Long-Term Relationships as Safeguards

Rafael Rob* and Huanxing Yang†

June 15, 2005

Abstract

We analyze a repeated prisoners' dilemma game played in a community setting with heterogeneous types. Some players are bad types, programmed to defect, others are good types, programmed to cooperate, and others yet choose actions to maximize their discounted payoffs. Players are also able to strategically choose whether to continue interacting with the same partner - form a long term relationship - or separate and seek a new partner. We show that the ability to endogenously form long term relationships facilitates the achievement of cooperative outcomes without information flows, without instability due to observational errors, and without a central coordinating device to synchronize players' actions. We also show that the heterogeneity of types helps, rather than hinders, cooperative behavior by inducing players to avoid bad types that inflict low payoffs on them and seek good (or opportunistic) types that bestow high payoffs.

JEL Classification numbers: C73, C78, D82.

Keywords: community games, information flows, heterogeneity of types, long term relationships, investment in human capital.

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104. Corresponding author. Email: rrob@econ.sas.upenn.edu. 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. The theory of repeated games has shown how inefficiencies that naturally arise in a game setting (as opposed to a competitive setting) can be rectified if the game is repeated, and if players are able to condition their present behavior on past behavior. This idea has been shown to work under a variety of circumstances, including the possibility that the game is played in a community setting, where players interact with varying opponents and, hence, where quick and personal retaliation are not feasible. This community setting scenario is explored in papers by Kandori (1992) and Ellison (1994), who find that although it is possible to sustain efficiency, the informational requirements needed to do so are not light. In particular, an individual needs to have information about behavior in interactions in which she did not participate. Alternatively, she may need to be able to synchronize her behavior with others' behavior by perfectly observing a public coordination device, or, else, the stability of efficient behavior cannot be guaranteed.

This paper continues this line of research by studying the role that the endogenous formation of long-term relationships may play in sustaining efficient behavior. Unlike previous analyses we consider the situation where the choice of a partner to interact with is (partly) endogenous, and hence where interaction is neither perfectly anonymous (as in the community setting) nor perfectly intimate (as in the traditional repeated game setting). Instead, a player, in addition to choosing her action, also chooses whether to keep interacting with her present partner, or seek a new partner. Another departure of our setting from previous analyses is that we accommodate the heterogeneity of types and, in particular, the presence of commitment types that are programmed to play specific actions.

What we find in this setting is that the community may be able to enforce efficient behavior, and make this behavior stable, while relying on minimal informational requirements. In particular, a player needs to know only what her present partner did in interacting with her, and not what the partner (or anyone else) did in interacting with others in the distant past. Furthermore, an individual does not have to synchronize her behavior with others' behavior by perfectly observing the outcome of a public coordination device. In this sense, what we show is that endogenously forming long-term relationships and keeping track of information that is internal to such relationships is sufficient to achieve efficient outcomes.

Another finding that emerges from our analysis is that heterogeneity of types helps, rather than hinders, sustain efficient outcomes. In essence, the fact that players have control over who to interact with and that different players are of different types, implies that players avoid interacting with types whose behavior might harm them, and seek interacting with types whose behavior might benefit them. If this objective (namely the search for a beneficial partner) is accomplished by cooperating with one's partner, then heterogeneity helps achieve cooperation and thereby raise players' payoffs.

In somewhat greater detail, we study a repeated prisoners' dilemma game in the context of a community with a continuum of agents. Each agent in the community is one of three types: either bad, which means she defects unconditionally (i.e., independent of her personal history), or good, which means she cooperates unconditionally, or she is an opportunist who chooses actions to maximize discounted payoffs. Players in the community are matched in pairs to play a prisoners' dilemma game in each period. An agent learns her opponent's action, and may choose to stay in a relationship with this opponent in the next period, or separate and be matched with another agent. We focus on a class of equilibria in this setting in which strategies are particularly simple: Strategies are such that an individual immediately separates from her partner if she encounters uncooperative behavior. In addition, an individual's choice of action is only conditioned on whether she is about to interact with her partner for the first time, or whether she has already interacted with him in the past.

Given this game and the class of strategies we focus on, our aim is to determine equilibrium behavior at each of the above two contingencies. More specifically, for any configuration of parameter values, we determine whether there is a pure and/or a mixed-strategy equilibrium. In doing so we link parameters values (payoffs in the payoff matrix, the discount factor, the rate of turnover in the community, and the configuration of types in the community) to behavior that is manifested in equilibrium. This link enables us then to establish comparative static properties of the equilibria.

Armed with these results we are able to be more precise about the intuitions we suggested earlier. For example we are able to show that the proportion of bad types must exceed some critical value (and must be no bigger than another critical value) to induce all opportunists to cooperate in equilibrium. We are also able to show that if the proportion of good types exceeds some critical value the dismal equilibrium in which players unconditionally defect no longer exists.

This contrast with standard results of the theory of repeated games, whereby the dismal equilibrium is the "easiest" one to construct. We offer a closed-form characterization of these critical values, and offer additional intuitions about these results in the body of the paper.

In an extension of the model we endogenize the configuration of types in the community by letting individuals invest in human capital, which expands the range of actions available to them (converting them from bad to opportunistic types). This extension enables us to study the interplay between investment in human capital and cooperative behavior, showing that more educated populace is positively correlated with more civil (or cooperative) behavior in the community. The extension also enables us to do welfare exercises, contrasting the equilibrium with a planner's optimum. This comparison identifies two kinds of departures between equilibrium and optimum. In one departure individuals under-invest because the fruits of their investments are partially enjoyed by others. In another departure individuals over-invest because of a conflict between ex-ante and ex-post incentives: On the one hand, it pays individuals to invest ex-ante to be "eligible" for the benefits of long-term relationships; on the other hand, having invested, an individual may defect because of a short-term gain. Because of that, some of the ex-ante investments are not utilized ex-post, which implies they are wasted from a social point of view.

Although this paper is intended as a theoretical exploration, anecdotal evidence suggests that the forces we identify here are of empirical relevance. One anecdote suggesting this comes from the banking industry and, in particular, the practice of "customer relationships." Roughly speaking, this practice is such that established customers, who pay back their loans on time, are able to enter into (or sustain) long-term relationships, and borrow at a lower interest rate or borrow a larger amount. On the other hand, new customers may have to pay a higher interest rate or borrow a smaller amount, and customers who are not current on their loans are denied credit and may have to turn to other institutions for future business, and pay a higher interest rate. Thereby, the promise of forming a long-term relationship and enjoying favorable terms, and the punishment of severing a relationship, having to start from scratch, and suffering unfavorable terms induces borrowers to behave honestly. Other examples in the same spirit 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 Apart from the community setting papers that we already mentioned, there is a small literature on "building trust" that we base our formulation on. This literature started in a little known paper by Dutta (1993), in which he shows that playing more and more cooperative actions over time is a way to gradually achieve efficient outcomes. This idea is significantly extended in Ghosh and Ray (1996) who incorporate (impatient) types into their framework, and refine the set of equilibria that arise based on the criterion of renegotiation proofness. Compared to those papers, this paper makes three contributions. First the component game we analyze is a standard prisoners' dilemma game with two actions and, therefore, with a limited scope for trust building and gradual convergence to cooperation. Instead, our focus is on the incentivating role that the heterogeneity of types plays. The second contribution is that we consider a richer framework with good types as well as bad and opportunistic types, and explore a wider class of equilibria. In doing so, we provide a full characterization of the set of pure and mixed-strategy equilibria, relate them to underlying parameters, and do comparative statics exercises. The third contribution is that we extend the model to study investment in human capital, how it interacts with cooperation, and what its welfare properties are.

Another paper that relates to our theme is Sobel (2002). He focuses, however, on the role of legal rules and does not deal with the heterogeneity of types. A different approach is taken by Tirole (1996) and Dixit (2003) who study community games appended with information intermediaries that make information available to players, which facilitates efficiency. Somewhat more tangential to our theme (although still relevant) are papers by Eeckhout (2002), Lindsey (2002) *et al.*, and Watson (2002).

Preview. The plan of the paper is as follows. The next section introduces our framework. In Section 3 we determine when the pure-strategy good equilibrium, in which opportunists 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 opportunists always 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 human capital, how it interacts with cooperative behavior, and what departures may exist between the equilibrium of this investment game and the social optimum. Some proofs are found in a technical appendix, while others are found in a working

paper version.

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), and each pair of partners play a two-stage game. In the first stage they play 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 first-stage game is specified momentarily.

After this stage, each partnership persists with probability ρ, and breaks with probability $1-\rho$. If a partnership persists, the two partners go into a simultaneous-move second-stage 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 breaks, both partners go into a pool of unmatched players. Players in this pool are randomly matched at the beginning of the next period, forming new partnerships. Consequently, the dissolution and re-formation of partnerships are partly exogenous and partly endogenous. No direct payoffs are associated with the second-stage game; its only role is to endogenize the decision whether to interact with the same individual in 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 first-stage 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.[1]

[1] A G-type player is either a "commitment type" (perhaps inherently moral), or has a payoff matrix that is

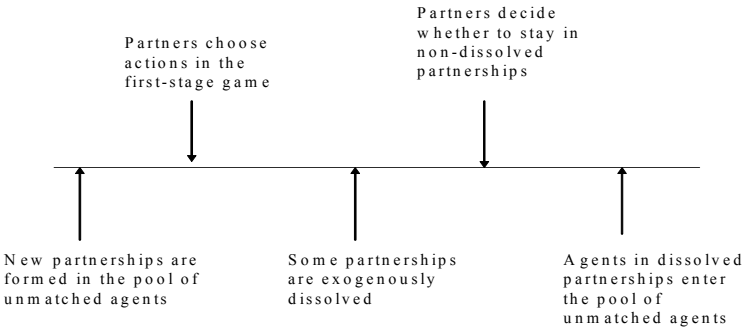

Figure 1: Time Line

	C	D
C	a	$-l$
D	b	0

Table 1: Payoff matrix of an O-type

We assume $b > a > 0$, $l > 0$, and $2a > b - l$. The first two restrictions say that this game, when played by two O-types, is a standard 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.[2] The discount factor is common to all players and is denoted by δ, where $\delta \in (0, 1)$.

obtained from Table 1 by subtracting a large number from the D row. Similarly, a B-type is either a commitment type (inherently immoral), or a large number is subtracted from the C row. The subtracted number is bigger than a, so that a B-type's payoff is negative at (C, C). Because of that, playing unconditional D is a dominant strategy for a B-type not only in the period game but also in the repeated game.

[2]The fact that G-types always choose C yet they are assumed to maximize their discounted payoff is reconciled as follows. In the first-stage game they choose C because C is a dominant action (in the period game) for G-types. In the second stage they may separate because they foresee a higher payoff from being matched with types who play C. Therefore, maximizing behavior is manifested in both stages.

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. We also assume that the **configuration of types**, (α, β, γ), are common knowledge.

Steady-state equilibria This is an infinitely repeated community game with incomplete information, so folk-theorem type arguments establish that there are many equilibria supported by a variety of repeated-game strategies. For example, Kandori's (1992) "contagious equilibrium," in which each player plays D forever if either he or one of his previous partners played D, is an equilibrium in our setting.

Rather than prove folk theorems, this paper focuses on a certain class of equilibria. This class is defined by two properties that strategies are required to satisfy, along with a specification of certain "initial conditions." To state these properties we first define the concept of a phase. A player is said to be in the **stranger** phase, denoted S, if he never interacted with his current partner (i.e., if he just entered into a new partnership). On the other hand, a player is said to be in the **friendly** phase, denoted F, if he interacted at least once with his current partner.[3] The first property that strategies are required to satisfy is that a player's action, or mixed strategy, in the first-stage game only depends on which of these two phases he is in, and on no other aspect of his personal history (this requirement rules out the contagious equilibrium). If this requirement is satisfied, we call the mapping from phases to actions in the first-stage game a **behavior pattern**. The second property that strategies are required to satisfy is that a player's action in the second-stage game is to separate if, and only if, at least one partner defected (in the first-stage game) at any point since the start of the partnership. These two properties, along with initial conditions (regarding at which phase each player is in initially), determine the measure of types in each phase at each point in time. The third requirement we impose is that these measures have settled to a steady state (at $t = 0$),

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

and, as such, remain constant through time. We refer to equilibria with these three properties as **steady-state equilibria**.

A few words are in order to explain why we focus on steady-state equilibria. The first (and obvious) reason is tractability. Indeed, as will be seen, we are able to fully characterize pure and mixed-strategy equilibria in this class, relate them - via closed-form expressions - to underlying parameters, and do comparative statics exercises. Arguably, one can expand our class of equilibria, while preserving tractability. For example, one may study strategies in which the *degree* of cooperation depends on the length of the relationship, i.e., where a player chooses C with a higher probability in a relationship that have lasted for a longer time. We found, however, that straightforward generalizations of this sort do not lead to new insights.[4] Another important property of steady-state equilibria is that learning about one's partner's type (and thus behavior) does not occur beyond the first period of a relationship. This is true because the distribution of types within the friendly phase is independent of the length of the relationship, and because perfect monitoring reveals - in the first period of interaction - all the information about one's partner's type that is ever going to be revealed. Therefore, one may view the class of equilibria we study as those for which behavior does not vary, if information does not vary.

Another (and perhaps more substantive) reason for focusing on steady-state equilibria is that they capture behavior that seems "realistic." Ordinarily (i.e., outside of the game-theory community), such behavior is explained using psychology or using emotionally charged language. For example, it would ordinarily seem that defection has the effect of "souring a relationship," triggering separation, and initiating a new relationship. But this is exactly what the separation strategy we focus on specifies. Likewise, it would seem that players view a new relationship as an opportunity for a "fresh start," and consequently would not let their past experience affect it. But this, again, is what a behavior pattern in our framework specifies. An important feature of our analysis is that there is no need to resort to explanations that are outside the purview of economics. Instead, the behavior we study is equilibrium behavior, so one may view it as purely driven by economic incentives and equilibrium reasoning.[5]

[4] These generalizations lead to the idea of gradual trust building, which has been explored by Dutta and Ghosh and Ray. Here, on the other hand, we are more interested in the role of heterogeneity.

[5] Non-economic explanations may still be "significant" or "popular," but as economists we focus on economic

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 first-stage game. In addition, we already specified that all player types separate in the second-stage 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). Given this, the only aspect of behavior that remains to be endogenously determined is the behavior-pattern of O-types in the first-stage game. This will be the focus of the analysis in the next sections.[6]

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 previously described separation strategy, induce a steady-state. The first step in the analysis is to determine this steady-state, i.e., determine the overall measure of agents in phase S, and its composition. To do that, 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 types in phase S is $x + \beta$, and the overall measure of types in phase F is $1 - x - \beta$. In the steady-state of the good equilibrium x must satisfy

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

explanations.

[6] In much of this analysis we hold other parameters constant, and determine the behavior pattern as a function of the configuration of types, (α, β, γ), only. Since $\alpha + \beta + \gamma = 1$, this is equivalent to the determination of behavior in terms of β and γ.

To interpret (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). Such solution to (1) is unique and, 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 $x(\beta)$. Given $x(\beta)$ and β we define the variable $y(\beta) \equiv \beta/x(\beta)$, or simply y, 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 agents choosing D to those choosing C (more precisely, the ratio of the *measures*). We next state a simple, but important, 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 of a one-shot deviation when in phase S. The equations defining these values are:

$$V_F = a + \delta[\rho V_F + (1-\rho)V_S] \tag{2}$$

$$V_S = \frac{x}{x+\beta}V_F + \frac{\beta}{x+\beta}(-l + \delta V_S) \tag{3}$$

$$V_F^d = b + \delta V_S \tag{4}$$

$$V_S^d = \frac{x}{x+\beta}(b + \delta V_S) + \frac{\beta}{x+\beta}(0 + \delta V_S). \tag{5}$$

As a representative of the logic on which these equations rest, consider the RHS of (2), which is the discounted payoff of an O-type at 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. All other value functions are based on a similar logic.

Equations (2) and (3) represent two linear equations in two unknowns, V_F and V_S, so one can explicitly solve them, i.e., express V_F and V_S in terms of model primitives. Doing so we get

$$V_F = \frac{(x+\beta-\delta\beta)a-\beta\delta(1-\rho)l}{(1-\delta)[x+\beta(1-\delta\rho)]} \tag{6}$$

$$V_S = \frac{xa-\beta(1-\delta\rho)l}{(1-\delta)[x+\beta(1-\delta\rho)]}, \tag{7}$$

where x is the solution to (1).

Incentive Constraints Above we considered the "mechanics" of the good equilibrium, computing the steady-state distribution, 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. The following two incentive constraints must be satisfied:

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

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

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 $l > 0$, 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 how to interpret (10), 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) shows 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.

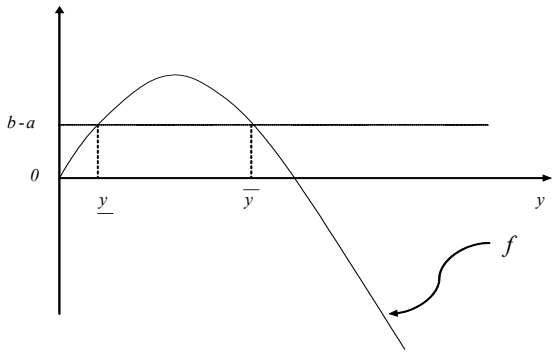

Figure 2: The graph of f

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

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

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 $\underline{y} = \underline{\beta}/x(\underline{\beta})$, and $\overline{\beta}$ is defined by $\overline{y} = \overline{\beta}/x(\overline{\beta})$. 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 $(0, 1)$. Also, 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 parameter values. Summarizing our analysis, we have the following result.

Proposition 1 *(i) The existence of the good equilibrium does not hinge on γ, the measure of good types. (ii) If (12) is not satisfied, then the good equilibrium does not exist. (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 $f(\frac{\beta}{x(\beta)}) = b - a$.*

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 ($\frac{\beta}{\beta+x} > \beta$ 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.[7] Without (a critical mass of) bad types this inducement/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 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[8]) 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.

Stability Having commented on the structure of incentives at the good equilibrium, let us now comment on its "resilience," and on how the good equilibrium compares in this regard to the contagious equilibrium à la Kandori (1992).

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 with probability $\varepsilon > 0$, even though the partner actually chose C. 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

[7] Another way of saying this is that the reward to playing C is that one eventually insulates oneself from bad types.

[8] 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.

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 another resolution to this non-resilience problem - within the context of the contagious equilibrium. 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. In our view, however, reliance on such device is a bit far fetched from a practical point of view. Indeed, it is hard to visualize a whole community relying on a central device to synchronize everyone's behavior. On the other hand, severing relationships, starting new ones, and observing behavior only within a relationship is a way to decentralize this outcome, which seems simpler and more realistic.

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 a good equilibrium is non-monotonic: When β is small the effect is positive (an increase in β widens the set of circumstances under which the good equilibrium exists), but when β is large the effect is negative.

Other comparative statics results are similarly derived. For instance, the effect of increasing δ is positive, i.e., (10) is satisfied under a wider set of circumstances. This mirrors conventional wisdom conveyed by folk theorems. On the other hand, 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* $\frac{a+l}{1-\delta} > b > (1-\delta)(a+l)$,[9] *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 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 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. The reason for this is that turnover introduces "fluidity" into the system, enabling movements from phase S to phase F and, thereby, generating incentives to play C.[10]

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 approach to other steady-state equilibria. Our analysis here expands the analysis in Section 3 in the sense that we unravel the structure of incentives at these other equilibria, and pin down the conditions under which they exist. More broadly, our

[9] This assumption is satisfied if δ is large enough or if $b = a + l$, which is the condition that the component game is a partnership game.

[10] 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 $\rho = 1$.

analysis here makes two points. The first point is that when the good equilibrium 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 second 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 C in phase S. Given the separation strategy, 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

$$(1-\rho)(\gamma - x) = x\rho\frac{x}{x+1-\gamma}. \tag{13}$$

Analogous to (1), the solution to (13) determines x as a function of γ. We let the ratio of non-good types to good type in phase S be $y \equiv \frac{1-\gamma}{x(\gamma)}$, which, as before, is also the ratio of agents choosing D to 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, which generates a different steady-state and different period payoffs. Making the requisite adjustments, the new value functions are:

$$V_F = b + \delta V_S \tag{14}$$

$$V_S = \frac{x}{x+1-\gamma} b + \delta V_S \tag{15}$$

$$V_F^d = a + \delta[\rho V_F + (1-\rho) V_S] \tag{16}$$

$$V_S^d = \frac{x}{x+1-\gamma}\{a + \delta[\rho V_F + (1-\rho) V_S]\} + \frac{1-\gamma}{x+1-\gamma}(-l + \delta V_S). \tag{17}$$

Given these value functions, the incentive constraints are:

$$\text{No deviation in phase } F \quad : \quad V_F - V_F^d \geq 0. \tag{18}$$

$$\text{No deviation in phase } S \quad : \quad V_S - V_S^d \geq 0. \tag{19}$$

Analyzing these constraints, we have the following result.

Lemma 3 *(i) (19) is redundant if (18) is satisfied. (ii) A bad equilibrium in which O-types defect in phase F exists if, and only if,*

$$b - a \geq \frac{1-\gamma}{x+1-\gamma} \delta \rho b. \tag{20}$$

Proof. See the working paper. ■

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 working paper version).

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

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

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's why two inequalities (rather than one) have to be satisfied in condition (21).

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

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

Existence of the bad equilibrium As we did with the good equilibrium, we transform condition (22) to a condition that involves only the primitive data. To this end we re-write the RHS of (22) in terms of y, giving us:

$$b-a \geq \frac{y}{1+y}\delta\rho b - y(1-\delta\rho)l \equiv f(y). \tag{23}$$

As can be readily seen, (23) 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 *(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, $\gamma \in [0,\underline{\gamma}] \cup [\overline{\gamma},1]$ (equivalently if $\gamma \notin (\underline{\gamma},\overline{\gamma})$), where $\underline{\gamma}$ and $\overline{\gamma}$ are found by solving $f(\frac{1-\gamma}{x(1-\gamma)}) = b-a$, 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 easiest 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 higher 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 F 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. Thus, the bad equilibrium does not exist when γ is in this range.[11]

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 $\beta = 1 - \gamma$. Also, Propositions 1 and 3 show that the presence of bad types can support the good equilibrium, while the presence of good types cannot. Analogously, the presence of good types can upset the bad equilibrium, while the presence of bad types cannot.

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 several types of mixed behavior patterns to consider. As we show in the working paper version, however, several of these behavior patterns do not give rise to equilibria, or give rise to equilibria that are behavior- and, hence,

[11] Another implication of Proposition 3 (analogous to Proposition 1) is that the presence of bad types has no bearing on the existence of the bad equilibrium. The reason for this, again, is that bad types and opportunistic types behave the same way in the bad equilibrium, so the incentive to deviate depends only on their overall measure or, equivalently, on the measure of good types

payoff-equivalent to equilibria we already considered. The only mixed behavior pattern that is not like this is where O-types mix in phase S and play C in phase F. Consequently, we focus now on this behavior pattern, 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.

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

$$(1-\rho)(\alpha - x_\alpha) = (1-\lambda)x_\alpha \rho \frac{(1-\lambda)x_\alpha + x_\gamma}{x_\alpha + x_\gamma + \beta} \tag{24}$$

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

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 agents playing D to agents playing C in phase S.[12]

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

$$V_F = a + \delta[\rho V_F + (1-\rho)V_S^C] \tag{26}$$

$$V_S^C = \frac{x}{z+\beta}V_F + \frac{z+\beta-x}{z+\beta}(-l + \delta V_S^C) \tag{27}$$

$$V_F^d = b + \delta V_S^C$$

$$V_S^D = \frac{x}{z+\beta}(b + \delta V_S^C) + \frac{z+\beta-x}{z+\beta}(0 + \delta V_S^C),$$

where the superscripts on V_S refer now to (candidate) equilibrium behavior, rather than to deviation from such behavior (while the superscript 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 satisfies. After some manipulations, we simplify these constraints as

[12] $\beta + z$ 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.

follows.

$$\text{No-deviation in phase } F\text{: } 0 \leq V_F - V_F^d \Leftrightarrow \frac{b-a}{\delta\rho} \leq V_F - V_S. \tag{28}$$

$$\text{Indifference in phase } S\text{: } V_S^D = V_S^C \Leftrightarrow V_F - V_S = \frac{b-a}{\delta\rho} + \frac{(z+\beta-x)l}{\delta\rho x}. \tag{29}$$

Since the RHS of (29) exceeds the RHS of (28), it suffices to require (29), which we re-write (after solving for V_F and V_S) as:

$$\frac{xa-(1-\delta\rho)(z+\beta-x)l}{(z+\beta)(1-\delta\rho)+\delta\rho x} = \frac{xb}{z+\beta}. \tag{30}$$

As before, we let $y \equiv \frac{\beta+z-x}{x}$ be the ratio of agents playing D to agents playing C in phase S. Since it plays an important role, the dependence of y on λ is made explicit here, $y(\lambda)$ (y continues to depend on β and γ, of course). Using the definition of y, equation (30) is re-written as

$$b-a = \frac{y}{1+y}\delta\rho b - y(1-\delta\rho)l \equiv f(y). \tag{31}$$

Existence of mixed-strategy equilibria We note that (31) 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 (31). From the discussion in Section 3 we know that if (12) is not satisfied, there are no roots to equation (31) and, hence, no mixed-strategy equilibria. Therefore, to proceed, we assume that (12) is satisfied.

Let us 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, (13). 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.[13] Therefore, as one varies λ 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 **completely** mixed-strategy equilibrium exists if and only if at least one of $\underline{y}$ or $\overline{y}$ is in $(\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$ ("completely" means that $0 < \lambda < 1$). Furthermore, a mixed-strategy (unless we state otherwise mixed means completely mixed) equilibrium is unique if exactly one of $\underline{y}$ or $\overline{y}$ is in $(\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$.

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

To be more precise about the set of circumstances under which a completely 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 completely mixed-strategy equilibrium "replicating" $\underline{y}$. 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, one can find a mixed-strategy equilibrium replicating $\overline{y}$. This gives us a complete characterization of when mixed-strategy equilibria exist as a function of underlying parameters. We summarize the analysis as follows.

Proposition 4 *If (12) is violated, there are no mixed-strategy equilibria. If (12) holds, then: (i) A completely mixed-strategy equilibrium exists if, and only if, there is a $\lambda \in (0, 1)$ so that (24), (25) and (31) are satisfied. (ii) This holds if and only if $\underline{y}$ or $\overline{y} \in (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$, which is equivalent to $(\beta, \gamma) \in [0, \underline{\beta}) \times [0, \overline{\gamma}) \cup [0, \overline{\beta}) \times [0, \underline{\gamma})$. (iii) A mixed-strategy equilibrium is unique if, and only if, exactly one of $\underline{y}$ or $\overline{y}$ is in $(\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$. (iv) In any mixed-strategy equilibrium, the ratio of agents playing D to agents playing C in phase S is either $\underline{y}$ or $\overline{y}$.*

Having shown the set of circumstances under which a mixed-strategy equilibrium can be constructed and how to compute it, let us comment now about 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 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. But such a "cycle" is exactly the set of circumstances under which mixed-strategies that constitute an equilibrium are constructed in general. Specifically, this is done 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 in phase S, opportunists' choices are consistent with each other's, i.e., 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 $\underline{\beta} = \beta + z - x > \beta$. 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 commitment 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 agents playing D to agents playing C in phase S is $\frac{\beta+z-x}{x} = \frac{\underline{\beta}}{x(\underline{\beta})}$. These two variables are independent of the particular value that (β, γ) assumes. 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 commitment 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 commitment types).

6 Classification of Equilibrium Outcomes

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.

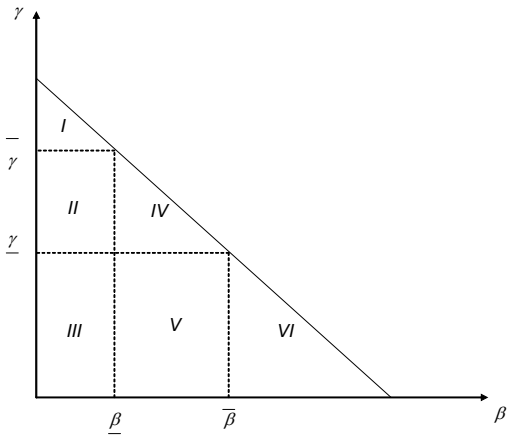

Figure 3: Classification of Equilibrium Outcomes

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 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 (to avoid tedious statements a region is exclusive of its boundaries). In regions I, III, and VI, the bad equilibrium exists, while the good equilibrium does not exist. In region IV, 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 completely 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. To do that, we consider four cases that exhaust the universe of possibilities.

Case 1 *Neither the good nor the bad equilibria exist (region II in the simplex).*

This case corresponds to $\frac{\beta}{x_g} < \underline{y}$ and $\underline{y} < \frac{1-\gamma}{x_b} < \overline{y}$. But then $\underline{y} \in (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$, i.e., there exists a $\lambda \in (0,1)$ so that $\frac{\beta+z-x}{x} = \underline{y}$. At the same time there is no $\lambda \in (0,1)$ so that $\frac{\beta+z-x}{x} = \overline{y}$, i.e., $\overline{y} \notin (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$. Therefore, there is only one mixed-strategy equilibrium in region II, and the phase S ratio of agents choosing D to those choosing C in it is $y = \underline{y}$.

Case 2 *The good equilibrium exists, but the bad equilibrium does not exist (region IV in the simplex).*

This case corresponds to $\underline{y} < \frac{\beta}{x_g}, \frac{1-\gamma}{x_b} < \overline{y}$. But, then, $\underline{y}, \overline{y} \notin (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$, which means there are no mixed-strategy equilibria.

Case 3 *The bad equilibrium exists but the good equilibrium does not exist (Regions I, III and VI in the simplex).*

In this case either $\frac{\beta}{x_g} < \underline{y}$ or $\overline{y} < \frac{\beta}{x_g}$, and either $\frac{1-\gamma}{x_g} < \underline{y}$ or $\overline{y} < \frac{1-\gamma}{x_b}$. Recalling that one must have $\frac{\beta}{x_g} < \frac{1-\gamma}{x_b}$, there are three sub-cases to consider.

(sub-case 3.1) $\frac{\beta}{x_g} < \underline{y}$ and $\frac{1-\gamma}{x_b} < \underline{y}$, which is region VI. Then, $\underline{y}, \overline{y} \notin (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$, so there are no mixed-strategy equilibria.

(sub-case 3.2) $\frac{\beta}{x_g} < \underline{y}$ and $\overline{y} < \frac{1-\gamma}{x_b}$, which is region III. Then, $\underline{y}, \overline{y} \in (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$, implying there are two mixed-strategy equilibria, replicating $\underline{y}$ and $\overline{y}$.

(sub-case 3.3) $\overline{y} < \frac{\beta}{x_g}$ and $\overline{y} < \frac{1-\gamma}{x_b}$, which is region I. Then, $\underline{y}, \overline{y} \notin (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$, so there are again no mixed-strategy equilibria.

Case 4 *Both the good and the bad equilibria exist (region IV in the simplex).*

In this case $\underline{y} < \frac{\beta}{x_g} < \overline{y}$ and $\overline{y} < \frac{1-\gamma}{x_b}$. Thus $\overline{y} \in (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$ and $\underline{y} \notin (\frac{\beta}{x_g}, \frac{1-\gamma}{x_b})$. This means there is a unique mixed-strategy equilibrium replicating $\overline{y}$.

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
Region I	Bad equilibrium	None
Region II	None	One replicating $\underline{y}$
Region III	Bad equilibrium	Two
Region IV	Good equilibrium	None
Region V	Both equilibria	One replicating $\overline{y}$
Region VI	Bad equilibrium	None

In summary, our analysis and Table 2 show 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 means that the good equilibrium exists for a range of β values. To simplify the notation, let $g(\beta) \equiv f(\frac{\beta}{x(\beta)})$. Then, the good equilibrium exists if and only if $g(\beta) \leq 2 = b - a$. The two roots of $g(\beta) = 2$ 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 reports how these numerical results fit into the classification of regions, and provides examples of mixed-strategy equilibria.

Table 3: Numerical Example

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

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 offer 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. 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.

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 III), 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 $\beta > \overline{\beta}$ (region VI), the unique steady-state equilibrium is the bad equilibrium.

Taking these three cases into account, social welfare takes the following form.

$$W(\beta) = \begin{cases} (1-z-\beta)a + (\beta+z-x)\frac{x}{z+\beta}b + x[\frac{x}{z+\beta}a - \frac{\beta+z-x}{z+\beta}l] & \text{if} \quad \beta < \underline{\beta} \\ (1-x-\beta)a + \beta\frac{x}{x+\beta}b + x[\frac{x}{x+\beta}a - \frac{\beta}{x+\beta}l] & \text{if} \quad \underline{\beta} \leq \beta \leq \overline{\beta} \\ 0 & \text{if} \quad \beta > \overline{\beta} \end{cases}, \tag{32}$$

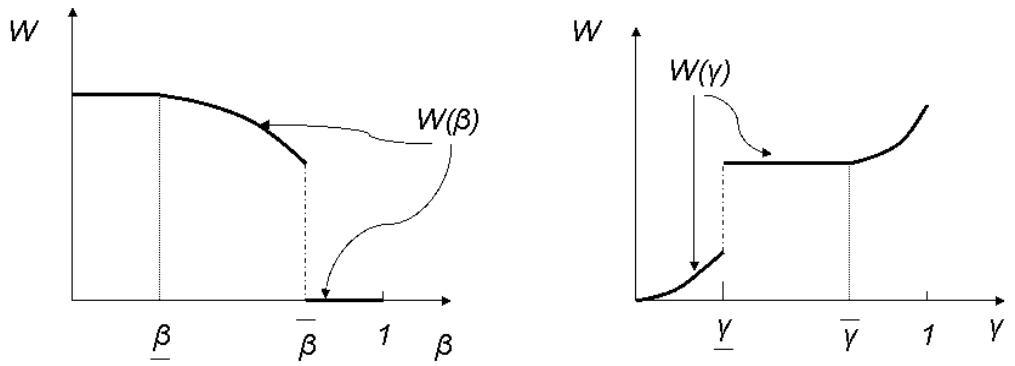

Figure 4: Welfare Measures

where x in the second line comes from the solution to (1), and x and z in the third line come from the solution to (24) and (25).

To elaborate on how (32) is arrived at, consider the middle term, $\underline{\beta} \leq \beta \leq \overline{\beta}$. Then, as stated above, welfare is evaluated at the good equilibrium. Opportunists in this equilibrium get 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 (32) we derive the following result, which is graphically illustrated in the left panel of Figure 4.

Lemma 5 *(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 $\beta > \overline{\beta}$, $W(\beta)$ is zero.*

Proof. See the Appendix. ■

The reason W is zero for $\beta > \overline{\beta}$ 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 level of cooperation and, hence, the average payoff in the community. Finally, the reason welfare is constant over $[0, \underline{\beta}]$ is that welfare (for each 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 $\beta > \overline{\beta}$, 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.

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

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

Analyzing this welfare function we derive the following result, which is proven in the working paper version, and is illustrated in the left panel of Figure 4.

Lemma 6 *(i) When $\gamma \in [0, \underline{\gamma}]$ or $\gamma \in [\overline{\gamma}, 1]$, $W(\gamma)$ 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 the social welfare of 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 Choice of Types

In this section we extend the model to the scenario in which individuals endogenously choose their types by investing in human capital. This extension enables us to address two issues: One is the interplay between investments in human capital and the level of cooperation in the community. The other is the comparison between the equilibrium outcome in the game in which individuals invest in human capital, based on their private returns, and the social optimum.

To motivate this extension consider the scenario in which a "partnership" is a team of professionals (say attorneys or accountants) that can reap higher benefits working as a team than the sum of benefits that partners may collect on their own.[14] To realize such benefits, team members must, however, be trained to execute team task, i.e., they must acquire human capital. Once they have been trained, they still face an incentive (or a moral hazard) problem inasmuch as they have the option to not cooperate, denying other team members the benefits of cooperation. To this point we have analyzed the incentive to cooperate in a community with heterogenous types, some trained and some untrained. In this section we analyze the investment in human capital problem, which determines how many individuals are trained in the first place.

To relate this scenario to the analysis thus far, we assume that initially all individuals are bad types,[15] and that each individual has the option of becoming an opportunist by investing in human capital (the effect of "investment" therefore is to expand the set of available actions). These investments take place initially, and are followed by the community game we have analyzed. In order to apply the analysis above, we continue to assume that types are unobserved. This assumption is, of course, more objectionable in the context of this extension because one may verify an individual's type simply by asking for a diploma and/or interviewing a candidate. Nonetheless, what we have in mind is that there are certain aspects of training and/or type that cannot be easily ascertained using such methods. For example, it is hard to know how "seriously" the individual took her training or how committed she is to apply the skills she acquired to team production. Hence, so

[14] More explicitly, some of the advantages of team prouction come from synergies, large projects that require the efforts of several people, or because team members exchange favors. These considerations are already reflected in the prisoners' dilemma game under study (because $2a > \max(b-l, 0)$).

[15] We briefly comment on the effect of having good types at the end of this section (also about segregating bad types?).

long as investment entails private information, and some residual uncertainty remains regarding the outcomes of investments, the forces we identify here remain relevant, although their effect may be attenuated.

We proceed using the method of backwards induction. As usual, the equilibrium outcome in the community game is what dictates incentives in the investment game. Taking this point of view, the reason an individual may invest in human capital is that this enables her to interact over the long-haul with other individuals that have invested too, reaping the benefits of team production. On the other hand, an individual that does not invest in human capital is deprived of the option of entering into a long-term relationship and enjoying the benefits of team production (she may still reap a short-term benefit before the relationship she is in is terminated). Whether this trade-off is such that some (or all) individuals invest depends of course on the level of cooperation in the community, which in turn depends on how many individuals invest. As stated earlier the aim of this section is to analyze this interplay between investments and cooperation.

To be more concrete, we assume that investment in human capital costs $c > 0$, which is the same for all individuals, and is solely borne by individuals that make the investments (no subsidies or surcharges to the acquisition of human capital). The timing of the extended game is as follows. Initially all individuals are B-types. Then, before the community game starts, each individual decides whether to invest in human capital (at cost c), or not. These decisions are made independently and 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 (this situation parallels the first welfare exercise of Section 7). We also assume that a steady-state is reached immediately,[16] and that individuals who invest are randomly assigned (at $t = 0$) to phases F or S according to the steady-state probabilities.

Before we proceed we note the existence of a degenerate equilibrium in which no one invests. This equilibrium arises because of investment externalities: it takes a critical mass of agents to invest to make it worth while for anyone to invest. In the sequel we focus on other equilibria.

[16] This assumption is justified if players are patient enough or the convergence to the steady-state is sufficiently fast.

Gross Return to Investment We are interested in determining the equilibrium outcome in the investment game. To this end we derive the gross return to investment in human capital, introducing 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.

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

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 (24) and (25), and $V_F(\beta)$ and $V_S(\beta)$ are derived from the solution to (26) and (27) under these values of x and z.

Let $\Delta(\beta)$ be the gross return to investment, which is the (discounted) 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 7 *(i)* $\Delta(\beta) \geq 0$ *for all* $\beta \in [0,1]$*; (ii)* $\Delta(\beta) = 0$ *for* $\beta \in (\overline{\beta}, 1]$*; (iii)* $\dot{\Delta(0)} > 0$*, and* $\Delta(\beta)$ *is increasing in* β *for* $\beta \in [0, \underline{\beta}]$*. (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 is 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 welfare is evaluated at the mixed-strategy equilibrium. Therefore, the aggregate behavior in the community (see comment 3 after Proposition 4) is constant in β, which implies $\pi^B(\beta)$, $V_S(\beta)$ and $V_F(\beta)$ are constant as well. As a consequence, the only effect of a decrease in β is that an O-type has a smaller probability of being assigned to phase F (at $t = 0$), which makes $\pi^O(\beta)$ and, consequently, $\Delta(\beta)$ smaller.

When $\beta \in [\underline{\beta}, \overline{\beta}]$ welfare is evaluated at the good equilibrium and, as suggested in Section 7, an increasing in β has a detrimental effect on payoffs, and more so on the payoff to opportunists in the friendly phase. Combining this with the effect identified in the previous paragraph, we conclude that a change in β triggers two opposing effects, which results in a (potentially) hump-shaped Δ curve over the domain $[\underline{\beta}, \overline{\beta}]$.

Figure 5 illustrates the content of Proposition 7 (ignore for now the horizontal line with height c).

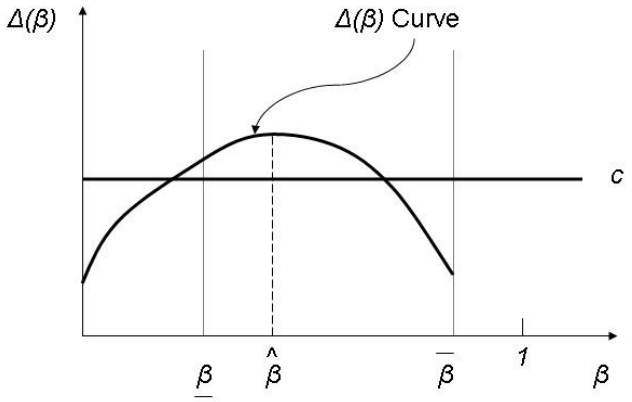

Figure 5: Endogenous Types

Equilibrium in the investment game Given the shape of Δ, as shown in Figure 5, an equilibrium in the investment game may be either interior (with some but not all individuals investing), in which case it is characterized by indifference between investing and not; or, it may be a corner

equilibrium (with all or none of the individuals investing), in which case it is characterized by a weak preference for the unanimously chosen alternative. In symbols, these possibilities are:

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

To determine which of these equilibria 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 (which depends on model parameters), pins down the type of equilibrium that occurs in the investment game.

Let's consider the possibility shown in Figure 5. Since $\Delta(\beta)$ is hump-shaped, there are (potentially) two intersection points, giving rise to two interior equilibria. We rule out the equilibrium at the higher intersection point, because it corresponds to an unstable equilibrium. Indeed, suppose that β is decreased (say) a little bit from this equilibrium value. Then, from Figure 5, at the perturbed point $\beta - \varepsilon$, $\Delta(\beta - \varepsilon) > c$, so more individuals invest in human capital, which further decreases β, drifting the system away from the original equilibrium value. On the hand, if we decrease β at the equilibrium with the lower intersection point, $\Delta(\beta - \varepsilon) < c$, so less individuals invest in human capital 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 point onwards on the stable one.

Turning to corner equilibria, Lemma 7 tells us that $\Delta(0) > 0$. 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 $c > \Delta(\widehat{\beta})$, then nobody invests.*

Proposition 5 shows that the level of human capital and the degree of cooperation in the community are positively correlated in equilibrium. Indeed, let's consider a small decrease in c. Then, the equilibrium measure of individuals investing in human capital either increases if this equilibrium is determined by the intersection of the c-line with the increasing portion of the Δ-curve, or stays constant if everyone is already investing. At the same time, the level of cooperation 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 equilibrium. Whatever combination of these possibilities materializes, a decrease in the exogenous variable c induces a non-negative correlation between the endogenous variables β and the degree of cooperation. As a consequence of this, the model predicts that in communities with more educated populace, people are more civil to each other.

Contrasting the Free entry equilibrium with the Social Optimum We contrast now the equilibrium in the investment game to a planner's optimum.

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

Proof. (i) If $c < \Delta(\underline{\beta})$ the equilibrium measure of individuals that invest exceeds $1 - \underline{\beta}$. But Lemma 5 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 7), if $c < \frac{W(\underline{\beta})}{1-\underline{\beta}}$, $S(\underline{\beta}) > S(1) = 0$, so no one investing cannot be socially optimal. On the other hand, since $\Delta(\widehat{\beta}) < c$, no one invests in the free-entry equilibrium. ■

Proposition 6 shows two departures of the equilibrium from the social optimum. On the one hand, individuals may under-invest in human capital because some of the benefit accrues to others who interact with them, and are able to realize higher payoffs in the community game. On the other hand, which might be more surprising, individuals may over-invest in human capital. This is because individuals first invest in human capital but then "undo" the investment by not cooperating.[17] It

[17] Another way to think about this is that the maximum cooperation level in the community is reached when there are $\underline{\beta} > 0$ bad types. Further decrease in β cannot increase the cooperation level, since to sustain cooperation a

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 acquire human capital because this entitles 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. Because of that investments in human capital 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 in human capital 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 there is a core of good types and, more specifically, that $\gamma \in [\underline{\gamma}, \overline{\gamma}]$. 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 when the measure of good types exceeds some critical mass. 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 coordinate on a more efficient outcome.

certain fraction of agents has to defect in the friendly phase. Therefore, if more agents than $1-\underline{\beta}$ invest in skill acquisition, some agents' investment are "reversed" (and are hence wasted) because of the structure of incentives in the community game.

References

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

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

[3] Eeckhout, J. "Minorities and Endogenous Segregation", 2002, CARESS Working Paper, University of Pennsylvania.

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

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

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

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

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

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

[10] Lindsey, J., Pollak, B. and Zeckhauser, R. "Free Love, Fragile Fidelity and Forgiveness: Social Conventions and Trust under Hidden Information", 2001, Yale University.

[11] Sobel, J. "For Better or Forever: Formal versus Informal Enforcement", 2002, Working Paper, University of California, San Diego.

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

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

[14] Watson, J. "Starting Small and Commitment", *Games and Economic Behavior*, 2002, 38, 176-199.

9 Appendix

Proof of Lemma 1

Proof of Proposition 2

Proof of Lemma 5

Proof of parts (i)-(iii) of Lemma 7

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, $\Delta(\beta) = \pi^O(\beta) - \pi^B(\beta) \geq 0$.

(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 $V_F(\beta) > V_S(\beta)$. 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 $V_F(0) > V_S(0) = 0$ and $\frac{1-z}{1-\beta}$ is positive, we have $\Delta(0) > 0$. ■

Proof of part (iv) of Lemma 7

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

$$\begin{aligned}
\Delta(\beta) &= \frac{x}{1-\beta}V_S(\beta) + (1 - \frac{x}{1-\beta})V_F(\beta) - \pi^B(\beta) \qquad (34)\\
&= V_S(\beta) - \pi^B(\beta) + (1 - \frac{x}{1-\beta})[V_F(\beta) - V_S(\beta)]\\
&= \frac{1}{1-\delta}[\frac{xa - \beta(1-\delta\rho)l}{x + \beta(1-\delta\rho)} - \frac{x}{x+\beta}b] + (1 - \frac{x}{1-\beta})\frac{\beta(a+l)}{x+\beta(1-\delta\rho)},
\end{aligned}$$

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 (34), we get

$$\Delta(y) = \frac{1}{1-\delta}[\frac{a - y(1-\delta\rho)l}{1+y(1-\delta\rho)} - \frac{1}{1+y}b] + \frac{\rho}{\rho+(1+y)(1-\rho)}\frac{y(a+l)}{1+y(1-\delta\rho)}. \tag{35}$$

Differentiating (35) and doing some algebra, we get:

$$\begin{aligned}
\Delta'(y) &= \frac{1}{1-\delta}[\frac{-(1-\delta\rho)(a+l)}{(1+y(1-\delta\rho))^2} + \frac{b}{(1+y)^2}] + \frac{\rho}{\rho+(1+y)(1-\rho)}\frac{(a+l)}{(1+y(1-\delta\rho))^2} \\
&\quad - \frac{\rho(1-\rho)}{(\rho+(1+y)(1-\rho))^2}\frac{y(a+l)}{1+y(1-\delta\rho)} \\
&= \frac{1}{(1-\delta)[1+y(1-\delta\rho)]^2} \times \\
&\quad \times \left\{ \frac{[b-(1-\delta\rho)(a+l)] + 2y(1-\delta\rho)[b-(a+l)] + (1-\delta\rho)[(1-\delta\rho)b-(a+l)]y^2}{(1+y)^2} \right. \\
&\quad \left. + \frac{\rho[1-(1-\rho)(1-\delta\rho)y^2]}{(\rho+(1+y)(1-\rho))^2}(a+l) \right\}.
\end{aligned} \tag{36}$$

We are going to show now that there is a $\widehat{y} \geq 0$ so that $\Delta'(y)$ is positive for $0 < y < \widehat{y}$ and negative for $y > \widehat{y}$, which implies that Δ has the desired hump shape property (if $\widehat{y} = 0$, Δ is increasing throughout). Since $\frac{1}{(1-\delta)[1+y(1-\delta\rho)]^2} > 0$, 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 $b \leq a+l$, it decreases in y and tends to $-\infty$ as $y \to \infty$. From these observations we infer that there are two points $y_1 \geq 0$ and $y_2 > 0$ so that the first term is positive for $y < y_1$ and negative for $y > y_1$, 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 $\frac{f_2(\widehat{y})}{g_2(\widehat{y})} > 0 > \frac{f_1(\widehat{y})}{g_1(\widehat{y})}$. 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:

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

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

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

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

(b) Consider the two terms of (36), 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 $V_S(y)-\pi^B(y) > 0$ for all $y \in (\underline{y}, \overline{y})$. Also, since $y_2 > y_1$, we have that the numerator of the second term of (36) is positive. Since the denominator of the second term is always positive, this term is positive as well, so altogether $\Delta'(\underline{y}) > 0$. Finally, since β and y are monotonically related, this implies $\Delta'(\underline{\beta}) > 0$. ■