Online appendix for "Cooperation through collective punishment and participation"

Contents

A Theoretical Analysis

A.1 Further Analysis of Peer-to-Peer Punishment with Commitment

This appendix analyses the case of standard punishment with commitment, as implemented in the experiment. In the first stage each player j has committed $p_{ji} \in \{0, P\}$ punishment points to player i, to be realised in case j receives a signal that i has not contributed, i.e. if j observes the signal $s_{ji} = 0$. In the second stage, each player i decides on a contribution $g_i \in \{0, e\}$. Consider the contribution decision of individual i. Her utility is

$$u_i\left(g_i, g_{-i}\right) = \max\left\{0, \alpha G + m - g_i - P\sum_{j\neq i}^{n} \mathbb{1}[s_{ji} = 0]p_{ji} - \beta\sum_{j\neq i}^{n} \mathbb{1}[s_{ij} = 0]p_{ij}\right\}.$$

Let

$$q = \Pr\left(s_{ij} = \bar{g}|g_j = \bar{g}\right) = \Pr\left(s_{ij} = 0|g_j = 0\right) > \frac{1}{2}.$$

In our experiment $q = 0.6$, $\alpha = 0.5$, $e = 20$, $m = 10$, $P = 15$, and $\beta = 1/3$. Moreover, in the experiments on standard private punishment with commitment the following obtained. (1) Each participant could decide whether to punish each other with $\bar{p} = 15$ points or not. Thus, each player can maximally use 45 points (to the other players). (2) Committed punishment points can only be conditioned on the noisy signals, not on total contributions. (3) The punishment points committed towards a player is known by that player before the contribution decisions are made. Under these assumption the following result obtains.

Proposition 1. *For all values of q there is a symmetric subgame perfect equilibrium in which no one ever contributes and no one commits to punish. If $q = 0.6$ then there does not exist a subgame perfect Nash equilibrium (SPNE) in which everyone contributes.*

It is elementary to demonstrate the existence of the symmetric subgame perfect equilibrium in which no one ever contributes and no one ever punishes. We prove the rest of the proposition by establishing that in any second stage subgame in which everyone has committed to punish if and only they obtain a signal of non-contribution it is the case that non-contribution gives a strictly higher payoff than contribution. Formally we establish:

Lemma 1. *Suppose it is the case that all of i's co-players commit to punish if an only if they obtain a signal of non-contribution. If $q = 0.6$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$ for any g_{-i} such that $\sum_{j\neq i}^{n} g_j \in \{0, 20, 40, 60\}$.*

Proof. Since all signals are independent we have

$$\begin{aligned}
\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= (1-q)^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - 25 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +3q\left(1-q\right)^2 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - 10 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +3q^2\left(1-q\right) \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 5 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +q^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 20 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right],
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] &= q^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - 15 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +3q^2\left(1-q\right) \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +3q\left(1-q\right)^2 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 15 - \frac{1}{3}\sum_{j\neq i}^{n} [s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +\left(1-q\right)^3 \mathbb{E}\left[\max\left\{0, \frac{1}{2}\sum_{j\neq i}^{n} g_j + 30 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}_{\{s_{ij}=0\}}p_{ij}\right\}\right],
\end{aligned}$$

Since $\sum_{j\neq i}^{n} g_j \in \{0, 20, 40, 60\}$, we consider four different main cases (all of which will contain four subcases).

Case 1. **No one else contributes.** If $\sum_{j\neq i}^{n} g_j = 0$ then

$$\begin{aligned}
\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 3q^2\left(1-q\right) \mathbb{E}\left[\max\left\{0, 5 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}[s_{ij}=0]p_{ij}\right\}\right] \\
&\quad +q^3 \mathbb{E}\left[20 - \frac{1}{3}\sum_{j\neq i}^{n} \mathbb{1}_{\{s_{ij}=0\}}p_{ij}\right],
\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] &= 3q\left(1-q\right)^2\mathbb{E}\left[15-\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}\right] \\ &\quad +\left(1-q\right)^3\mathbb{E}\left[30-\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}\right].\end{aligned}$$

Since $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}_{\{s_{ij}=0\}}p_{ij}\in\{0,5,10,15\}$, there are four subcases to consider.
(1.1) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=15$; we have, if $q=0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 5q^3=1.08 \\ &> 0.96=15\left(1-q\right)^3=\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

(1.2) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=10$; we have, if $q=0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 10q^3=2.16 \\ &< 2.72=15q\left(1-q\right)^2+20\left(1-q\right)^3=\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

(1.3) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=5$; we have, if $q=0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 15q^3=3.24 \\ &< 4.48=30q\left(1-q\right)^2+25\left(1-q\right)^3=\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

(1.4) Conditional on $\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=0$; we have, if $q=0.6$,

$$\begin{aligned}\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] &= 15q^2\left(1-q\right)+20q^3=6.48 \\ &> 6.24=45q\left(1-q\right)^2+30\left(1-q\right)^3=\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right].\end{aligned}$$

Putting (1.1)-(1.4) together we have

$$\begin{aligned}&\mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]-\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] \\ =\ & q^3\mathbb{E}\left[u_i\left(20, g_{-i}\right)-u_i\left(0, g_{-i}\right)\middle|\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=15\right] \\ &+3q^2\left(1-q\right)\mathbb{E}\left[u_i\left(20, g_{-i}\right)-u_i\left(0, g_{-i}\right)\middle|\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=10\right] \\ &+3q\left(1-q\right)^2\mathbb{E}\left[u_i\left(20, g_{-i}\right)-u_i\left(0, g_{-i}\right)\middle|\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=5\right] \\ &+\left(1-q\right)^3\mathbb{E}\left[u_i\left(20, g_{-i}\right)-u_i\left(0, g_{-i}\right)\middle|\frac{1}{3}\sum_{j\neq i}^{n}\mathbb{1}[s_{ij}=0]p_{ij}=0\right]\end{aligned}$$

For $q = 0.6$ we use the calculations from above to obtain

$$\begin{aligned} & \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right] - \mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] \\ = & \left(0.6\right)^3\left(1.08 - 0.96\right) \\ & +3\left(0.6\right)^2\left(1 - 0.6\right)\left(2.16 - 2.72\right) \\ & +3\left(0.6\right)\left(1 - 0.6\right)^2\left(3.24 - 4.48\right) \\ & +\left(1 - 0.6\right)^3\left(6.48 - 6.24\right) \\ = & -0.55776 \end{aligned}$$

Thus if $\sum_{j\neq i}^{n} g_j = 0$ and $q = 0.6$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
The remaining cases, defined by $\sum_{j\neq 1}^{n} g_j$ being equal to 20, 40, and 60 respectively, are completely analogous.
Case 2. If $\sum_{j\neq i}^{n} g_j = 20$ and $q < 0.64268$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
Case 3. If $\sum_{j\neq 1}^{n} g_j = 40$ and $q < 0.6042$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
Case 4. If $\sum_{j\neq 1}^{n} g_j = 60$ and $q < 0.60422$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$.
Combining cases 1-4 we find that if $q = 0.6$ then $\mathbb{E}\left[u_i\left(0, g_{-i}\right)\right] > \mathbb{E}\left[u_i\left(20, g_{-i}\right)\right]$. □

A.2 Relaxing Selfishness and Rationality

As a theoretical benchmark we have used selfish materialistic preferences and subgame perfect Nash equilibrium (SPNE). This implies that each agent acts to maximize their own monetary payoff and agents have correct beliefs about the game and the consequences of their actions. One may wish to relax the assumption that players are selfish money-maximizers, since it has been argued that negative reciprocity is an important motivation in social dilemmas with punishment opportunities (Fehr and Gächter, 2000, 2002). This can be modelled by assuming that people derive utility from the act of punishing non-contribution, or derive utility from decreasing the payoff of non-contributors. Alternatively, one may relax the perfect information and rationality assumption. In what follows we discuss both alternatives in more detail.

A.2.1 Relaxing the Selfishness Assumption: Negative Reciprocity

We model negative reciprocity as follows. If player j has not contributed, then player i suffers a disutility of $\rho_i \pi_j$, where π_j denotes the material payoff to player j (in the current round). For simplicity suppose $\rho_i = \rho \geq 0$ for all i. We analyze the different punishment regimes as follows:

Collective punishment without commitment. Consider a candidate SPNE which prescribes that everyone contributes, and prescribes that in case a single player does not contribute then collective punishment should be implemented. We do not need to specify what happens in case more than one person fails to contribute, since we only need to make sure that there are no profitable unilateral deviations from full contribution. Let $\hat{\pi}_i := \alpha G + e - g_i$ denote the material payoff absent punishment. If j does not contribute then player i who is chosen to decide on collective punishment will find it profitable to punish if

$$\hat{\pi}_i - P\left(1 + \beta\right) - \rho\left(\alpha G + e - P\left(1 + \beta\right)\right) \geq \hat{\pi}_i - \rho\left(\alpha G + e\right) \Longleftrightarrow \rho \geq 1.$$

Here we have ignored limited liability since for our parameters payoffs will not be negative

when there is a single non-contributor. It turns out that a player is willing to implement collective punishment if she is has sufficiently reciprocal preferences. More precisely, the marginal benefit of decreasing the payoff of a non-contributor, ρ, has to be larger than the marginal cost of doing so, 1.

Collective punishment with commitment. Since full contribution was an SPNE in the model without negative reciprocity, adding negative reciprocity does not alter the predictions.

Peer-to-peer punishment without commitment. Consider a candidate SPNE in which everyone contributes. Upon observing that total contributions fall short of full contributions, players may use their private signals to form beliefs about which players have failed to contribute. If signals were perfect (unlike in our experiment), then signals could be taken at face value. In this case, when i receives as signal that j has not contributed, then player i believes that player j has indeed not contributed and punishes j if

$$\pi_i - g_i - P\beta - \rho\left(\alpha G + e - P\right) \geq \pi_i - \rho\left(\alpha G + e\right) \Longleftrightarrow \rho \geq \beta.$$

That is, the marginal benefit of decreasing the payoff of a non-contributor (ρ) has to be larger than the marginal cost of doing so (β). Since signals are not perfect, we should interpret $\rho \geq \beta$ as a necessary condition.

Peer-to-peer punishment with commitment. It still holds that for $q = 0.6$ there is no SPNE in which everyone contributes. As before, the problem does not stem from a lack of willingness to punish, but from the fact that noisy signals make punishment an ineffective deterrence.

In total, when players are negatively reciprocal (in the particular way we have assumed) the conditions for a full contribution equilibrium are most favorable under Collective punishment with commitment ($\rho \geq 0$), since a full cooperation equilibrium requires at least $\rho \geq \beta$ (if it exists) under Peer-to-peer punishment without commitment, requires $\rho \geq 1$ under Collective punishment without commitment, and does not exist under Peer-to-peer punishment with commitment (i.e. no ρ allows for full contribution).

A.2.2 Relaxing the Rationality Assumption: Perceived Indefinite Repetition

Suppose agents are completely selfish but boundedly rational in the sense that they (incorrectly) think of the game as if there was an indefinite number of rounds, with a continuation probability $\delta \in (0, 1)$ rather than (correctly) a fixed number of rounds. As is well-known, indefinite repetition can induce selfish players to cooperate in social dilemmas provided that the continuation probability δ is high enough. The threshold δ^* above which cooperation can be maintained can be used as a measure of how conducive the environment is to cooperation, with a lower threshold δ^* implying that cooperation is easier to achieve. In the following we analyze the different punishment regimes. We restrict attention to simple grim-trigger strategies which use the stage game Nash equilibrium payoff as a threat to incentivize punishment.

Collective punishment without commitment. First, consider the following candidate SPNE strategy s^{CP}: *(i) in the first round contribute and implement collective punishment*

if there was not full contribution, (ii) contribute and punish if and only if in each previous round either everyone contributed or collective punishment was implemeted. Suppose we are in a subgame where in each previous round either everyone contributed or collective punishment was implemented (or we are in the initial round). Complying with s^{CP} at the contribution step is profitable if[1]

$$\frac{4\alpha e}{1-\delta} \geq (3\alpha+1)\, e - P\,(1+\beta) + \frac{\delta 4\alpha e}{1-\delta} \iff \frac{40}{1-\delta} \geq 30 + \frac{\delta 40}{1-\delta},$$

which holds for all δ. Complying with s^{CP} at the punishment step is profitable if

$$\begin{aligned} \alpha G + e - g_i - P\,(1+\beta) + \frac{4\alpha e}{1-\delta} \quad &\geq \quad \alpha G + e - g_i + \frac{\delta e}{1-\delta} \\ &\iff \quad \frac{40\delta}{1-\delta} \geq 20 + \frac{20\delta}{1-\delta} \\ &\iff \quad \delta \geq \delta^{CP} := 1/2 \end{aligned}$$

In all other subgames s^{CP} proposes playing the SPNE of the stage game, hence it prescribes a SPNE of any such subgame. Thus, s^{CP} is a SPNE iff $\delta \geq \delta^{CP} := 1/2$.

Now consider a different candidate SPNE strategy s^{NoP}: *(i) in the first round contribute (ii) contribute if and only if in each previous round everyone contributed (iii) never punish.* Suppose we are in a subgame where in each previous round everyone contributed (or we are in the initial round). Complying with s^{NoP} at the contribution step is profitable if

$$\begin{aligned} \frac{4\alpha e}{1-\delta} \quad &\geq \quad (3\alpha+1)\, e + \frac{\delta e}{1-\delta} \\ &\iff \quad \frac{40}{1-\delta} \geq 50 + \frac{20\delta}{1-\delta} \\ &\iff \quad \delta \geq \delta^{NoP} := 1/3. \end{aligned}$$

Complying with s^{NoP} at the punishment step is always profitable. In all other subgames s^{NoP} proposes playing the SPNE of the stage game, hence it prescribes a SPNE of any such subgame. Thus, s^{NoP} is a SPNE iff $\delta \geq \delta^{NoP} := 1/3$.

We see that under collective punishment without commitment, punishment does not make any difference when the players treat the game as indefinitely repeated: the critical value of δ is lowest for strategy s^{NoP} which does not rely on punishment.

Collective punishment with commitment. Since full contribution was an SPNE in the model without negative reciprocity, introducing indefinite repetition does not alter the predictions: punishment and full contribution is still an equilibrium (for all levels of δ).

Peer-to-peer punishment without commitment. Note that players in our experiment

[1]Here $\frac{4\alpha e}{1-\delta}$ is the payoff when everyone cooperates indefinitely starting in the current period, and $\frac{\delta 4\alpha e}{1-\delta}$ is the payoff when everyone cooperates indefinitely starting in the next period, whereas $(3\alpha+1)\, e$ is the payoff in the contribution step when one does not contribute, and $P\,(1+\beta)$ is the loss due to collective punishment.

only have information about whether they were punished and whether they punished. They do not have information about punishment among other players. This limits the effectiveness of repeated game strategies to incentivise punishment. Still we can use strategy s^{NoP} from above. As before, s^{NoP} is SPNE if $\delta \geq \delta^{NoP} := 1/3$.

Peer-to-peer punishment with commitment. Players may ignore the punishment option and simply use strategy s^{NoP}, which is SPNE iff $\delta \geq \delta^{NoP} := 1/3$.

In total, when players treat the game as indefinitely repeated full contribution levels are possible in all treatments. The conditions for a full contribution equilibrium are most favorable under Collective punishment with commitment ($\delta \in (0, 1)$) than unde the other regimes ($\delta \in (1/3, 1)$).

B Implementation of the Experiment

The experimental sessions were conducted at [blinded for peer review] . Ethical approval was obtained by the [blinded for peer review] Participants were recruited using Orsee Greiner (2004). Subjects received written instructions and were allowed to ask questions before the experiment which were answered in private. The experiments were programmed using zTree Fischbacher (2007). Overall, 340 participants participated in our main treatments. This includes the strong punishment treatment reported on in the Appendix We had 12 or 13 groups (independent observations) in the treatments without commitment and around double that size (23 groups) in the treatments with commitment. The reason to have more observations with commitment is that for some analyses we split the data in the commitment treatments by whether punishment was committed or not (Section 4.2) and we wanted to have an adequate sample size for each of these cases. Sessions lasted approximately 1 hour 45 min, including instructions, a short post experimental questionnaire and payment of participants. On average participants were paid approximately $ 24, including a show up fee of $4.5 .

C Sample Instructions

We provide the Instructions for treatment C-COMM. Instructions for other treatments are available on authors' webpages or upon request.

General information

You are about to participate in a decision making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in cash at the end of the experiment.

This set of instructions is for your private use only. During the experiment you are not allowed to communicate with anybody. In case of questions, please raise your hand. Then we will come to your seat and answer your questions. Any violation of this rule excludes you immediately from the experiment and all payments.

Throughout the experiment you will make decisions about amounts of tokens. At the end of the experiment all tokens you have will be converted into pounds at the exchange rate 1 pounds for 150 token and paid you in cash in addition to the show-up fee 2.50 pounds.

All your decisions will be treated confidentially both during the experiment and after the experiment. This means that none of the other participants will know which decisions you made.

Experimental Instructions

The experiment will consist of 50 decision making periods. Each period consists of two stages. At the beginning of the experiment, you will be randomly matched with three other people in this room. Therefore, there are 4 people, including yourself, participating in your group. You will be matched with the same people during the entire experiment. None of the participants knows who is in which group.

In each period you, and each other person in your group, will be given an endowment of 20 tokens. In each period you will be asked to either place your endowment in a private account or a group account.

Your private account already has 10 tokens in each period. If you place your endowment in the private account, the private account will have 30 tokens at the end of the period. If you do not place your endowment in the private account, the private account will have 10 tokens at the end of the period. This means that the private account has a return of 1. Nobody except yourself benefits from your private account. The tokens that you place in the group account are summed together with the tokens that the other three members of your group place in the group account. Every member of the group benefits equally from the tokens in the group account. Specifically, the total amount of tokens placed in the group account by all group members is doubled and then is equally divided among the four group members. Hence, your share of the group account at the end of the first stage is

2∗ (sum of tokens in the group account) / 4

(PLEASE TURN OVER) ■

Before you decide whether to allocate your tokens to the private or to the group account, you will be asked to choose whether you would like to introduce a subtraction mechanism. Specifically, you will be asked to make four choices: whether you would like to introduce the mechanism if the total number of tokens in the group account is 0, 20, 40 or 60. All group members will make this choice simultaneously. However, only the choice of one randomly selected group member will be relevant. This means that your choice is relevant with a 25 percent chance.

If the subtraction mechanism is introduced, then it will automatically subtract tokens from all group members in case there is at least one group member who did not place their tokens in the group account. In this case it will subtract 15 tokens from each group member. In addition, if the mechanism is activated, this has a cost of 5 tokens per group member. The following table illustrates the relation between your cost in tokens and the amount of tokens that are taken away from every member of your group (including you):

Tokens subtracted	Cost for you
15	5

If all group members place their tokens in the group account, then the subtraction mechanism does not subtract tokens from anyone and there is no cost. You will know whether the subtraction mechanism was implemented for each of the four cases before you make your choice. In the second stage of each period you will be informed for each group member whether they placed their tokens in the private or group account. However, for each group member, this information is only correct with a 60% chance. It is wrong with a 40% chance.

You will also receive one piece of information that is always correct. In particular we will tell you how much money was placed in the group account.

At the end of each period, you will be informed about

- the size of the group account.
- your share of the group account (remember it is the same for all group members).
- the size of your private account.
- whether tokens were subtracted from you.
- your total earnings in this period.

This information is always correct.

All other participants will receive exactly the same instructions. Your total income in the end of the experiment is equal the sum of earnings you obtained in each period. At the end of the experiment there will be a short questionnaire for you to fill in.

This is the end of the instructions. If you have any questions please raise your hand and an experimenter will come by to answer them

D Sample Properties

Table 7 contains some balancing checks. There are about 65 percent women in our experiment, but there are no substantial nor statistically significant differences across our treatments. The average age of our participants is between 21.5-23 years across our main treatments, again without substantial nor statistically significant differences. We asked participants about which class they feel most they belong to. Around 36-38 percent of participants associate with working class (class 1), 34-35 percent with lower middle class (class 2), 24-25 percent with upper middle class (class 3) and only 0-3 percent with upper class (class 4). There are no systematic differences across treatments with the exception of treatment **C** where somewhat fewer participants identify as upper middle class. Also note that the R2 is low in all regressions. Treatments explain less than 3 percent of the variation in these parameters.

Table 7: Balancing checks

	(1) gender	(2) age	(3) class1	(4) class2	(5) class3	(6) class4	(7) risk	(8) trust	(9) rec	(10) recn
S	-0.052	0.948	-0.104	0.135	-0.020	-0.010	-0.385	-1.104*	-0.885*	-1.667
	(0.0929)	(1.138)	(0.094)	(0.088)	(0.064)	(0.035)	(0.240)	(0.624)	(0.493)	(1.377)
S-COMM	-0.014	0.014	-0.059	0.102	-0.010	-0.031	-0.510**	-0.916*	-0.488	-1.804
	(0.077)	(1.055)	(0.092)	(0.087)	(0.067)	(0.029)	(0.243)	(0.434)	(0.440)	(1.164)
C	0.055	0.252	0.125	0.002	-0.115*	-0.012	-0.127	-1.053*	-0.031	-1.500
	(0.081)	(1.334)	(0.113)	(0.097)	(0.062)	(0.034)	(0.261)	(0.544)	(0.511)	(1.260)
C-COMM	-0.091	-0.692	-0.048	0.102	-0.054	0.001	-0.107	-1.079***	-0.401	-0.196
	(0.080)	(0.983)	(0.087)	(0.070)	(0.056)	(0.034)	(0.258)	(0.399)	(0.432)	(1.265)
Constant	0.656***	22.09***	0.375***	0.344***	0.250***	0.031	6.531***	11.69***	17.78***	9.500***
	(0.061)	(0.874)	(0.077)	(0.061)	(0.044)	(0.029)	(0.180)	(0.326)	(0.360)	(1.109)
Observations	316	316	316	316	316	316	316	316	316	316
R-squared	0.011	0.009	0.023	0.010	0.009	0.009	0.013	0.015	0.015	0.023

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: OLS regression of questionnaire variables on treatment dummies. The baseline is the no punishment (**N**) treatment. Robust standard errors clustered at the matching group level are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

We also elicited attitudes towards risk, trust, positive reciprocity (rec) and negative reciprocity (recn) at the end of the experiment. Here we do see some significant differences. Most notably, participants in all our treatments with punishment possibilities (**S**, **S-COMM**, **C** and **C-COMM**) are less trusting compared to treatment **N**. Note, though, that as this outcome is elicited at the end of the experiment it is not exogenous to the treatments. It is plausible that participants who experienced punishment report to be less trusting afterwards. We should also note that there is no significant difference in trust across our punishment treatments

E Additional Treatments and Results

E.1 Inefficacy of Strong Punishment

It has been suggested by Ambrus and Greiner (2012) that severe peer-to-peer punishment may result in relative high contribution and welfare levels under imperfect monitoring. In their *s*trong punishment treatment participants may subtract more punishment points from the punished participants at a lower cost. Importantly, their framework differs from the present work in featuring a substantially lower noise rate.[2] In particular, their setting featured a one sided noise rate[3] of 10%, thus making the correct assessment of one's opponent's action the rule rather than the exception.

In order to i) contrast the efficacy of collective punishment to the benchmark of strong punishment but also to ii) assess the role of strong punishment in the present high noise environment we ran a strong peer-to-peer punishment treatment similar to the one in Ambrus and Greiner (2012), where participants could subtract 30 punishment points at the cost of 5 from each other. The treatment is identical to treatment **S** except for the fact that instead of the 1:3 punishment technology now a 1:6 punishment technology is used. We ran one session with this treatment with 24 participants in six groups. The strong punishment treatment **S-Strong** was exactly the same as the punishment treatment with the exception that now 5 tokens bought 30 punishment points, implying $\beta = \frac{1}{6}$ as in the strong punishment treatments of Ambrus and Greiner (2012).

Figure 7 reports contribution rates and resulting payoffs. While contributions (left panel) are slightly higher under strong punishment, payoffs (right panel) are substantially lower. Both of these differences are statistically significant ($p < 0.05$, for payoffs its $p < 0.01$). The profit-ratio discussed in Section 4.3 equals only 0.34 in the case of strong punishment and again this is statistically different from that of the peer-to-peer punishment treatments.

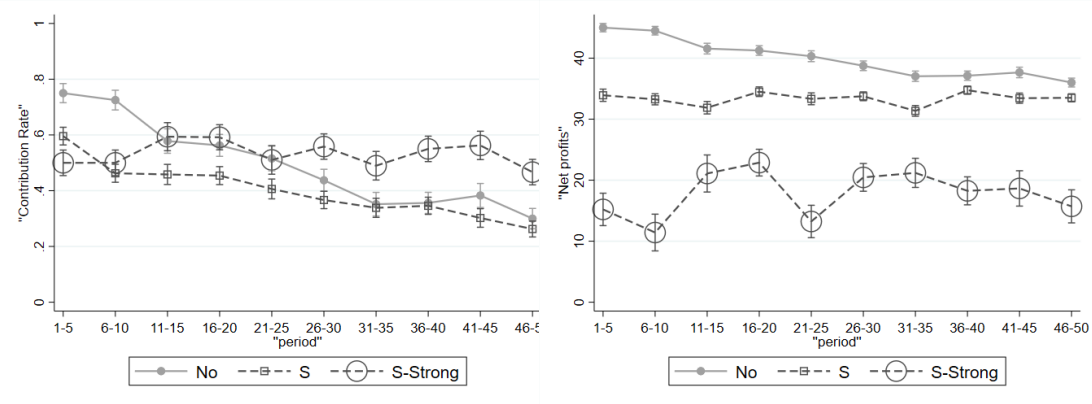
Result 7. *While contribution levels under strong punishment are higher than those under peer-to-peer punishment or in the absence of the punishment, payoff levels are substantially lower than in the peer-to-peer- or no- punishment case.*

One of the conclusions of Ambrus and Greiner (2012) is that the negative welfare implications associated with peer-to-peer punishment under imperfect monitoring may be offset by increasing the severity of punishment, resulting in welfare levels comparable to those where no punishment is available. In the case of high noise levels this is not the case. Simply providing participants with a larger stick results in vendetta-like dynamics, featuring excessive punishment and substantially lower welfare levels than those

[2]Further, their work features groups of three participants and a more fine-tuned punishment technology where participants can vary the severity of punishment. In contrast, in our setting participants can either punish or not, thus enabling a straightforward implementation of commitment. In addition and in contrast to Ambrus and Greiner (2012), per round payoffs cannot not be negative in our setting.

[3]That is non-contributors were always identified as non-contributors and contributors were incorrectly labelled as non-contributors in 10% of all cases.

Figure 7: Contributions and payoffs under strong punishment

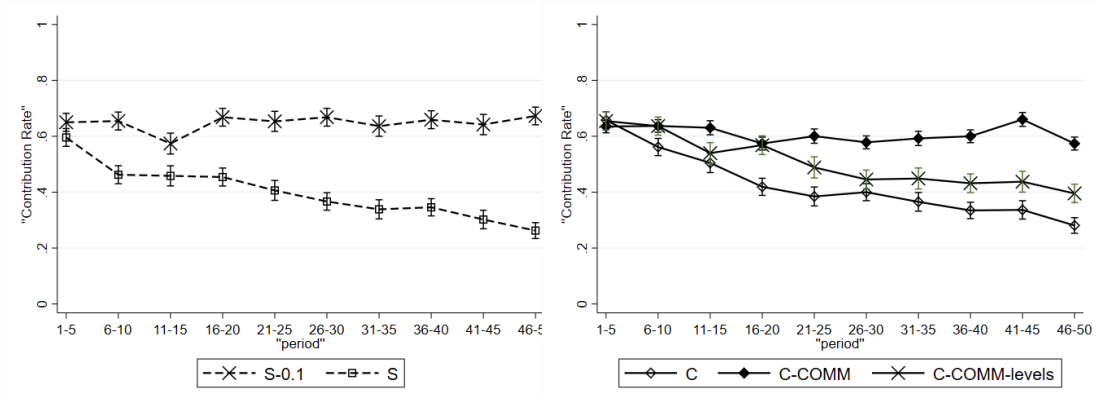


under peer-to-peer punishment; we identify a substantial positive correlation between punishments in periods t and $t-1$ ($\rho = 0.50$, $p = 0.00$).

E.2 Other treatments

At the beginning of this experimental study we also conducted a standard punishment treatment with a lower noise level of 0.1 (**S-0.1**). We had 44 participants (11 groups) in this treatment. Unlike some of the earlier literature, we found, however, that in our setting cooperation rates were high under standard punishment when noise levels are low. Hence, we found that with low noise levels standard punishment was successful in solving the free-rider problem. As we are interested in situations where standard punishment fails to solve this problem, we decided subsequently to conduct the entire experiment with high noise levels. The left panel in Figure 8 compares contribution rates under standard punishment with low (0.1) and high (0.4) noise.

Figure 8: Contribution Rates Additional Treatments



Apart from this treatment we conducted a few more treatments. After we finished all sessions, we were wondering how effective C-COMM would be if we allowed participants to condition the punishment mechanism on more than just total payoffs (assume they could observe whether 1,2, or 3 people contributed). We conducted such a hybrid treatment and found that cooperation rates lie in between those of **C** and **C-COMM** (right panel of

Figure 8). Last, we also conducted some sessions with no as well as standard punishment where participants were given the wrong information. We are not reporting on those but are happy to share results upon request.