. In addition, one readily verifies that $y_1 < y_2$, so that φ is positive for $[0, y_1]$ and negative for $[y_2, \infty)$.

It remains to analyze the behavior of φ over (y_1, y_2). By continuity, there exists a $\widehat{y} \in (y_1, y_2)$ so that $\varphi(\widehat{y}) = 0$. To show that $\widehat{y}$ is unique, which would bring the proof to a conclusion, it suffices to prove that $\varphi'(\widehat{y}) < 0$.

Since $y_1 < y_2$, we know that $\frac{f_1(\widehat{y})}{g_1(\widehat{y})} < 0 < \frac{f_2(\widehat{y})}{g_2(\widehat{y})}$. This implies $\left(\frac{f_2(y)}{g_2(y)}\right)' |_{y=\widehat{y}} < 0$, so it suffices to show that $\left(\frac{f_1(y)}{g_1(y)}\right)' |_{y=\widehat{y}} < 0$. Now,

$$\left(\frac{f_1(y)}{g_1(y)}\right)' = \frac{f_1' g_1 - f_1 g_1'}{g_1^2} < 0 \Longleftrightarrow f_1' g_1 < f_1 g_1'.$$

Substituting in for f_1 and g_1, leaves us with the following inequality to prove:

$$\begin{aligned} & (1+y)^2 \{2(1-\delta\rho)[b-(a+l)] + 2(1-\delta\rho)[(1-\delta\rho)b-(a+l)]y\} \\ < \; & 2(1+y)\left\{[b-(1-\delta\rho)(a+l)] + 2y(1-\delta\rho)[b-(a+l)] + (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y^2\right\}. \end{aligned}$$

Dividing both sides of this inequality by $2(1+y)$, we need to show that:

$$\begin{aligned} & [b-(1-\delta\rho)(a+l)] + 2y(1-\delta\rho)[b-(a+l)] + (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y^2 \\ > \; & (1+y)\{(1-\delta\rho)[b-(a+l)] + (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y\} \\ = \; & (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y^2 + (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y + (1-\delta\rho)[b-(a+l)]y \\ & +(1-\delta\rho)[b-(a+l)] \\ = \; & (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y^2 + (1-\delta\rho)[(2-\delta\rho)b-2(a+l)]y + (1-\delta\rho)[b-(a+l)]. \end{aligned}$$

Looking at the two ends of this inequality, and comparing term by term establishes that this inequality holds.

(b) Consider the two terms of (65), evaluated at $\underline{y}$. The first term is equivalent to $\frac{d(V_S(y)-\pi^B(y))}{dy}$, which is positive at $\underline{y}$ because $V_S(\underline{y})-\pi^B(\underline{y}) = 0$ and $0 < V_S(y)-\pi^B(y)$ for all $y \in (\underline{y}, \overline{y})$. Also, since $y_1 < y_2$, we have that the numerator of the second term of (65) is positive. Since the denominator of the second term is always positive, this term is positive as well, so altogether $0 < \Delta'(\underline{y})$. Finally, since β and y are monotonically related, this implies $0 < \Delta'(\underline{\beta})$. ■