

# Sorting and Sustaining Cooperation\*

Nick Vikander

November 2010

## Abstract

This paper look at selfish types and conditional cooperators working together in teams. Players knows that lay-offs will occur at a fixed future date, which creates incentives similar to those in a finitely repeated prisoners' dilemma.

The results show that the equilibrium with the most cooperation tends to be a sorting equilibrium, where players reveal their types so that conditional cooperators can identify and cooperate with one another. Changes in parameter values that in most situations would make cooperation more attractive, such as an increase in the discount factor, the fraction of conditional cooperators or the pay-off to reciprocated cooperation, can actually reduce equilibrium cooperation if they decrease an egoist's incentive to sort.

## 1 Introduction

A wealth of evidence suggests that many people will cooperate in team or group interactions if they expect others to do the same. The presence of these conditional cooperators can help explain cooperation in settings where free-riding is possible.<sup>1</sup>

Alongside conditional cooperators, a significant group of rational, selfish people may still choose to free ride on the work of others. For this reason, group composition can have an important effect on the extent of cooperation (see Gächter 2007 for a discussion). In a heterogeneous group, conditional cooperators may experience free-riding or believe it is likely to occur. As they do not want to be taken advantage of, this can cause cooperation to break down.

Putting like-minded, cooperative people together can therefore be an effective way of sustaining cooperating (Gächter and Thöni 2005). Once separated

---

\*I would like to thank my supervisor Maarten Janssen for helpful comments, as well as participants in the RES 2009 Conference, SMYE 2009, and seminar participants in Rotterdam, Amsterdam and Vienna. A previous version of this paper was entitled "The Breakdown of Morale".

<sup>1</sup>See for example Keser and Van Winden (2000), Fischbacher, Gächter and Fehr (2001), Frey and Meier (2004). Gächter (2007) gives an overview.

from selfish types, conditional cooperators can be confident that they can cooperate without being taken advantage of. This sorting of conditional cooperators from selfish types, as a means to sustain cooperation in teams, is the subject of this paper.

One way a firm may separate selfish from cooperative types is through the recruitment process. If employees will later work together in teams where there is an opportunity to free ride, firms may offer contracts designed to attract cooperative candidates (Kosfeld and Von Siemens 2009). In teams that are already formed, managers may try to separate out free riders from the group in order to improve work morale, for example by firing the “bad apples” (Bewley 1999). If group formation is endogenous, sorting can again promote cooperation by allowing cooperative people who identify each other to work together (Page, Putterman and Unel 2005). For example, employees may be able to change the composition of a self-managed team by forcing out those who shirk (Barker 1993).

This paper looks at how sorting can help sustain cooperation in teams, where interactions are repeated and players are either selfish or conditional cooperators. Sorting refers to a situation where, at some point, different types take different actions, so that conditional cooperators are then able to identify and work with one another.

Teams work on a productive task in each period, where players can either cooperate or defect. The incentives for selfish players, called egoists, are given by a prisoners’ dilemma, while those for conditional cooperators are given by a coordination game, and type is private information.

At the beginning of the game, players know that some of them will be laid off at a fixed future date. That is, after period  $T_0$ , a randomly selected group of players will leave the game. When the fraction of players leaving is sufficiently large, these lay-offs will cause egoists to defect in period  $T_0$ . If sorting has not occurred in a previous period, then conditional cooperators may also defect in period  $T_0$  because of the possibility that their teammate is an egoist. Cooperation in all prior periods may then break down by backwards induction. The issue considered in this paper is to what extent cooperation can be sustained in the periods prior to the lay-offs.

The first result is that, under most circumstances, sorting of types is good for cooperation. That is, unless players are sufficiently patient or the fraction of conditional cooperators is close enough to one, the equilibrium with the most cooperation will be a sorting equilibrium. A sorting equilibrium is one where, in some period  $t \leq T_0 - 1$ , different types take different actions, which allows conditional cooperators to reveal their type. Conditional cooperators can then identify each other, rematch into new teams and work together in subsequent periods. They are able to cooperate, regardless of the upcoming lay-offs, because they no longer worry that their teammate will defect in period  $T_0$ .<sup>2</sup>

---

<sup>2</sup>If conditional cooperators could not rematch, then sorting may be less effective in sustaining cooperation. However, as I discuss in Section 5, if a sorting equilibrium exists, then the parameter values for which it involves more cooperation than any non-sorting equilibrium remain the same as if players could rematch.

The second result is that a change in parameter values that in most situations would make cooperation more attractive, such as an increase in the discount factor, the fraction of conditional cooperators, or the pay-off to reciprocal cooperation, can actually reduce the amount of equilibrium cooperation. The reason is that such an increase can cause a sorting equilibrium to break down.

For a sorting equilibrium to exist, two incentive constraints must hold. One of these is the sorting constraint for egoists, who must be willing to defect in a period where others cooperate in order to reveal their type. An egoist can also deviate by mimicking a conditional cooperator and cooperating in that period, and increasing the incentive to cooperate make this deviation more attractive. If changes in parameter values cause an egoist's sorting constraint to be violated, then sorting will no longer be possible. In particular, I show that the effect of a change in the discount factor or the fraction of conditional cooperators on the amount of equilibrium cooperation can be non-monotonic: a large increase in the parameter can then increase equilibrium cooperation, but a smaller increase may have the opposite effect.

This paper does not assume that the fraction of conditional cooperators is particularly small. Indeed, experimental and field evidence suggests that conditional cooperation is relatively widespread. For example, Frey and Meier (2004) show that students are more likely to contribute to a charity fund if they believe that many others have done the same, and the effect is substantial. Andreoni and Samuelson (2006) find evidence of a range of types, from selfish player to conditional and committed cooperators, which helps explain how the distribution of pay-offs across periods can affect cooperation in a twice repeated prisoners' dilemma.

When the probability of lay-offs is sufficiently high, the strategic situation facing players prior to period  $T_0$  is similar to that in Kreps et al. (1982), where some types view the stage game as a prisoners' dilemma and other behavioral types view it as a coordination game. The main issue there is the extent to which cooperation can be sustained even if the fraction of behavioral types is small. Conlon (2003a) and Conlon (2003b) derive an explicit solution to this problem, where players use mixed strategies and the probability of defection varies with each passing period. In contrast, I restrict attention to pure strategies, which means that sorting will play a major role in promoting cooperation. This in turns means that changes in parameters that would seemingly encourage cooperation may actually be counterproductive. Furthermore, I allow for rematching and for the possibility that some players remain in the game after period  $T_0$ , which both affect the incentive to play particular equilibrium strategies.

The main theme of this paper, that sorting between different types can promote cooperative and productive outcomes, relates to other work showing that firms may want to distinguish between intrinsically motivated and selfish workers. For example, Delfgauw and Dur (2007) show that a firm wanting to fill a vacancy may not want to post a high wage, even though this will generate more applications. In fact, the firm faces a trade-off because a high wage will also reduce the average intrinsic motivation of applicants, as less motivated people

will also apply. Kosfeld and Von Siemens (2007) and Kosfeld and Von Siemens (2009) consider a quite similar set-up to the current paper, where workers are either selfish or conditional cooperators, and where they pair up into teams. They look at a one-shot game, and show that a separating equilibrium may exist where conditional cooperators all sort to particular firms and selfish types to others. The firms with conditional cooperators offer a lower wage and earn strictly positive profits, and entry is not profitable because a higher wage would attract selfish types. Important differences are that the current paper looks at sorting within existing team, rather than at hiring decisions, and where all players' incentive come from repeated interactions with teammates.

The idea that parameter changes that seemingly make cooperation more attractive may actually be counterproductive relates more generally to the issue of signaling and crowding out of intrinsic motivation. A number of papers show that offering material incentives may reduce pro-social behavior by leading selfish types to mimic intrinsically motivated types. Benabou and Tirole (2006) examine a framework where intrinsically motivated types are willing to take a pro-social action in part because it allows them to signal their type. Material incentives can lead selfish types to also take this action, thereby reducing its signaling value. Janssen and Mendys-Kamphorst (2004) make a similar point in an evolutionary framework, where they show that moderate financial incentives can cause a social norm of contributing to a public good to break down. Again, the problem is that financial incentives can lead selfish types to imitate intrinsically motivated types, so that the later cannot distinguish themselves through their contribution. Other papers highlight different reasons why explicit incentives can lead to lower performance, such as signaling that the person offering the incentives is not worth impressing (Ellingsen and Johannesson 2008), or signaling something negative about the agent's ability to do the task (Benabou and Tirole 2003).

Finally, Gächter (2007) suggests that interactions between selfish types and conditional cooperators can be related to the concept of work morale. As noted by Bewley (1999), increased turnover or a sense of unfairness may cause morale to break down, and managers also worry that low morale can be contagious. In this context, viewing sustained cooperation as high morale suggests the interpretation that sorting of types can help prevent morale from breaking down.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 derives the equilibrium with the most cooperation, and shows under which circumstances this will be a sorting equilibrium. Section 4 looks at comparative statics, how changes in certain parameter values can affect equilibrium cooperation. Section 5 looks at how robust the results are to not allowing players to rematch, and to players being uncertain about exactly when, or even if, lay-offs will occur. Section 6 then concludes. All proofs can be found in the appendix.

## 2 The Model

There are a countably infinite number of players, indexed by  $i \in \mathbb{N}$ . Players differ only in type, where player  $i$  has type  $\theta_i = \{\theta_C, \theta_E\}$ . Players of type  $\theta_C$  are conditional cooperators and players of type  $\theta_E$  are egoists. The ex-ante probability that any player is a conditional cooperator is  $\lambda \in (0, 1)$ , and type is private information.

Time is discrete and indexed by  $t = 1, 2, \dots$ . In each period, players work in teams of two on a productive task. Each player's pay-off depends on his action and the action of his teammate. Teammates play a stage game, where each player can choose to cooperate (C) or to defect (D). If  $i$  is the row player and  $j$  the column player, pay-offs are:

	C	D
C	$a + \Gamma_i, a + \Gamma_j$	$c, b$
D	$b, c$	$d, d$

The material incentives are  $a, b, c$  and  $d$ , and they give a prisoners' dilemma:  $b > a > d > c$ , and  $2a > b + c$ . The term  $\Gamma_i$  is player  $i$ 's intrinsic utility from cooperation, if his partner cooperates as well. I assume  $\Gamma_i = 0$  if  $\theta_i = \theta_E$ , and  $\Gamma_i = \Gamma > b - a$  if  $\theta_i = \theta_C$ . Egoists only value material incentives, while conditional cooperators obtain extra utility if their cooperation is reciprocated. Two conditional cooperators who knew each other's type would play a coordination game.<sup>3</sup> Denote player  $i$ 's period  $t$  action by  $a_{it} \in \{C, D\}$ , and the vector of all players' period  $t$  actions by  $\mathbf{a}_t = \{a_{it}\}_{i \in \mathbb{N}}$ .

Players have a common discount factor  $\delta \in (0, 1)$ , and each player also knows that lay-offs will occur at the end of period  $T_0$ . A randomly selected fraction  $1 - \delta_0$  of players will then leave the game, while the remaining fraction  $\delta_0 \in [0, 1)$  of players will remain in the game. If a player leaves the game, then his pay-off is zero in all periods  $t \geq T_0 + 1$ .

If players  $i$  and  $j$  are teammates, then I say these players are matched together. Let  $m_{it}$  denote the player with whom  $i$  is matched at the start of period  $t$ . Here,  $m_{it} = j$  means that players  $i$  and  $j$  play the stage game together in period  $t$ . For any given period for which player  $i$  is in the game, there must be some other player  $j$  with whom he is matched. Let the vector  $\mathbf{m}_t = \{m_{it}\}_{i \in \mathbb{N}}$  describe the set of matches, or teams, at the start of period  $t$ .

Each player observes all other players' actions, and with whom they were matched, in all previous periods. That is, when choosing  $a_{it}$ , player  $i$  has observed the full history of play up to and including period  $t - 1$ . Denote this history by  $h_{t-1}$ , where  $h_0 = \mathbf{m}_1$  and  $h_t = h_{t-1} \times \mathbf{a}_t \times \mathbf{m}_{t+1}$ . A pure strategy  $s_i$  for player  $i$  is a function which, for any history  $h_t$ , selects an action  $a_{it} \in \{C, D\}$ . I am interested in pure strategy equilibria, where I denote the strategy of conditional cooperators by  $s_C$  and the strategy of egoists by  $s_E$ .

<sup>3</sup>The material pay-offs correspond to a particular situation with joint production technology, team incentives and costly effort: output is  $2d$  when neither player exerts effort,  $2b$  when one player exerts effort, and  $2(b + a - c)$  when both players exert effort, where the cost of effort is  $b - c$  and each player receives half the team output as pay-off.

At the start of period  $t$ , player  $i$  has a belief about the type of any other player  $j$ , given history  $h_{t-1}$ . Denote that belief by  $\mu_i(j|h_{t-1}) \in [0, 1]$ , which is the probability that player  $i$  ascribes to player  $j$  being a conditional cooperator. All players will hold the same belief, so I can drop the subscript  $i$ . I just write  $\mu_{jt}$  for the belief about player  $j$  at the start of period  $t$ , where the dependence on a particular history is implicit.

Players are randomly matched before the start of period 1. At the end of period  $T_0$ , all players who remain in the game but whose teammates have been laid off are randomly rematched with one another. Players remain with their initial teammates for all  $t \leq T_0$ , unless there is some such period where, on the equilibrium path, different types take different actions. That is, unless there is a period  $t \leq T_0$  where the candidate equilibrium has conditional cooperators all cooperate and egoists all defect, or vice versa. Denote the first such period, if it exists, by  $t'$ . Then an exogenous matching mechanism takes players who have taken the same period  $t'$  action, and matches them together for period  $t' + 1$ :  $m_{it'+1} = j$  if and only if  $a_{it'} = a_{jt'}$ . Teams then stay together for all  $t' + 1 \leq t \leq T_0$ .<sup>45</sup>

I look for symmetric, pure strategy perfect Bayesian equilibria, where all players of the same type use the same strategy. Each player's strategy is optimal given the strategies of other players and given beliefs. Beliefs are consistent with player strategies, in the sense of following from Bayes' rule whenever possible. I look for equilibria where, on the equilibrium path, both types cooperate in all periods  $t \geq T_0 + 1$ . I assume players are sufficiently patient to sustain this cooperation:

$$\delta \geq \frac{b-a}{b-d}. \quad (1)$$

If there are multiple equilibria, then I am interested in the equilibrium with the most cooperation in periods  $t \leq T_0$ . By this, I mean the equilibrium which maximizes  $\sum_{t=1}^{T_0} x_t$ , where  $x_t = 1$  if both types cooperate in period  $t$ , and either  $\lambda$  or  $1 - \lambda$  if only conditional cooperators or egoists cooperate.

### 3 Analysis

I assume throughout the paper that players use trigger strategies. Let  $s_C$  denote the strategy of conditional cooperators and  $s_E$  the strategy of egoists for a

---

<sup>45</sup>While this matching technology may not seem realistic, it is sufficient for the purposes at hand. It allows conditional cooperators to be matched up together when their type is revealed, after they sort in period  $t'$ . Because both types use trigger strategies, the matching mechanism will not affect players' incentive to deviate in any other period.

<sup>5</sup>There are a countably infinite number of players, so that on the equilibrium path, both the set of players who cooperate in period  $t'$  and the set of players who defect are also countably infinite. Each player can therefore be matched with someone who took the same period  $t'$  action. The situation would be different if all but an odd number of one type deviated by changing their period  $t'$  action. This issue will not be a concern, however, since I will only consider unilateral deviations from the candidate equilibrium.

given candidate equilibrium. Suppose player  $i$  takes action  $a_{it'}$  after history  $h_{t'}$ , but neither  $s_C$  nor  $s_E$  prescribes this particular action for player  $i$  after this particular history. Then both  $s_C$  and  $s_E$  has every other player  $j$  defect in all later periods  $t \geq t' + 1$ . The punishment is self-enforcing because defection is a Nash equilibrium of the stage game for both types. Trigger strategies will imply that the only deviations which can be profitable are those that increase a player's immediate pay-off, or those that mimic the equilibrium action of the other type.

The first result is that full cooperation can be sustained if the fraction of players laid off after period  $T_0$  is sufficiently small.

**Proposition 1.** *An equilibrium exists where both types cooperate in all periods if and only if  $\delta_0 \in [\delta_0^*, 1]$ , where*

$$\delta_0^* = \frac{1 - \delta}{\delta} \frac{b - a}{a - d}.$$

The intuition for the result is straightforward. If all players remained in the game after period  $T_0$ , so  $\delta_0 = 1$ , then the game would be stationary with an infinite horizon. Then, by (1), players are patient enough to sustain an equilibrium with cooperation in all periods. An egoist's incentive to deviate in some period  $t \leq T_0$  increases if he knows that he may be laid off and leave the game after period  $T_0$ . Again by (1), players are sufficiently patient so that an egoist's best deviation is to defect in period  $T_0$ , so that if he leaves the game he can avoid the punishment altogether. If the fraction of players laid off is sufficiently high, then the strategic situation facing an egoist resembles that of a finitely repeated prisoners' dilemma, and the egoist will prefer to defect.

The proof of Proposition 1 shows that in any candidate equilibrium where both types cooperate in all  $t \geq T_0 + 1$ , no player has an incentive to deviate in any of these periods. From now on, I only consider equilibria of this sort. To avoid being repetitive, I will often describe a candidate equilibrium simply in terms of actions in the first  $T_0$  periods. It should be implicitly understood that on the equilibrium path, both types always cooperate in all  $t \geq T_0 + 1$ .

I now turn to sorting equilibria, by which I mean the following.

**Definition.** A *sorting equilibrium* is one where, on the equilibrium path, conditional cooperators and egoists take different actions in at least one period  $t \leq T_0 - 1$ .

In a sorting equilibrium, different types take actions at some point prior to the last period before the lay-offs. I will sometimes refer to other equilibria as non-sorting equilibria.<sup>6</sup>

---

<sup>6</sup>I consider an equilibrium where players pool in periods  $t \leq T_0 - 1$  and then different types choose different actions at  $t = T_0$  to be a non-sorting equilibrium. With sorting equilibria, I want to focus on how choosing different actions in some period  $t'$  can facilitate cooperation in periods  $t' + 1 \leq t \leq T_0$ . This is not possible if different types first choose different actions in period  $T_0$ .

The next result addresses the issue of when sorting will promote cooperation. More specifically, if cooperation in all periods is not possible, it shows conditions under which the equilibrium with the most cooperation will be a sorting equilibrium. This will be the case if the fraction of conditional cooperators is below a certain threshold.

**Proposition 2.** *Suppose  $\delta_0 < \delta_0^*$ , as given in Proposition 1. If both of the following conditions hold*

$$\lambda \geq \frac{d - c}{a + \Gamma - b - c + d}, \quad (2)$$

$$\lambda \geq \frac{1}{\delta(b - d)} \left[ (b - a) - \delta_0 \frac{\delta^2}{1 - \delta} (a - d) \right], \quad (3)$$

*then the equilibrium with the most cooperation has both types cooperate in all periods, except for egoists who defect in period  $T_0$ .*

*If (2) is violated, then there is a unique non-sorting equilibrium, where both types defect for all  $t \leq T_0$ . If (3) is violated but (2) is not, then there is second non-sorting equilibrium, where both types defect for all  $t \leq T_0$  except for conditional cooperators who cooperate in period  $T_0$ . Both of these equilibria have less cooperation than any sorting equilibrium.*

The proposition shows that when the fraction of conditional cooperators is sufficiently high, then almost full cooperation can be sustained in equilibrium. Both types cooperate in all periods, except for egoists who defect in period  $T_0$ . This equilibrium exists when both (2) and (3) are satisfied, in which case sorting is not needed to sustain cooperation.

When (2) holds, the fraction of conditional cooperators is high enough to make these types willing to cooperate in period  $T_0$ . The reason is that each conditional cooperator believes it sufficiently likely that his period  $T_0$  teammate is of the same type. This condition is necessary to avoid the breakdown of all cooperation in periods  $t \leq T_0 - 1$ , in a non-sorting equilibrium. If (2) does not hold, then both types will defect in period  $T_0$ , and since egoists expect this, they will defect in period  $T_0 - 1$ . Conditional cooperators must also defect in period  $T_0 - 1$  in a non-sorting equilibrium, and the logic of backwards induction then implies that no cooperation is possible in any  $t \leq T_0$ .

The proposition shows that (2) is necessary to avoid the breakdown of all cooperation in periods  $t \leq T_0 - 1$  in a non-sorting equilibrium, but it is not sufficient. In a candidate non-sorting equilibrium where both types cooperate in all  $t \leq T_0 - 1$ , an egoist knows that his teammate may defect in period  $T_0$ . The egoist may be tempted to defect one period earlier, when he knows both types will cooperate. The egoist will then be punished in period  $T_0$ , but this punishment is only effective if his teammate would have cooperated on the equilibrium path. That is the case when many players are conditional cooperators, so when  $\lambda$  is large. For the deviation not to be profitable, (3) must hold.

A sorting equilibrium always involves cooperation in multiple periods, because it allows conditional cooperators to reveal their type. After player sort in

period  $t'$ , the matching rule pairs conditional cooperators together so they can cooperate in all periods  $t \geq t' + 1$ . Even though egoists are left to defect, but there is still some cooperation in periods  $t', \dots, T_0$ .

The next result shows under what conditions a sorting equilibrium will exist, and the extent to which players will cooperate.

**Proposition 3.** *A sorting equilibrium where different types first take different actions in period  $t' \leq T_0 - 1$  exists if and only if  $t'$  satisfies the two following conditions:*

$$\lambda b + (1 - \lambda)d - \lambda(a + \theta_0) - (1 - \lambda)c \leq \frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t'})(a + \theta_0 - d), \quad (4)$$

$$\frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t' - 1})(a - d) + \delta^{T_0 - t'}(b - d) - \delta_0 \frac{\delta^{T_0 - t' + 1}}{1 - \delta}(a - d) \leq \lambda b + (1 - \lambda)d - \lambda a - (1 - \lambda)c \quad (5)$$

If a sorting equilibrium exists, let  $t'_H$  and  $t'_L$  be the largest and smallest values of  $t'$  which satisfy both (4) and (5). Consider the following condition:

$$\lambda \geq \frac{1}{\delta(b - d)} \left[ (b - a) - \delta_0 \frac{\delta^{T_0 - t' + 2}}{1 - \delta}(a - d) \right] \quad (6)$$

If  $t'_H$  satisfies (6), then the sorting equilibrium with the most cooperation has conditional cooperators cooperate for all  $t \leq T_0$ , while egoists cooperate for all  $t \leq t'_H - 1$  and defect for all  $t'_H \leq t \leq T_0$ . If  $t'_H$  does not satisfy (6), then the sorting equilibrium with the most cooperation has conditional cooperators defect for all  $t \leq t'_L - 1$  and cooperate for all  $t'_L \leq t \leq T_0$ , while egoists defect for all  $t \leq T_0$ .

The proposition shows that for a sorting equilibrium to exist, two incentive constraints must be satisfied. Each constraint concerns a player's incentive to follow his equilibrium strategy in period  $t'$ , the period when different types take different actions: conditional cooperators must have an incentive to cooperate, given by (4), while egoists must have an incentive to defect, given by (5).

If players do follow their equilibrium strategies up to and including period  $t'$ , then conditional cooperators will be rematched together in period  $t' + 1$  and earn  $a + \Gamma$  up until period  $T_0$ . Egoists will be rematched together and earn  $d$  in each of these periods. All players remaining in the game after period  $T_0$  will then cooperate in all later periods and earn, depending on their type, either  $a$  or  $a + \Gamma$ .

When deciding whether to cooperate in period  $t'$ , a conditional cooperator must weigh what may be an immediate cost of cooperation against its future benefit. The left-hand side of (4) gives the immediate cost of cooperation. It is positive whenever (2) from Proposition 2 is violated, so when defection yields a higher period  $t'$  pay-off than cooperation. The right-hand side of (4) gives the future benefits from cooperation, as a conditional cooperator will earn  $a + \Gamma$

in all periods  $t' + 1 \leq t \leq T_0$ . If a conditional cooperator defected in period  $t'$ , then players would believe he was an egoist and he would earn  $d$  in each of these periods. Since the right-hand side of (4) is strictly positive, this condition is easier to satisfy than (2). A conditional cooperator is more willing to reveal his type when he knows that doing so will increase his pay-off in future periods. Note that a conditional cooperator's best deviation is to fully imitate an egoist's equilibrium strategy, so he will not be punished after period  $T_0$ . That is why the incentive constraint (4) does not depend on  $\delta_0$ .

In contrast, an egoist must weigh the immediate benefit of defecting against its future cost. The right-hand side of (5) gives the immediate benefit from defecting, which is strictly positive since defection is an egoist's dominant strategy in the stage game. The left-hand side gives the future cost of defecting, as an egoist will earn  $d$  in all periods  $t' + 1 \leq t \leq T_0$ . If an egoist instead cooperated in period  $t'$ , then players would mistake him for a conditional cooperator and he would be able to earn  $a$  each future period. An egoist is more willing to reveal his type when  $t'$  is large, since there are then fewer periods where he must suffer the lower pay-off.

In fact, an egoist who deviates can do strictly better than cooperating in all  $t \geq t' + 1$ . Instead, he prefers to cooperate up to and including period  $T_0 - 1$ , and then defect. The egoist is then punished, because he is no longer mimics a conditional cooperator's equilibrium strategy, but this only affect him if he remains in the game. That is why the incentive constraint (5) depends on  $\delta_0$ . The reason that this particular deviation gives the highest pay-off is that, by (1), an egoist is sufficiently patient to cooperate until period  $T_0 - 1$ , and then, by  $\delta_0 < \delta_0^*$ , he is sufficiently likely to leave the game to defect in period  $T_0$ .

A sorting equilibrium exists if there is a value of  $t'$  which satisfies the incentive constraints for both types, (4) and (5). However, the amount of cooperation in the sorting equilibrium depends on another constraint, (6). If the fraction of conditional cooperators is sufficiently high, so that (6) holds, then a sorting equilibrium will exist where both types cooperate in all periods before sorting,  $t \leq t' - 1$ . The reason for this is similar to the argument behind (3) in Proposition 2. If an egoist suspects his teammate may defect in period  $t'$ , then he may be tempted to deviate by defecting a period earlier, when he knows his teammate will cooperate. He will then be punished in period  $t'$ , but this punishment is only effective if his period  $T_0$  teammate is likely a conditional cooperator who would have cooperated in equilibrium.

An egoist who deviates cannot be punished in periods  $t' + 1 \leq t \leq T_0$ , because his pay-off in the candidate equilibrium was just  $d$ . However, he can be punished again after period  $T_0$ , if he remains in the game, since then he would have earned  $a$  in each period of the candidate equilibrium. This reasoning implies that (6) is easier to satisfy when players sort later in the game, since the punishment after period  $T_0$  then weighs more heavily. So if (6) holds for any  $t'$  satisfying (4) and (5), a sorting equilibrium exists where conditional cooperators cooperate in all periods, and egoists only defect from  $t' \leq t \leq T_0$ . Types sort as late as possible,  $t' = t_{max}$ , to maximize the number of periods where all players cooperate.

If (6) does not hold for  $t' = t_{max}$ , then any non-sorting equilibrium will be

non-monotonic. Both types will defect in all periods before players sort, and then conditional cooperators will cooperate and egoists will continue to defect from  $t' \leq t \leq T_0$ . The sorting equilibrium with the most cooperation will be when types sort as early as possible,  $t' = t_{min}$ .

## 4 Comparative Statics and Equilibrium Cooperation

I now consider how a change in various parameters can affect the amount of equilibrium cooperation. I first establish the following lemma, which will be used in proving the remaining propositions. Consistent with the discussion following Proposition 3, it says that the incentive constraint of conditional cooperators is easier to satisfy if players sort early in the game, while the incentive constraint for egoists is easier to satisfy if players sort later in the game. Moreover, if there is a period where one constraint is satisfied with equality, then the other constraint will also be satisfied.

**Lemma.** *If (4) holds for  $t'$ , then it also holds for all  $t \leq t' - 1$ . If (5) holds for  $t'$ , then it holds for all  $t \geq t' + 1$ .*

*Any  $t'$  which satisfies (4) with equality strictly satisfies (5), and any  $t'$  which satisfies (5) with equality strictly satisfies (4).*

The remaining results show that an increase in parameters that in most most situations would make cooperation more attractive may actually reduce the amount of cooperation that can be sustained in equilibrium. Depending on the circumstances, an increase in the discount factor,  $\delta$ , the fraction of conditional cooperators,  $\lambda$ , or the pay-off to reciprocated cooperation,  $a$ , may all increase the returns to cooperation. However, for a sorting equilibrium to exist, an egoist must be willing to defect in some period where conditional cooperators cooperate. Increasing the returns to cooperation can increase an egoist's incentive to deviate from a sorting equilibrium, since he prefers to mimic a conditional cooperator. If (5) is then violated, the sorting equilibrium breaks down, and this can reduce the amount of equilibrium cooperation.

It is difficult to draw general conclusions about whether a change in parameter values will increase or decrease equilibrium cooperation. That is because the incentive of a conditional cooperator to follow his equilibrium strategy is in many ways opposite to that of an egoist. Making cooperation more attractive may make (4) easier to satisfy and (5) more difficult to satisfy, while making cooperation less attractive has the opposite effect.

What the following results do, however, is show that parameter values do exist such that the relationship between  $\delta$  or  $\lambda$  and equilibrium cooperation is non-monotonic. A large increase in  $\delta$  or  $\lambda$ , so that either is close enough to one, will increase the amount of equilibrium cooperation. However, a smaller increase in  $\delta$  or  $\lambda$  may actually cause equilibrium cooperation to break down.

**Proposition 4.** *Suppose that  $(b - a) - (d - c)$  is small but strictly positive,  $b - a < (a - d)(1 - \delta_0)$  and  $\delta_0 > 0$ . Then for  $a + \Gamma$  sufficiently close to  $b$ , there exist  $\delta_1 < \delta_2 < \delta_3 < 1$  such that the equilibrium with the most cooperation is as follows:*

*If  $\delta \in [\delta_1, \delta_2]$ , then it is a sorting equilibrium where different types first take different actions in period  $t' = T_0 - 1$ .*

*If  $\delta \in (\delta_2, \delta_3)$ , then both types defect in all periods  $t \leq T_0$ .*

*If  $\delta \in [\delta_3, 1]$ , then both types cooperate in all periods.*

A high value of  $\delta$  gives a conditional cooperator a larger incentive to play his strategy in a sorting equilibrium, since he must trade off the current cost of cooperation against its future benefit. In contrast, a high value of  $\delta$  gives an egoist a lower incentive to play his equilibrium strategy, because he must trade off the current benefit from defecting against its future cost. This means that an increase in  $\delta$  makes (4) easier to satisfy but (5) more difficult to satisfy. Depending on parameter values, this can cause a sorting equilibrium to break down.

The assumptions on parameters lead to the following situation. For lower values of  $\delta$ , the equilibrium with the most cooperation is a sorting equilibrium where different types take different actions at  $t' = T_0 - 1$ . For intermediate values of  $\delta$ , this sorting equilibrium breaks down, and the only non-sorting equilibrium has all players defect in all  $t \leq T_0$ . For high values of  $\delta$ , an equilibrium exists where players cooperate in all periods. The effect of  $\delta$  on equilibrium cooperation is therefore non-monotonic.

If  $\delta_0 > 0$ , so that not all players leave the game after period  $T_0$ , and if players are sufficiently patient, then an equilibrium clearly exists where players cooperate in all periods. This is the case when  $\delta \in [\delta_3, 1]$ , which is equivalent to  $\delta_0 \geq \delta_0^*$  from Proposition 1. The assumption that  $a + \Gamma$  is sufficiently close to  $b - a$  ensures that (2) is violated, so that the equilibrium with the most cooperation for lower values of  $\delta$  is indeed a sorting equilibrium. This amounts to assuming that conditional cooperators are only willing to reveal their type if doing so generates some future reward. The other assumptions imply that there is an interval  $(\delta_2, \delta_3)$  for which no sorting equilibrium exists, and that (1) is still satisfied for all values of  $\delta$  under consideration.

I now consider a similar situation with  $\lambda$ , the fraction of conditional cooperators.

**Proposition 5.** *Suppose  $\delta = 1$  and  $\delta_0 = 0$ . Then for  $b - d < d - c$ , and  $a + \Gamma$  sufficiently close to  $b$ , there exist  $\lambda_1 < \lambda_2 < \lambda_3 < 1$  such the equilibrium with the most cooperation is as follows:*

*If  $\lambda \in [\lambda_1, \lambda_2]$ , then it is a sorting equilibrium where different types first take different actions in period  $t' = T_0 - 1$ .*

*If  $\lambda \in (\lambda_2, \lambda_3)$ , then both types defect in all periods  $t \leq T_0$ .*

*If  $\lambda \in [\lambda_3, 1]$ , then it is a non-sorting equilibrium where conditional cooperators cooperate in period  $t' = T_0$ . For  $\lambda$  sufficiently close to one, both types cooperate in all periods except for egoists who defect at  $t = T_0$ .*

The reason that an increase in the fraction of conditional cooperators can reduce an egoist's incentive to sort is because  $b - d < a - c$ . This assumption, implied by  $b - d < d - c$ , means that an egoist's immediate pay-off to cooperation is higher if he expects his teammate to cooperate as well. An egoist is therefore more tempted to deviate from a sorting equilibrium if he believes his teammate is a conditional cooperator, as (5) is then more difficult to satisfy for larger values of  $\lambda$ . An increase in  $\lambda$  can then cause a sorting equilibrium to break down.

Sorting is not an issue when  $\lambda$  is sufficiently close to 1, because then the equilibrium with the most cooperation is not a sorting equilibrium. By Proposition 2, both types can then cooperate in all periods, except for egoists who defect in period  $T_0$ . In fact, the assumption on parameters ensures no sorting equilibrium exists whenever (2) is satisfied, so whenever  $\lambda \in [\lambda_3, 1]$ , or for slightly lower values of  $\lambda$ . Assuming  $b - d < d - c$  ensures that a sorting equilibrium exists for values of  $\lambda$  close enough to zero, so that sorting is possible for  $\lambda \in [\lambda_1, \lambda_2]$ , with  $\lambda_2 < \lambda_3$ .

The final result shows that an increase in the pay-off from reciprocated cooperation,  $a$ , can actually decrease the amount of equilibrium cooperation.

**Proposition 6.** *Suppose that either (i)  $\delta_0 = 0$ , the conditions for Proposition 4 are satisfied, and fix  $\delta = \delta_2$ , or (ii) the conditions for Proposition 5 are satisfied and fix  $\lambda = \lambda_2$ . Then the equilibrium with the most cooperation is a sorting equilibrium where different types first take different actions in period  $t' = T_0 - 1$ . If there is a marginal increase in  $a$ , then the equilibrium with the most cooperation has both types defect in all periods  $t \leq T_0$ .*

An increase in the value of  $a$  has an unambiguously positive effect on a conditional cooperator's incentive to play his equilibrium strategy in a sorting equilibrium. However, looking at (5) shows that an increase in  $a$  has two different effects on an egoist's incentive to play his equilibrium strategy.

First, if an egoist deviates, then an increase in  $a$  increases his pay-off in periods  $t' \leq t \leq T_0 - 1$ . An egoist who deviates can earn  $a$  if his period  $t'$  teammate cooperates, after which he can continue to earn  $a$  as long as he himself continues to cooperate. Second, an increase in  $a$  increases an egoist's pay-off in periods  $t \geq T_0$  when he plays his equilibrium strategy. This is because all players can cooperate after the lay-offs, which effectively increases the effect of the punishment following a deviation. The net effect of an increase of  $a$  on an egoist's incentive to deviate from a sorting equilibrium is therefore ambiguous. Assuming  $\delta_0 = 0$ , so that all players leave the game after period  $T_0$ , ensures that the only the first effect exists. An increase in  $a$  then increases an egoist's incentive to deviate, and can cause a sorting equilibrium to break down.

## 5 Robustness

I now examine the extent to which the results are robust to not allowing players to rematch, and instead requiring that they stay with their initial teammates

as long as both remain in the game. I also look at the impact of assuming that players do not know precisely when, or even if, lay-offs will occur.

One attractive feature of a sorting equilibrium, as described in Section 3, is that it allows all conditional cooperators to cooperate in multiple periods: between the period where players sort, and the period where lay-offs occur. The reason this cooperation is possible is because of rematching. Conditional cooperators can identify each other when different types first take different actions in period  $t'$ . If conditional cooperators can then rematch with one another, then they can coordinate on cooperation in all later periods, regardless of the lay-offs in period  $T_0$ .

A natural question, then, is to what extent sorting can sustain cooperation if players cannot rematch. Rematching might not be feasible because of administrative costs involved in making changes to the production process, or it might take a certain transition period before employees are able to work productively with one another.

Not allowing for rematching would clearly decrease the amount of cooperation in a sorting equilibrium. If different types first take different actions in period  $t'$ , then a conditional cooperator who discovers that he is working with an egoist would have to remain in the same team. If  $\delta_0 < \delta_0^*$ , as given in Proposition 1, then the conditional cooperator will expect the egoist to defect in period  $T_0$ . He will then defect himself in that period, and by backwards induction both players will defect as of period  $t' + 1$ .

Sorting will still allow some conditional cooperators to cooperate between period  $t' + 1$  and period  $T_0$ , but only those who discover that their teammate is actually a conditional cooperator as well. This amounts to a fraction  $\lambda^2$  of players, so less than the fraction  $\lambda$  who would be able to cooperate if rematching were possible. This suggests that from a quantitative point of view, the extent to which sorting can sustain cooperation is lower if players cannot be rematched.

That being said, from a qualitative point of view, not allowing for rematching will leave the main results of the previous sections unchanged. Propositions 1 and 2 will remain exactly the same, including conditions (2) and (3) which describe when the equilibrium with the most cooperation will be a sorting equilibrium. In any sorting equilibrium without rematching, there will be at least one period  $t < T_0$  where all conditional cooperators cooperate, and at least one other period prior to  $T_0$  where a fraction  $\lambda$  of conditional cooperators, those who discover their teammate is of the same type, will cooperate again. This is more cooperation than in a non-sorting equilibrium where, at best, conditional cooperators will cooperate in period  $T_0$ .

One important consequence of not allowing for rematching is that the incentive constraints in a sorting equilibrium, given by (4) and (5) from Proposition 3, will change. These constraints consider a period  $t'$  where all conditional cooperators are expected to cooperate and all egoists to defect, and describe each type's incentive to actually take its equilibrium action.

A conditional cooperator now has a lower incentive to cooperate in period  $t'$ . Whenever (2) is violated, cooperation carries an immediate cost relative to defection, because the conditional cooperator may be taken advantage of by an

egoist. Cooperation also carries a future benefit, to the extent that conditional cooperators expect to cooperate together in future periods. However, without rematching, the expected benefit is reduced to a fraction  $\lambda$  of what it was before, which is the probability that the conditional cooperator discovers that his teammate is of the same type he is. Condition (4) therefore becomes

$$\lambda b + (1 - \lambda)d - \lambda(a + \theta_0) - (1 - \lambda)c \leq \lambda \left[ \frac{\delta}{1 - \delta} (1 - \delta^{T_0 - t'}) (a + \theta_0 - d) \right], \quad (7)$$

where the only difference is the factor  $\lambda$  now on the right-hand side.

By a similar reasoning, an egoist now has a higher incentive to defect in period  $t'$ . Defection gives an immediate benefit, because it is the egoist's dominant strategy in the stage game, but it also carries an implicit future cost. By defecting, the egoist foregoes the opportunity to mimic and work with a conditional cooperator in future periods and earn a higher pay-off. The implicit cost of defecting is now a fraction  $\lambda$  of what it was before, which is the probability that the egoist's teammate is a conditional cooperator. Condition (5) therefore becomes

$$\lambda \left[ \frac{\delta}{1 - \delta} (1 - \delta^{T_0 - t' - 1}) (a - d) + \delta^{T_0 - t'} (b - d) - \delta_0 \frac{\delta^{T_0 - t' + 1}}{1 - \delta} (a - d) \right] \leq \lambda b + (1 - \lambda)d - \lambda a - (1 - \lambda)c, \quad (8)$$

where the only difference is the factor  $\lambda$  now on the left-hand side.

For a given  $t'$ , not allowing for rematching makes an egoist's incentive constraint easier to satisfy, and a conditional cooperator's incentive constraint more difficult to satisfy. Conditional cooperators are now only willing to sort for (weakly) lower values of  $t'$ , while egoists are now willing to sort for (weakly) lower values of  $t'$  for which they were not willing to sort before. This will tend to push sorting, if it does occur, to earlier in the game. Comparative statics, of the type described in Section 4, will continue to hold. In particular, comparing (8) with (5) shows that an increase in  $\lambda$  is now more likely to increase an egoist's incentive to deviate from a sorting equilibrium, and thus cause equilibrium cooperation to break down.

The above discussion suggests that allowing for rematching after sorting will not unambiguously increase the amount of equilibrium cooperation. For a given sorting equilibrium, rematching increases cooperation, since all conditional cooperators are then able to cooperate with one another. However, allowing for rematching decreases an egoist's incentive to sort. An egoist knows that if he deviates by imitating a conditional cooperator, rematching will allow him to obtain a higher pay-off in future periods. If this violates the egoist's incentive constraint, this can cause a sorting equilibrium to break down.<sup>7</sup>

<sup>7</sup>Although I do not pursue the idea here, it is worth noting that a principal's ability to rematch players after sorting could raise possible commitment issues. A principal may want to convince players that rematching will not occur, so that egoists are willing to reveal their type. However, after types are revealed, the principal could renege on his promise, and pair

I have also assumed throughout the paper that players know with certainty that lay-offs will occur after period  $T_0$ . It is arguably more realistic to assume that the timing of lay-offs is uncertain, but there is a period of time where players believe lay-offs are more likely. Moreover, players may well know that the period of time with possible lay-offs is approaching, before it actually begins. This might correspond to an expected announcement on firm performance or a discussion of restructuring, where the accompanying uncertainty could last some time.<sup>8</sup>

I now assume that lay-offs may occur at most once, and after any one of  $N \geq 1$  periods. The probability of lay-offs occurring after any given period  $T_0, T_0 + 1, \dots, T_0 + N - 1$  is  $p$ , conditional on lay-offs not having occurred before. If lay-offs do occur in one of these periods, then the fraction of players who remain in the game is again  $\delta_0$ . When  $p = 1$ , this set-up reduces to the original model.

The key point in the previous analysis was that the fraction of players laid-off must be sufficiently large to cause an equilibrium where both types cooperate in all periods to break. If that is the case, then in any equilibrium, egoists will defect in period  $T_0$ . Sorting can then help sustain cooperation leading up to period  $T_0$  by allowing conditional cooperators to identify one another.

Similar results will continue to hold under uncertainty about lay-offs, as long as an equilibrium where both types cooperate in all periods cannot be sustained. The following proposition shows this will be the case as long as the probability of lay-offs in a given period is sufficiently high, since then egoists will defect in period  $T_0$ . Moreover, the minimum probability of lay-offs needed for an equilibrium with full cooperation to break down will decrease if there is an increase in the number of periods where lay-offs may occur.

**Proposition 7.** *Suppose that lay-offs may occur in at most one of the  $N$  following periods:  $T_0, \dots, T_0 + N - 1$ . The probability of lay-offs after any one of these periods, conditional on there being no lay-offs before, is  $p \in [0, 1]$ . If lay-offs do occur, then a fraction  $1 - \delta_0$  of players leave the game before the following period.*

*Consider a candidate equilibrium where both types cooperate in all periods. If  $\delta_0 < \delta_0^*$ , then there exists a critical value  $p_N < 1$  such that, for all  $p \in (p_N, 1]$ , an egoist has a strict incentive to defect in period  $T_0$ . The critical value  $p_N$  is increasing in  $\delta_0$  and  $\delta$ , decreasing in  $(b - a)/(a - d)$ , and decreasing in  $N$  at a decreasing rate. For  $N = 1$ , we have  $p_1$  which satisfies*

$$\frac{b - a}{a - d} = \frac{\delta}{1 - \delta}(1 - p)(1 - \delta_0) + \delta_0 \frac{\delta}{1 - \delta},$$

*and as  $N$  tends to infinity,  $p_N$  tends to  $p_\infty$  which satisfies*

---

conditional cooperators together to cooperate in all future periods. Even if egoists then respond by defecting, this will not have a large effect on profits as egoists would also have defected on the equilibrium path until period  $T_0$ .

<sup>8</sup>As in the main analysis, we must have  $T_0 \geq 2$  for sorting to play any role. If  $T_0 = 1$ , then by its very definition, no sorting equilibrium can exist. It is therefore essential that players know that the increased probability of lay-offs will begin at a *future* date.

$$\frac{b-a}{a-d} = \frac{\delta}{1-\delta(1-p)}(1-p)(1-\delta_0) + \delta_0 \frac{\delta}{1-\delta}.$$

The intuition behind the proposition is as follows. An increase in  $p$  means that lay-offs are more likely, which decreases the probability an egoist will remain in the game to be punished after deviating in period  $T_0$ . A player who deviates is also more likely to suffer this punishment as the size of the lay-offs increases, and will place a higher weight on the punishment if he is patient. In both cases, as  $\delta_0$  or  $\delta$  increases, the minimum value of  $p$  needed to convince an egoist to deviate will also increase.

An egoist's immediate gain from defecting is  $(b-a)$ , and the implicit price he must pay in any subsequent period where he is punished is  $a-d$ . An increase in  $b-a/b-d$  will therefore increase his incentive to deviate. The same goes for an increase in  $N$ , as this increases the probability that he will leave the game. In both cases, the minimum value of  $p$  needed to convince an egoist to deviate will decrease.

To see how an increase in  $N$  can decrease the critical value  $p_N$ , consider the following example:  $b-a=3$ ,  $a-d=2$ ,  $\delta=2/3$  and  $\delta_0=1/2$ . These parameters imply  $p_1=0.5$  and  $p_\infty=1/4$ , which can be interpreted as follows.

For an equilibrium with cooperation in all periods to break down, lay-offs need not occur with certainty in period  $T_0$ ; it is sufficient that they occur with probability  $1/2$ . If that is not the case, but the probability of lay-offs is greater than  $1/4$ , then an equilibrium with full cooperation will still break down if the period of time where lay-offs might occur is sufficiently large.

Moreover, adding just a single extra period where lay-offs may occur, so going from  $N=1$  to  $N=2$ , will already have an important impact on  $p_N$ . For this example,  $p_2$  is approximately 0.35, so the probability of lay-offs in a period need only be slightly higher than one third. If  $p$  exceeds this critical value, then egoists will defect in period  $T_0$ , and sorting may be useful to help sustain cooperation.

## 6 Conclusion

This paper has considered a situation where cooperation within teams can break down if players foresee that lay-offs will occur at a fixed future date. The results show that unless players are sufficiently patient or the fraction of conditional cooperators is sufficiently high, then the equilibrium with the most cooperation will be a sorting equilibrium. Sorting allows players to reveal their type, so that conditional cooperators can identify and cooperate with one another. Changes in parameter values that make cooperation more attractive may actually be counterproductive, as they may increase an egoist's incentive to deviate and cause a sorting equilibrium to break down.

For future work, it may also be interesting to explore how sorting can promote cooperation when incentives come not only from repeated team interac-

tions, but in an organizational setting. The issue would then be under what circumstances a principal should induce agents to sort in the face of upcoming lay-offs, and what mechanism he should use.

## Appendix

*Proof of Proposition 1.* Consider a candidate equilibrium where, on the equilibrium path, both types cooperate in all periods  $t \geq T_0 + 1$ , and let  $s_C$  and  $s_E$  be the associated strategies. Say no player has deviated in any period  $t \leq T_0$ . I show that no player has an incentive to deviate in any period  $t \geq T_0 + 1$ .

A conditional cooperator has no incentive to deviate in any  $t \geq T_0 + 1$ , because he earns the maximum pay-off  $a + \Gamma$  in each period. An egoist can deviate by defecting in some period  $t' \geq T_0 + 1$ , after which he is punished. The deviation is not profitable if

$$b + \frac{\delta}{1 - \delta}d < \frac{1}{1 - \delta}a.$$

The right-hand side is the egoist's pay-off as of period  $t'$  in this candidate equilibrium, discounted as of that period. The left-hand side is his pay-off as of period  $t'$  from the deviation, since he is punished by the trigger strategy as of period  $t' + 1$ . The inequality is equivalent to

$$\delta \geq \frac{b - a}{b - d},$$

which holds by (1).

Now say that in this candidate equilibrium, on the equilibrium path, both types also cooperate in all  $t \leq T_0$ . Once again, a conditional cooperator has no incentive to deviate because he earns the maximum pay-off in each period. The pay-off for an egoist in this candidate equilibrium is

$$\sum_{t=1}^{T_0} \delta^{t-1}a + \delta_0 \frac{\delta^{T_0}}{1 - \delta}a.$$

The second term is an egoist's expected pay-off as of period  $T_0 + 1$ , since the probability he remains in the game after the lay-offs is  $\delta_0$ .

Say an egoist deviates by defecting in some period  $t \leq T_0$ , and let  $t'$  be the first such period. His pay-off from the deviation is

$$\sum_{t=1}^{t'-1} \delta^{t-1}a + \delta^{t'-1}b + \sum_{t=t'+1}^{T_0} \delta^{t-1}d + \delta_0 \frac{\delta^{T_0}}{1 - \delta}d.$$

The egoist earns  $b$  in the period where he first defects and then  $d$  in all later periods where he is punished. This pay-off is increasing in  $t'$  if  $a + \delta b \geq b + \delta d$ , which holds by (1). The most profitable such deviation is therefore to choose  $t' = T_0$ , so that the egoist is only punished if he remains in the game. The deviation is profitable if

$$a + \delta_0 \frac{\delta}{1 - \delta}a < b + \delta_0 \frac{\delta}{1 - \delta}d,$$

which is equivalent to

$$\frac{1 - \delta}{\delta} \frac{b - a}{a - d} < \delta_0.$$

The left-hand side is just  $\delta_0^*$ , so no player has an incentive to deviate if and only if  $\delta \in [\delta_0^*, 1]$ .

*Proof of Proposition 2.* Since  $\delta < \delta_0^*$ , Proposition 1 implies there is no equilibrium where both types cooperate in all periods. Consider a candidate equilibrium where, on the equilibrium path, both types cooperate in all periods, except for egoists who defect in period  $T_0$ .

A conditional cooperator has no incentive to deviate from this candidate equilibrium in any  $t \leq T_0 - 1$ , because he earns the maximum pay-off  $a + \Gamma$  in each period. The pay-off for a conditional cooperator in this candidate equilibrium as of period  $T_0$ , discounted as of that period, is

$$\lambda(a + \Gamma) + (1 - \lambda)c + \delta_0 \frac{\delta}{1 - \delta} (a + \Gamma).$$

Both types have cooperated in all  $t \leq T_0 - 1$ , so the probability that any player's teammate in period  $T_0$  is a conditional cooperator is just the prior,  $\lambda$ . In this candidate equilibrium, conditional cooperators cooperate and egoists defect in period  $T_0$ . The first two terms therefore give a conditional cooperator's expected period  $T_0$  pay-off.

A conditional cooperator who deviates by defecting in period  $T_0$  is not punished in later periods, because he mimics an egoist's period  $T_0$  equilibrium action. His pay-off from the deviation, as of period  $T_0$ , is

$$\lambda b + (1 - \lambda)d + \delta_0 \frac{\delta}{1 - \delta} (a + \Gamma).$$

A conditional cooperator does not have an incentive to deviate if

$$\lambda b + (1 - \lambda)d \leq \lambda(a + \Gamma) + (1 - \lambda)c$$

which is equivalent to (2). The pay-off for an egoist in this candidate equilibrium is

$$\sum_{t=1}^{T_0-1} \delta^{t-1} a + \delta^{T_0-1} [\lambda b + (1 - \lambda)d] + \delta_0 \frac{\delta^{T_0}}{1 - \delta} a.$$

The terms in square brackets give an egoist's expected pay-off in period  $T_0$ . The probability his period  $T_0$  teammate is a conditional cooperator who will cooperate is  $\lambda$ , while the probability he is an egoist who will defect is  $1 - \lambda$ .

It cannot be profitable for an egoist to deviate by cooperating in period  $T_0$ , since defection is his dominant strategy in the stage game. If an egoist has a profitable deviation, it will be to defect earlier than expected, in some period  $t \leq T_0 - 1$ . Let  $t' \leq T_0 - 1$  be the first period where he defects. The pay-off for an egoist from this deviation is

$$\sum_{t=1}^{t'-1} \delta^{t-1} a + \delta^{t'-1} b + \sum_{t=t'+1}^{T_0} \delta^{t-1} d + \delta_0 \frac{\delta^{T_0}}{1 - \delta} d.$$

The egoist increases his period  $t'$  pay-off from  $a$  to  $b$ , but he is then punished in all subsequent periods. The pay-off from this deviation is increasing in  $t'$  if  $a + \delta b \geq b + \delta d$ , which holds by (1). The most profitable deviation is therefore to first defect in period  $t' = T_0 - 1$ . That gives pay-off

$$\sum_{t=1}^{T_0-2} \delta^{t-1} a + \delta^{T_0-2} b + \delta^{T_0-1} d + \delta_0 \frac{\delta^{T_0}}{1-\delta} d.$$

An egoist does not have an incentive to deviate if

$$b + \delta d + \delta_0 \frac{\delta^2}{1-\delta} d \leq a + \delta [\lambda b + (1-\lambda)d] + \delta_0 \frac{\delta^2}{1-\delta} a,$$

which is equivalent to (3).

An equilibrium cannot exist where both types cooperate in all periods, except for conditional cooperators who defect in period  $T_0$ . In this candidate equilibrium, an egoist could profitably deviate by defecting in period  $T_0$ . This would increase his immediate pay-off, because defection is his dominant strategy in the stage game, and he would not be punished because he mimics a conditional cooperator's period  $T_0$  equilibrium action. This implies that when (2) and (3) both hold, the equilibrium with the most cooperation is one where both types cooperate in all periods, except for egoists who defect in period  $T_0$ .

An equilibrium always exists where both types defect for all  $t \leq T_0$ . No player has an incentive to deviate by cooperating in any period  $t' \leq T_0$ , because doing so would reduce his immediate pay-off from  $d$  to  $c$  and trigger a punishment. If (2) is violated, then this is the only non-sorting equilibrium.

If (3) is violated but (2) is not, then another non-sorting equilibrium exists where both types defect for all  $t \leq T_0$ , except for conditional cooperators who cooperate in period  $T_0$ . No player has an incentive to deviate by cooperating in any period  $t' \leq T_0$ , because doing so would reduce his immediate pay-off from  $d$  to  $c$  and trigger a punishment. An egoist has no incentive to deviate in period  $T_0$  because defection is his dominant strategy in the stage game, while a conditional cooperator has no incentive to deviate in period  $T_0$  because (2) holds.

I now show that when (3) is violated but (2) is not, no another non-sorting equilibrium exists. In any such equilibrium, there must be some period  $t' \leq T_0 - 1$  where both types cooperate. An egoist who deviates by defecting in period  $t'$  is punished in all subsequent periods. Because this is a non-sorting equilibrium, we know that, on the equilibrium path, either both types cooperate or both types defect in period  $t$ , for each  $t' + 1 \leq t \leq T_0 - 1$ . The punishment only reduces the egoist's period  $t$  pay-off if, on the equilibrium path, both types would have cooperated in period  $t$ . An egoist therefore has the lowest incentive to deviate from a candidate equilibrium where both types cooperate for all  $t \leq T_0 - 1$ . But because (3) is violated, an egoist has a profitable deviation from that candidate equilibrium.

Now suppose that a sorting equilibrium exists. By the definition of a sorting equilibrium, there is some period  $t' \leq T_0 - 1$  where different types first take different actions. By Bayes' rule, players then update their beliefs to  $\mu_{it+1} = 1$  if player  $i$  is a conditional cooperator and  $\mu_{it'+1} = 0$  if player  $i$  is an egoist. The matching rule pairs each conditional cooperator with a teammate who is also a conditional cooperator, and these players can then cooperate for all  $t \geq t' + 1$ . Since  $t' \leq T_0 - 1$ , any sorting equilibrium yields more cooperation than an equilibrium where both types defect for all  $t \leq T_0 - 1$ .

*Proof of Proposition 3.* Consider a candidate sorting equilibrium where different types first take different actions in period  $t' \leq T_0$ . Say that in this period, egoists cooperate and conditional cooperators defect. The pay-off for an egoist as of period  $t'$  in this candidate equilibrium is

$$\lambda c + (1 - \lambda)a + \sum_{t=1}^{T_0-t'} \delta^t d + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} a.$$

The first two terms are an egoist's expected pay-off in period  $t'$ . There has been no period  $t \leq t'$  where different types have taken different actions, so the probability that any player's teammate in period  $T_0$  is a conditional cooperator who will defect is just the prior,  $\lambda$ .

By Bayes' rule, players then update their beliefs to  $\mu_{it+1} = 1$  if player  $i$  is a conditional cooperator and  $\mu_{it'+1} = 0$  if player  $i$  is an egoist. The matching rule pairs each conditional cooperator with a teammate who is also a conditional cooperator, and these players can then cooperate for all  $t \geq t' + 1$ . Each egoist is paired with another egoist, and by backwards induction they defect in all  $t' + 1 \leq t \leq T_0$ .

By a similar reasoning, the pay-off for a conditional cooperator as of period  $t'$  in this candidate equilibrium is

$$\lambda d + (1 - \lambda)b + \sum_{t=1}^{T_0-t'} \delta^t (a + \Gamma) + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} (a + \Gamma). \quad (9)$$

An egoist can deviate by defecting in period  $t'$ , and mimicking a conditional cooperator's strategy in all subsequent periods. An egoist's pay-off as of period  $t'$  from this deviation is given by (9) with  $\Gamma$  set equal to zero. This is strictly higher than an egoist's pay-off in the candidate equilibrium, for each period  $t \leq t' \leq T_0$ . Hence, in any sorting equilibrium, conditional cooperators must cooperate and egoists must defect in period  $t'$ .

Now consider a candidate sorting equilibrium where conditional cooperators cooperate and egoists defect in period  $t'$ , and whether any player has an incentive to deviate in any  $t \geq t'$ . The pay-off for a conditional cooperator as of period  $t'$  in this candidate equilibrium is

$$\lambda(a + \Gamma) + (1 - \lambda)c + \sum_{t=1}^{T_0-t'} \delta^t (a + \Gamma) + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} (a + \Gamma). \quad (10)$$

The pay-off for an egoist as of period  $t'$  in this candidate equilibrium is

$$\lambda b + (1 - \lambda)d + \sum_{t=1}^{T_0-t'} \delta^t d + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} a. \quad (11)$$

A conditional cooperator has no incentive to deviate in any  $t \geq t' + 1$ , because he earns the maximum pay-off  $a + \Gamma$  in each period. The only deviation which could be profitable in any  $t \geq t'$  is to mimic an egoist and defect in period  $t'$ . The pay-off for a conditional cooperator as of period  $t'$  from this deviation is

$$\lambda b + (1 - \lambda)d + \sum_{t=1}^{T_0-t'} \delta^t d + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} (a + \Gamma), \quad (12)$$

which is the same as (11), but with  $a$  replaced by  $a + \Gamma$ .

An egoist who defects in period  $t'$  has no incentive to deviate in any period  $t' + 1 \leq t \leq T_0$ , because doing so would reduce his immediate pay-off from  $d$  to  $c$  and trigger a punishment. If an egoist has a profitable deviation in any period  $t \geq t'$ , then it must involve mimicking a conditional cooperator's equilibrium action and cooperating in period  $t'$ . The matching rule then pairs the egoist with a conditional cooperator, who believes that he is a conditional cooperator with probability one.

An egoist who deviates by cooperating in period  $t'$  knows that his teammate will cooperate as long as he continues to mimic a conditional cooperator's equilibrium strategy, which is to cooperate in each period. The egoist's incentive to defect in any period  $t \geq t' + 1$  is therefore the same as in a candidate equilibrium where both types cooperate in all periods. By the same argument as in the proof of Proposition 1, the most profitable deviation is to cooperate for all  $t' \leq t \leq T_0 - 1$  and then defect in period  $T_0$ .

The pay-off for an egoist as of period  $t'$  from this deviation is

$$\lambda a + (1 - \lambda)c + \sum_{t=1}^{T_0-t'-1} \delta^t a + \delta^{T_0-t'} b + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} d. \quad (13)$$

Comparing (10) and (12), a conditional cooperator does not have an incentive to deviate in any  $t \geq t'$  if

$$\lambda b + (1 - \lambda)d + \sum_{t=1}^{T_0-t'} \delta^t d \leq \lambda(a + \Gamma) + (1 - \lambda)c + \sum_{t=1}^{T_0-t'} \delta^t (a + \Gamma)$$

which is just (4).

Comparing (11) to (13), an egoist does not have an incentive to deviate in any  $t \geq t'$  if

$$\lambda b + (1 - \lambda)d + \sum_{t=1}^{T_0-t'} \delta^t d + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} a \leq$$

$$\lambda a + (1 - \lambda)c + \sum_{t=1}^{T_0-t'-1} \delta^t a + \delta^{T_0-t'} b + \delta_0 \frac{\delta^{T_0-t'+1}}{1 - \delta} d,$$

which is equivalent to (5). Therefore no player has an incentive to deviate in any  $t \geq t'$  if and only if both (4) and (5) hold.

Now say (4) and (5) hold, and consider a candidate sorting equilibrium where both types defect in all periods  $t \leq t' - 1$ . No player has an incentive to deviate by cooperating in any period  $t' \leq t' - 1$ , because doing so would reduce his immediate pay-off from  $d$  to  $c$  and trigger a punishment. There is therefore an equilibrium where conditional cooperators defect for all  $t \leq t' - 1$  and cooperate for all  $t' \leq t \leq T_0$ , while egoists defect for all  $t \leq T_0$ .

Now consider a candidate sorting equilibrium where there is some period  $t'' \leq t' - 1$  where both types cooperate. This will be an equilibrium if and only if an egoist does not have an incentive to deviate by defecting in any such period  $t''$ .

An egoist who defects in period  $t''$  is punished in all future periods. This punishment only reduces his pay-off in periods where he would have obtained more than  $d$  in the candidate equilibrium. This implies that an egoist's incentive to deviate in period  $t''$  is lowest for a candidate equilibrium where both types cooperate for all  $t \leq t' - 1$ .

The pay-off for an egoist in this candidate equilibrium is

$$\sum_{t=1}^{t'-1} \delta^{t-1} a + \delta^{t'-1} (\lambda b + (1-\lambda)d) + \sum_{t=t'+1}^{T_0} \delta^{t-1} d + \delta_0 \frac{\delta^{T_0}}{1-\delta} a \quad (14)$$

The pay-off for an egoist from deviating in period  $t''$  is

$$\sum_{t=1}^{t''-1} \delta^{t-1} a + \delta^{t''-1} b + \sum_{t=t''+1}^{T_0} \delta^{t-1} d + \delta_0 \frac{\delta^{T_0}}{1-\delta} d$$

The pay-off from this deviation is increasing in  $t''$  if  $a + \delta b \geq b + \delta d$ , which holds by (1). The most profitable deviation is therefore to set  $t'' = t' - 1$ , which yields

$$\sum_{t=1}^{t'-2} \delta^{t-1} a + \delta^{t'-2} b + \sum_{t=t'+1}^{T_0} \delta^{t-1} d + \delta_0 \frac{\delta^{T_0}}{1-\delta} d \quad (15)$$

An egoist does not have an incentive to deviate if (15)  $\leq$  (14), which is equivalent to (6). No player then has an incentive to deviate from this candidate equilibrium, where conditional cooperators cooperate for  $t \leq T_0$ , and egoists cooperate for  $t \leq t' - 1$  and defect for  $t' \leq t \leq T_0$ .

Note that the right-hand side of (6) is decreasing in  $t'$ , so it is easiest to satisfy for  $t' = t'_{max}$ . This means that when  $t'_{max}$  satisfies (6), the sorting equilibrium with the most cooperation is where both types cooperate until  $t = t'_{max}$ , and then different types take different actions. If  $t'_{max}$  does not satisfy (6), then the sorting equilibrium with the most cooperation is where both types defect until  $t = t'_{min}$ , and then different types take different actions.

*Proof of Lemma.* To establish the first part, it is sufficient to show that the right-hand side of (4) and the left-hand side of (5) are both decreasing in  $t'$ . For (4), this is clearly the case because of the term  $(1 - \delta^{T_0-t'})$ . The left-hand side of (5) can be rewritten as

$$\frac{\delta}{1-\delta} (1 - \delta^{T_0-t'}) (a-d) + \delta^{T_0-t'} (b-a) - \delta_0 \frac{\delta^{T_0-t'+1}}{1-\delta} (a-d) \quad (16)$$

where  $\delta^{T_0-t} (a-d)$  has been added to the first term and subtracted from the second term. Substituting  $t' + 1$  for  $t'$  gives

$$\frac{\delta}{1-\delta} (1 - \delta^{T_0-t'-1}) (a-d) + \delta^{T_0-t'-1} (b-a) - \delta_0 \frac{\delta^{T_0-t'}}{1-\delta} (a-d)$$

and subtracting (16) from this expression yields

$$\frac{\delta}{1-\delta} (-\delta^{T_0-t'-1} + \delta^{T_0-t}) (a-d) + (\delta^{T_0-t'-1} - \delta^{T_0-t'}) (b-a) + \delta_0 \frac{\delta}{1-\delta} (-\delta^{T_0-t'-1} + \delta^{T_0-1}) (a-d).$$

Dividing by  $\delta^{T_0-t'-1} - \delta^{T_0-1} > 0$  gives

$$-\frac{\delta}{1-\delta} (a-d) + (b-a) - \delta_0 \frac{\delta}{1-\delta} (a-d),$$

which is negative if

$$b - a \leq \frac{\delta}{1 - \delta}(a - d)(1 + \delta_0).$$

To minimize the right-hand side, I set  $\delta_0 = 0$  and choose the lowest value of  $\delta$  so that (1) is still satisfied,  $\delta = (b - a)/(b - d)$ . The right-hand side then reduces to  $b - a$ , so the inequality is satisfied. The left-hand side of (5) is therefore decreasing in  $t'$ .

Rewriting (4) gives

$$\lambda b + (1 - \lambda)d - \lambda a - (1 - \lambda)c \leq \frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t'})(a + \theta_0 - d) + \lambda\Gamma.$$

Since the left-hand side of (5) is equal to (16), to establish the second part of the lemma it is sufficient to show that

$$\frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t'})(a - d) + \delta^{T_0 - t'}(b - a) - \delta_0 \frac{\delta^{T_0 - t' + 1}}{1 - \delta}(a - d) \leq \frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t'})(a + \theta_0 - d) + \lambda\Gamma.$$

The left-hand side is decreasing in  $\delta_0$ , while the right-hand side is increasing in  $\Gamma$ . I set  $\delta_0 = 0$  and let  $\Gamma$  tend to  $b - a$  to give

$$\delta^{T_0 - t'}(b - a) \leq \frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t'})(b - a) + \lambda(b - a),$$

$$[\delta^{T_0 - t'} - \frac{\delta}{1 - \delta}(1 - \delta^{T_0 - t'}) - \lambda](b - a) \leq 0,$$

which holds if

$$[(\frac{\delta^{T_0 - t'} - \delta}{1 - \delta}) - \lambda](b - a) < 0.$$

This inequality holds because the term in the square brackets is negative.

*Proof of Proposition 4.* From Proposition 1, an equilibrium exists where both types cooperate in all periods if  $\delta_0 \geq \delta_0^*$ . By the definition of  $\delta_0^*$ , this is equivalent to

$$\delta \leq \frac{b - a}{b - a + \delta_0(a - d)} \quad (17)$$

Define  $\delta_3$  as the right-hand side of this inequality, where  $\delta = \delta_3$  implies  $\delta_0 = \delta_0^*$ . Both types can cooperate in all periods if  $\delta \in [\delta_3, 1]$ , where  $\delta_3 < 1$  by  $\delta_0 > 0$ .

Consider (5) evaluated at  $t' = T_0 - 1$ . That is

$$\delta(b - d) - \delta_0 \frac{\delta^2}{1 - \delta}(a - d) \leq \lambda b + (1 - \lambda)d - \lambda a - (1 - \lambda)c. \quad (18)$$

Evaluating (18) at  $\delta = \delta_3$  gives

$$\delta_3(b - d) - \delta_3(b - a) \leq \lambda b + (1 - \lambda)d - \lambda a - (1 - \lambda)c, \quad (19)$$

because  $\delta_0^*(a - d)[\delta/(1 - \delta)] = b - a$ . Rearranging gives

$$\delta_3(a - d) \leq \lambda(b - a) + (1 - \lambda)(d - c).$$

I now show that this inequality does not hold. Substituting for  $\delta_3$  from (17) yields

$$\left[ \frac{b-a}{b-a+\delta_0(a-d)} \right] (a-d) \leq \lambda(b-a) + (1-\lambda)(d-c).$$

Since  $b-a > d-c$ , it is sufficient to show that

$$\left[ \frac{1}{b-a+\delta_0(a-d)} \right] (a-d) > 1,$$

which is true because  $b-a < (1-\delta_0)(a-d)$ . Lemma 1 then implies that (5) must also be violated for all  $t' \leq T_0 - 2$ , so that when  $\delta = \delta_3$ , no sorting equilibrium exists.

Although (18) does not hold for  $\delta = \delta_3$ , it clearly does hold for sufficiently small  $\delta$ . Moreover, the left-hand side of (18) is concave in  $\delta$ . This implies there is a single value of  $\delta \in (0, \delta_3)$  such that (18) holds with equality. Define this value of  $\delta$  as  $\delta_2$ , so no sorting equilibrium exists when  $\delta \in (\delta_2, \delta_3)$ .

Since (18) holds with equality at  $t' = T_0 - 1$  when  $\delta = \delta_2$ , Lemma 1 implies that (4) must strictly hold at  $t' = T_0 - 1$  when  $\delta = \delta_2$ . Moreover, (18) also holds for all  $\delta \leq \delta_2$ . By continuity, this means that there is some  $\delta_1 < \delta_2$  such that a sorting equilibrium exists where different types first take different action in period  $t' = T_0 - 1$  for  $\delta \in [\delta_1, \delta_2]$ .

I now show that (1) holds for all  $\delta \geq \delta_1$ , where it will be sufficient to show that (18) strictly holds when evaluated at  $\delta = (b-a)/(b-d)$ . Substituting and rearranging gives

$$(1-\lambda)(b-a) - (1-\lambda)(d-c) \leq \delta_0 \frac{\delta^2}{1-\delta} (a-d).$$

The right-hand side is strictly positive, and because  $(b-a) - (d-c)$  is small, the inequality holds.

Finally, (2) is violated for  $\Gamma$  close enough to  $b-a$ , since  $b > a$ ,  $d > c$  and  $\lambda < 1$ . By Proposition 2, this implies that for all  $\delta < \delta_3$ , any non-sorting equilibrium has both types defect for all  $t \leq T_0$ . In particular, this is the case for  $\delta \in (\delta_2, \delta_3)$ .

*Proof of Proposition 5.* Define  $\lambda_3$  as the value of  $\lambda$  for which (2) holds with equality. This implies

$$\lambda_3 b + (1-\lambda_3)d - \lambda_3 a - (1-\lambda_3)c = \lambda_3 \Gamma. \quad (20)$$

Consider (5) for  $t' = T_0 - 1$ , which is given by (18), with  $\delta = 1$  and  $\delta_0 = 0$ .

$$b-d \leq \lambda b + (1-\lambda)d - \lambda a - (1-\lambda)c. \quad (21)$$

Letting  $\lambda = \lambda_3$  and substituting from (20) into (21) yields

$$b-d \leq \lambda_3 \Gamma.$$

The left-hand side is strictly greater than  $b-a$ , since  $a > d$ . (21) is therefore violated for  $a + \Gamma$  sufficiently close to  $b$ . Moreover, the right-hand side of (21) is decreasing in  $\lambda$ , as  $b-d < d-c$  implies  $b-d < a-c$ . This means that (21) is also violated for all  $\lambda > \lambda_3$ .

By Lemma 1, (5) is also violated for all  $t' \leq T_0 - 2$  whenever  $\lambda \in [\lambda_3, 1]$ , so no sorting equilibrium exists for any value of  $t'$ . The equilibrium with the most cooperation is therefore a non-sorting equilibrium as described in Proposition 2.

Pay-offs from any given strategy are continuous in  $\lambda$ , so (5) is also violated for all  $t' \leq T_0 - 1$  whenever  $\lambda$  marginally less than  $\lambda_3$ . No sorting equilibrium exists, and since (2) is violated the only equilibrium is for both types to defect in all  $t \leq T_0$ .

I showed that (21) does not hold for  $\lambda = \lambda_3$ , but it must hold for  $\lambda$  sufficiently close to zero because  $b - d < d - c$ . This implies there is some  $\lambda_2 \in (0, \lambda_3)$  such that (21) is violated when  $\lambda > \lambda_2$ , but holds with equality when  $\lambda = \lambda_2$ . By Lemma 1, (4) must then strictly hold for  $t' = T_0 - 1$  when  $\lambda = \lambda_2$ . A sorting equilibrium therefore exists when  $\lambda = \lambda_2$ , where different types first take different actions in period  $T_0 - 1$ . Again, because pay-offs are continuous in  $\lambda$ , (5) still holds for  $\lambda$  marginally less than  $\lambda_2$ . This means that there is some  $\lambda_1 < \lambda_2$  such that the sorting equilibrium exists for all  $\lambda \in [\lambda_1, \lambda_2]$ .

*Proof of Proposition 6.* The claim about the equilibrium with the most cooperation before the marginal increase in  $a$  follows directly from Propositions 4 and 5, where this part of Proposition 4 did not depend on  $\delta_0 > 0$ .

Consider (5) for  $t' = T_0 - 1$ , when  $\delta_0 = 0$ :

$$\delta(b - d) \leq \lambda b + (1 - \lambda)d - \lambda a - (1 - \lambda)c.$$

By the definition of  $\delta_2$  and  $\lambda_2$ , this condition holds with equality when  $\delta = \delta_2$  or  $\lambda = \lambda_2$ . The left-hand side is independent of  $a$ , while the right-hand side is decreasing in  $a$ . So keeping  $\delta$  fixed at  $\delta_2$  or  $\lambda$  fixed at  $\lambda_2$ , the condition will be violated by a marginal increase in  $a$ . By Lemma 1, (5) will then also be violated for all  $t' \leq T_0 - 2$ . Proposition 3 then implies that a sorting equilibrium does not exist.

The proofs of Propositions 4 and 5 showed that (2) was violated at  $\delta = \delta_2$  and  $\lambda = \lambda_2$ . Pay-offs are continuous in  $a$ , so (2) must also be violated after a marginal increase in  $a$ . By Proposition 2, the equilibrium with the most cooperation therefore has both types defect in all periods  $t \leq T_0$ .

*Proof of Proposition 7.* Consider an egoist's expected pay-off in some period  $t$  of the candidate equilibrium, with  $t \geq T_0 + 1$ . The expected pay-off is  $a$ , appropriately discounted and multiplied by the probability that a player will still be in the game in that period.

Define  $i = t - T_0$ . The probability that a player will still be in the game in period  $t \in \{T_0 + 1, \dots, T_0 + N\}$ , is  $(1 - p)^i + [1 - (1 - p)^i]\delta_0$ ; it is the probability that lay-offs have not yet occurred, plus the probability they have occurred but the player has remained in the game. For any  $t \geq T_0 + N + 1$ , the probability is the same as for  $t = T_0 + N$ :  $(1 - p)^N + [1 - (1 - p)^N]\delta_0$ .

A player's expected pay-off as of period  $T_0$ , discounted as of that period, is therefore

$$a + \sum_{i=1}^N \left\{ (1 - p)^i + [1 - (1 - p)^i]\delta_0 \right\} \delta^i a + \sum_{i=N+1}^{\infty} \left\{ (1 - p)^N + [1 - (1 - p)^N]\delta_0 \right\} \delta^i a.$$

An egoist who defects in period  $T_0$  and is punished by the trigger strategy in later periods obtains

$$b + \sum_{i=1}^N \left\{ (1 - p)^i + [1 - (1 - p)^i]\delta_0 \right\} \delta^i d + \sum_{i=N+1}^{\infty} \left\{ (1 - p)^N + [1 - (1 - p)^N]\delta_0 \right\} \delta^i d.$$

This deviation is profitable if

$$b - a > (a - d) \left\{ \sum_{i=1}^N [(1-p)^i (1-\delta_0) + \delta_0] \delta^i + (a - d) \sum_{i=N+1}^{\infty} [(1-p)^N (1-\delta_0) + \delta_0] \delta^i \right\} \quad (22)$$

Plugging in  $p = 1$  gives

$$b - a > (a - d) \frac{\delta_0 \delta}{1 - \delta},$$

which holds by  $\delta_0 < \delta_0^*$ . Plugging in  $p = 0$  gives

$$b - a > (a - d) \frac{\delta}{1 - \delta},$$

which is violated by (1). The right-hand side of (22) is decreasing in  $p$ , since  $\delta_0 < 1$ . This implies that there is a critical value  $p_N$ , with  $0 < p_N < 1$ , such that an egoist has an incentive to deviate if and only if  $p \in (p_N, 1]$ .

Say  $p = p_N$ , so the left hand side and the right-hand side of (22) are equal. An increase  $(b - a)/(a - d)$  will then mean that (22) holds, and so the egoist has an incentive to deviate. A smaller probability of lay-offs is then sufficient to convince an egoist to defect in period  $T_0$ , so  $p_N$  decreases.

Say again  $p = p_N$ , and consider an increase in  $\delta_0$ . This increases the right-hand side of (22), since  $(1 - p)^i < 1$ , as does an increase in  $\delta$ . The egoist then has a strict incentive to play his equilibrium strategy, so  $p_N$  increases.

Subtracting the right-hand side of (22) from the same expression, but with  $N$  replaced by  $N + 1$ , and canceling terms gives

$$[(1 - p)^{N+1} - (1 - p)^N] \sum_{t=N+1}^{\infty} \delta^t (1 - \delta_0) (a - d).$$

This expression is negative, since  $(1 - p)^{N+1} < (1 - p)^N$ , but it is increasing in  $N$ . The right-hand side of (22) is therefore decreasing in  $N$  at a decreasing rate. This implies that  $p_N$  is also decreasing in  $N$ , at a decreasing rate.

Setting  $N = 1$  in (22) and replacing the inequality by an equality gives

$$\frac{b - a}{a - d} = \frac{\delta}{1 - \delta} (1 - p)(1 - \delta_0) + \delta_0 \frac{\delta}{1 - \delta},$$

which defines  $p_1$ . As  $N$  tends to infinity, (22) tends to

$$b - a > (a - d) \sum_{i=1}^{\infty} [(1 - p)^i (1 - \delta_0) + \delta_0] \delta^i,$$

Replacing the inequality by an equality and substituting for the value of the geometric series gives

$$\frac{b - a}{a - d} = \frac{\delta}{1 - \delta(1 - p)} (1 - p)(1 - \delta_0) + \delta_0 \frac{\delta}{1 - \delta}.$$

which gives the value of  $p_{\infty}$ .

## References

- [1] Andreoni, J. and L. Samuelson (2006). “Building Rational Cooperation.” *Journal of Economic Theory* 127: 117–154.
- [2] Barker, J.R. (1993). “Tightening the Iron Cage: Concertive Control in Self-Managing Teams.” *Administrative Science Quarterly* 38: 408–437.
- [3] Benabou, R. and J. Tirole (2003). “Intrinsic and Extrinsic Motivation.” *Review of Economic Studies* 70: 489–520.
- [4] Benabou, R. and J. Tirole (2006). “Incentives and Prosocial Behavior.” *American Economic Review* 96: 1652–1678.
- [5] T. Bewley (1999). *Why Wages Don’t Fall During a Recession*. Harvard University Press.
- [6] JR. Conlon (2003a). “Hope Springs Eternal.” *Journal of Economic Theory* 112: 35–65.
- [7] JR. Conlon (2003b). “Gang of Four Revisited.” *Mimeo, University of Mississippi*.
- [8] Delfgaauw, J. and R. Dur (2007). “Signaling and Screening of Workers’ Motivation.” *Journal of Economic Behavior & Organization* 62: 605–624.
- [9] Ellingsen, T. and M. Johannesson (2008). “Pride and Prejudice: The Human Side of Incentive Theory.” *American Economic Review* 98: 990–1008.
- [10] Fischbacher, U., Gächter, S. and E. Fehr (2001). “Are People Conditionally Cooperative? Evidence From a Public Goods Experiment.” *Economic Letters* 71: 397–404.
- [11] Frey, B. and S. Meier. (2004). “Social Comparisons and Pro-social behavior: Testing “Conditional Cooperation” in a field experiment.” *American Economic Review* 94: 1717-1722.
- [12] S. Gächter (2007). “Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications.” Published in *Economics and Psychology - A Promising New Cross-Disciplinary Field*. MIT Press.
- [13] Gächter, S. and C. Thöni (2005). “Social Learning and Voluntary Cooperation Among Like-Minded People.” *Journal of the European Economic Association* 3: 303–314.
- [14] Janssen, MCW and E. Mendys-Kamphorst (2004). “The Price of a Price: on the Crowding out and in of Social Norms.” *Journal of Economic Behavior & Organization* 55: 377–395.
- [15] Kezer, C. and F. van Winden (2000). “Conditional Cooperation and Voluntary Contributions to Public Goods.” *The Scandinavian Journal of Economics* 102: 23–39.
- [16] Kosfeld, M. and F. Von Siemens (2007). “Competition, Cooperation and Corporate Culture.” *IZA Discussion Paper No. 2927*.
- [17] Kosfeld, M. and F. Von Siemens (2009). “Worker Self-Selection and the Profits from Cooperation.” *Journal of the European Economic Association* 7: 573–582.
- [18] Kreps, D., Milgrom, P., Roberts, J. and R. Wilson (1982). “Rational Cooperation in the Finitely Repeated Prisoners’ Dilemma.” *Journal of Economic Theory* 27: 245–252.
- [19] Page, T., Putterman, L. and B. Unel (2005). “Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency.” *Economic Journal* 115: 1032–1053.