Enforcing Cooperation in Public Goods Games: Is One Punisher Enough?

Abstract

We experimentally investigate a finitely repeated public goods game setting where, in each round, access to sanctioning power is exclusively awarded to one single player per group. We show that our central 'Top Contributors as Punishers' institution – a mechanism by which a player needs to be the highest contributor in her group in order to earn the right to sanction others – is extremely effective in raising cooperation and welfare due to turnover in the top contributor role and to top contributors' willingness to substantially sanction others. Our findings yield implications for the design of mechanisms intended to foster cooperation in social dilemma environments.

JEL Classification: C72; C91; D02 ; D63.

Keywords: Public Goods Games; Cooperation; Top Contributors; Peer Punishment; Behavioral Mechanism Design.

APA Keywords: Cooperation; Public Good Provision.

PsycINFO Classification code: 2260, 3020.

1. Introduction

Explaining the emergence and sustainability of human cooperation in social dilemma environments, where a strong temptation to free ride on others exists, has long been a core problem for social scientists. In the last years, an increasing number of economic experiments have been contributing to shed light on the issue by investigating the role that *institutions* can play in enhancing cooperation (see e.g. Yamagishi, 1986, and Casari and Plott, 2003). Since in social dilemmas the maximization of social welfare conflicts with individual payoff maximization, the role of sanctioning institutions aimed at penalizing deviant behavior has been extensively explored (Ostrom et al., 1992). On the whole, so far, laboratory studies have concentrated on two broad classes of punitive mechanisms, tackling the problem from two different angles: *decentralized* and *centralized* punishment.

Under voluntary, decentralized punishment, players are usually free to sanction each other arbitrarily and this institutional arrangement turned out to be extremely successful in stabilizing cooperation rates over time, due to many participants' willingness to engage in (costly) punishment of opportunistic behavior (see Fehr and Gächter's (2000; 2002) pathbreaking studies). However, the experimental literature has recently identified a 'dark side' of unrestricted peer punishment, shedding light on some serious drawbacks of this peer-based sanctioning mechanism. First, there are inefficiencies due to lack of coordination among potential punishers (O'Gorman et al., 2009). Second, in many cases this institutional arrangement may undermine the scope for self-governance, as sanctioning may take the form of misdirected, 'antisocial' punishment – that is, low contributors inefficiently meting out sanctions on high contributors (Herrmann et al., 2008; Gächter and Herrmann, 2011; Hauser et al., 2014).[1] As a consequence of the waste in punishment points due to both miscoordination and antisocial punishment, recent work documents that the success of 'vigilante justice' in enforcing cooperation comes at a substantial cost: unless we consider a significantly longer time horizon (Gächter et al., 2008), *average earnings* turn out to be *lower* than in the absence of sanctioning options (Denant-Boemont et al., 2007; Dreber et al., 2008). This is a major shortcoming of

[1] A further problem with discretionary sanctioning is that when multiple stages of punishment are allowed, so that immunity of sanctioners from reprisals is removed, counterpunishment and feuds are likely to be triggered, limiting, once again, successful self-governance and leading, eventually, to a demise of cooperation (Denant-Boemont et al., 2007, Nikiforakis, 2008 and Nikiforakis and Engelmann, 2011).

unrestricted punishment, as it risks turning into a wasteful, inefficient activity for those communities or organizations that adopt it (Nosenzo and Sefton, 2014).

Thus, it would be natural to think that an alternative, viable solution could be to delegate the sanctioning power to a *single, external* enforcer. With a Hobbesian 'Leviathan' entitled to monitor individuals' behavior and wield a 'sword' against free riders, no coordination problems in meting out sanctions would arise. However, even a purely centralized solution appears to be largely unsatisfactory under some important respects. A first reason has to do with the *informational* dimension (see on this also Baldassarri and Grossman, 2011). In many jobs, due to lack of physical proximity, employers cannot observe the exact contribution provided by each worker to the production of total output (Mas and Moretti, 2009). The underlying argument is that in many socio-economic contexts the relevant knowledge is dispersed and a decentralized system is better able to detect it and fulfill its potential, compared to a centralized one. Next, even apart from informational problems, monitoring individuals can be extremely *costly*. In this regard, recent studies also highlight the importance of potentially significant 'hidden costs of control': experimental evidence indicates that many agents reduce their performance as a response to the principal's controlling decision, likely due to agents perceiving the latter being as a signal of distrust and a limitation of their choice autonomy (Falk and Kosfeld, 2006).

In light of such serious drawbacks characterizing 'extreme' (i.e. purely decentralized and purely centralized) punishment-based incentive schemes, in this paper we study the power of an intermediate solution in which the sanctioning power is concentrated in the hands of a *single player* (like in classic centralized mechanisms), but this sole punisher, far from being externally appointed, is *a member of the group* (like in decentralized mechanisms). The reason why we conjectured that this hybrid *peer-to-peer centralized* mechanism could work is that in principle it combines the key advantages of centralized and decentralized sanctioning institutions.

On the whole, our experimental analysis is aimed at studying the effects on cooperation of the introduction of peer-based centralized punishment mechanisms (instead of purely decentralized ones) that differ from one another depending on the criteria used to identify the sole punisher in the groups.

This goal has been pursued by designing an experiment consisting of five treatments. In our baseline, punishment is potentially *widespread*, in the sense that more than one player in each group may sanction others in each period (see Fehr and Gächter, 2000; 2002), while in the other four mechanisms sanctioning power is *concentrated* in the hands of one player only in each group. These four single punisher treatments are characterized by (1) one punisher who is randomly selected in each period (O'Gorman et al., 2009) and is also immune from punishment, (2) one punisher who is selected in each period on the basis of her contribution behavior (i.e., (one of) the top contributor(s)) and is also immune from punishment, (3) one punisher who is selected every four periods on the basis of her contribution behavior (i.e., (one of) the top contributor(s)) and is also immune from punishment and (4) one randomly selected punisher who cannot punish the top contributor, who, therefore, is immune from sanctioning (though not entitled to sanction).

The reason why we opted for this experimental design is twofold. First, we aimed at comparing the performances of alternative peer-based centralized mechanisms based on single punishers. Second, we sought to shed light on single punishers' motivations towards cooperation and sanctioning. To achieve these goals, we first compared our decentralized punishment baseline with two single punisher treatments: one where single punishers are randomly selected (1) and one where only (one of) the top contributor(s) can punish and is immune from punishment (2). This comparison aimed at understanding whether it is passing from decentralized punishment to centralized sanctioning based on single punishers *per se* that makes the difference or whether the *specific criterion* through which single punishers are selected (i.e. random vs. contribution-based) also matters. Then, by comparing (2) and (3), i.e., a treatment in which only (one of) the top contributor(s) can punish but, unlike in (2), punishment and immunity last for four periods, we try to distinguish between two broad classes of explanations: are players actually motivated to significantly contribute to the (first-order) public good and punish low contributors or are they driven by the desire to become immune from punishment and/or enjoy punishment *per se* (i.e., regardless of its being targeted at lower or higher contributors)? Finally, in order to disentangle search for immunity from willingness to punish, we introduced a treatment (4) in which a randomly selected player in each period is allowed to sanction others but this right cannot be exerted against the top contributor, who therefore is immune.

Our findings indicate that the 'Top Contributors as Punishers' mechanism effectively fosters cooperation and results in a level of earnings equal to that of the mechanism based on random selection of the punisher (1) and higher than that of the unrestricted, decentralized punishment mechanism (that is our baseline). Next, the comparison between (2) and (3) revealed that the success of the ‘Top Contributors as Punishers’ institutional arrangement occurs *despite* the widespread presence of players who, once in power, both free ride on others’ contributions and punish to a significant extent not only low contributors but also high contributors (i.e. engage in so called antisocial punishment). Thus we could exclude that the success of our mechanism is driven by a significant presence of subjects who are willing to act in favor of the group. The most plausible explanatory factors were instead the willingness to punish others and the search for immunity motive. In this regard, our findings from the latter treatment, i.e. (4), led us to rule out that individuals’ race to become the top contributor in our core treatments is driven by the search for immunity motive.

The remainder of the paper is structured as follows. Section 2 reviews the closely related literature. Section 3 illustrates the experimental design. In Section 4 we report and discuss our core findings. Section 5 concludes the paper.

2. Related literature

As to prior experiments based on exogenous restrictions on the number of peer punishers, the most closely related work is the following. Three papers (O’Gorman et al., 2009; Carpenter et al., 2012, and Nosenzo and Sefton, 2014), in line with our study, examine the finitely repeated public goods game framework and depart from Fehr and Gächter’s (2000) seminal article by exogenously imposing that *only one punisher per group* can emerge in each round. The presence of such restriction significantly differentiates these papers from most previous work on peer punishment.[2]

A potential advantage of a single punisher mechanism is clearer accountability and lack of coordination problems at the punishment level (O’Gorman et al., 2009; Nosenzo and Sefton, 2014). O’Gorman et al. (2009) explore a solution in which responsibility for punishment is borne by one specific, designated and randomly selected individual. They find that, under a single punisher institution, cooperation can be successfully sustained and reaches levels comparable

[2] Other experimental work based on restricted sanctioning (though not on the appointment of single punishers) includes Carpenter (2007a) and Eriksson et al. (2013). For a recent study focusing on a stochastic two-person prisoner’s dilemma environment where only cooperators can punish non-cooperators, see Xiao and Kunreuther (2016).

with that maintained when punishment is potentially widespread. Further, group earnings levels are higher: a sole punisher solution reduces inefficient losses, as sanctioning efforts are not unnecessarily duplicated. Carpenter et al. (2012) also have a treatment where the monitoring and sanctioning power is concentrated in the hands of one randomly selected group member. However, unlike O'Gorman et al. (2009), they show that under this form of restricted punishment both contributions and average earnings are *lower* than under unrestricted sanctioning. The third closely related study (Nosenzo and Sefton, 2014), by means of a five-treatment experimental design, analyzes both reward and punishment schemes, under both centralized and decentralized systems. In one of their central monitoring treatments, all punishment power is concentrated in the hands of one, randomly assigned group member and this was the same group member in all ten periods. The authors show that concentrating punishment power *reduces* the effectiveness of punishment, compared to (unrestricted) peer punishment. Hence, their findings are similar to Carpenter et al.'s results: in both papers, unlike in O'Gorman et al., concentrating the power in the hands of a single punisher is not an effective means to sustain cooperation. Therefore, these three papers on the whole present mixed evidence on the effectiveness of single punishers in raising cooperation and earnings levels.

A further related paper is Andreoni and Gee (2012), who investigate an enforcement device based on authority delegation to a third party which they term the 'hired gun' mechanism. Compared to classic unrestricted peer punishment, the hired gun mechanism acts as a low cost device that fosters cooperation and welfare. Unlike their mechanism, in our central treatment we focus on a peer-to-peer punishment institution, rather than an external third party.[3]

Our study has some similarity also with recent experimental work focusing on the *legitimacy* of sanctioning institutions. Faillo et al. (2013) investigate a sanctioning mechanism where access to sanctioning power is restricted in each group as they exogenously impose that, in each round, player *i* can punish player *j* only insofar as *i* contributes more than *j* in the same round (see on this also Farjam et al., 2015).[4] The present paper differs from Faillo et al. (2013) along two key dimensions. First, we pass from the analysis of (potentially) multiple punishers to

[3] A second key difference between our design and Andreoni and Gee's study is that, as we clarified above, in two of our treatments, punishers, far from being randomly selected, have to *obtain* this right by being the highest contributors in their group.

[4] This article presents the results of an agent-based simulation aimed at investigating the effectiveness of different punishment mechanisms (included the legitimate punishment institution studied by Faillo et al., 2013) in fostering cooperation within a population of adaptive agents in which indirect reciprocity is already at work.

a *single punisher* institution in fostering cooperation and welfare, by comparing our findings with the ones obtained in the aforementioned stream of literature focusing on single punisher mechanisms, where no coordination problems arise but only one player at a time has to carry the burden of the second-order public good (i.e. punishment) provision. Secondly, as we made clear in the Introduction, another important focus of the current study is the analysis of individual players' motivations underlying their contribution and punishment behavior throughout the game.

Since our central 'Top Contributors as Punishers' mechanism tests the idea that a race occurs among group players to become the highest contributor in each round, our work also relates to the stream of literature on contests that, in the last years, has been dealing with the role of *competition* to mitigate free riding in social dilemma environments. In fundraising, it is common to have charities organizing raffles or auctions to raise money and awarding non-pecuniary prizes (such as the renaming of a building) to their most generous donors only. So far, experimental studies have investigated – mainly in the lab, but also, more recently, in the field – a series of prize-based fund-raising schemes – from all-pay and winner-pay auctions to lotteries. On the whole, they provide mixed evidence over the relative performance of alternative mechanisms in raising contributions to a public good (Carpenter et al., 2008; Schram and Onderstal, 2009; Onderstal et al., 2013, Samek and Sheremeta, forthcoming).[5] First-price all-pay auctions – in which the player exerting the highest effort wins the prize with certainty (unlike in lotteries) and all bidders have to pay their bid – appear to be the fund-raising scheme that mostly resembles our 'Top Contributors as Punishers' mechanism. However, it is important to highlight that the key difference between the two schemes is that while in all-pay auctions the highest contributor is awarded a *prize*, in our institution the top contributor is immune from others' sanctioning and earns the right to incur *costs* to sanction others.

[5] Our work has also some similarities with the large literature on *tournament-based* incentive schemes in organizational settings. In particular, we are close to the body of research exploring the effectiveness of reward-based rank-order tournaments extensively used in real world settings by managers as motivational tools to encourage employees to compete and work harder than their colleagues (Konrad, 2009). In so called 'reward tournaments', the most productive agent in the team receives the best prize, usually consisting of bonuses (for example, for the 'employee of the month') or promotions (Lazear and Rosen, 1981).

3. Experimental design

3.1. Treatments

Our experiment consists of the five treatments illustrated in the Introduction and described below. In all of them, participants played a finitely repeated public goods game with punishment options for 20 periods. In every period, the experimental game consists of two decision stages: at stage 1 (contribution stage), players choose how much to contribute to the public good and, at stage 2 (punishment stage), they have access to punishment options. In each treatment, participants are informed about these features of the game to be played.

Our starting point is the comparison between the performances of a *decentralized* vs. *centralized* peer punishment institution. Like in O'Gorman et al. (2009), we compare a baseline with unrestricted punishment that replicates Fehr and Gächter's design (2000) (hereafter, UP) with a treatment in which in each round, for each group, only one of the members is randomly assigned the right to punish the other players (random single punisher treatment; hereafter, SR).

Next, we are interested in comparatively investigating the performances of different *peer-based centralized* mechanisms, in which punishment power is concentrated in the hands of one group member only. In particular, we extend to peer-based centralized settings the idea that punishment has to be 'legitimate' (Faillo et al., 2013), comparing a random punisher selection mechanism to an endogenous selection mechanism in which subjects must *gain the right to punish* (and, consequently, are granted *immunity* from others' punishments), by contributing more than the other members of their group (restricted single punisher or 'Top Contributors as Punishers' mechanism; hereafter, TOP).

Given the nature of our key selection mechanism, in the restricted single punisher (TOP) treatment we might observe a higher average contribution than in the random single punisher (SR) either because subjects are motivated to contribute and punish lower contributors, aiming at increasing the level of cooperation of the group, or because of subjects' desire to become immune from punishment and/or enjoy punishment of others per se (i.e. regardless of its being targeted at lower or higher contributors).

In order to discriminate between these two alternative explanations, we introduce a fourth treatment in which the right to punish is assigned to the top contributor (like in TOP), but lasts for *four periods* rather than for one period only (we label this treatment as FOUR).[6] If, after being appointed, in the following three periods in which the entitled punisher is in charge, (1) her contributions significantly drop and (2) punishment points are not used to increase the level of contribution of the group, then finding higher contributions in the TOP treatment should be explained as the consequence of the desire to gain immunity and/or enjoy punishment of others *per se*, rather than by the willingness to act at a cost in favour of their group and contribute to the provision of the public good in the two decision stages over time.

Next, to shed further light on players' motivations and disentangle these two motives (i.e. search for immunity and willingness to sanction others), by keeping direct comparability with the SR and the TOP treatments, we run a fifth treatment, that we termed Random TOP. As in SR, the punisher is randomly selected; however, she cannot punish the member who gave the highest contribution of the group. If the level of contribution observed in this treatment is the same as in SR, then we can exclude that immunity is a relevant driver of the behavior observed in TOP: to increase their probability to become immune, subjects must increase their contribution, and this probability is equal to one when they contribute all their endowment. If the search for immunity is relevant, we should then observe, on average, a level of contribution significantly higher than that of the SR treatment.

In the next subsections, the five treatments are presented in detail.

3.1.1. Unrestricted punishment: UP

In the UP treatment, punishment is unrestricted and subjects are provided with full information, that is there is feedback about *all* their group co-players' individual contributions. This is a replication of the standard linear *VCM* with punishment and partner protocol (Fehr and Gächter, 2000), where everyone can freely punish everyone else in the group. In stage 1, each participant receives a fixed amount e =20 of tokens and has to decide whether she wants to invest or not an amount $g_i \leq e$ into a public project. Decisions are made simultaneously and with no

[6] The mechanism at work in FOUR captures the key features of social environments in which top contributors only have the right to sanction but, unlike in TOP, they are in charge for a given period of time. This allows us to check whether top contributors contribute a lot and punish players who contribute less than they do also in non-key periods (i.e. in periods different from the ones in which they have access to punishment power). For example, Kosfeld and Rustagi's (2015) experimental field evidence indicates that, once in power, group leaders engage not only in punishment of free riders, but also, to a significant extent, in *antisocial punishment* (i.e. punishment of cooperators).

information about peers' choices. At the end of stage 1, each participant is informed about her current earnings, which are calculated by the computer in the following way:

$$\pi_i = (20 - g_i) + 0.4\sum_{j=1}^{4} g_j$$

In stage 2, subjects are informed about the contribution by the other members of their group and can decide to assign between 0 and 10 punishment points to any of them. Points assignment is costly and costs are charged according to a convex cost function as in Fehr and Gächter's study (Table 1).

[TABLE 1]

Each point that a subject receives reduces her earnings at stage 1 by 10%, with 100% as the maximum total reduction. Punishment is anonymous: subjects who get punished do not know the identity of the punisher. Each participant's net earnings at the end of stage 2 are given by her earnings at the end of stage 1 minus the costs of assigned and received punishment points: they are calculated by the computer and each participant sees her cumulative net earnings on the screen at the end of each round.

3.1.2. Randomly selected single punisher (SR)

The SR treatment differs from UP as here punishment power is concentrated in the hands of *one subject* only in each round. As before, subjects are provided with full information on the contribution levels of their peers in the group. However, in the punishment stage only one *randomly selected* participant obtains the right to punish their peers. The random draw is repeated in each period, potentially assigning sanctioning power to different subjects over time (like in O'Gorman et al., 2009). Like in UP, the punisher can decide to assign between 0 and 10 punishment points to any peer and costs are charged according to the same cost function.

3.1.3. Restricted single punisher (TOP): the 'Top Contributors as Punishers' mechanism

The key feature of the TOP treatment is that *in each round* only the 'Top Contributor' is allowed to sanction her group co-players.[7] A participant is classified as Top Contributor when her contribution is the highest in the group ($g_i > g_{-i}$). In this treatment, in stage 2, the Top

[7] It is important to make clear that, in order to minimize so called 'experimenter demand effects' (Zizzo, 2010), we never used loaded terms such as 'punishment', 'top contributor' and 'free riding' during the experiment.

Contributor *only* is given the opportunity to punish other subjects in the group by assigning a certain amount of points.

In case the highest amount is contributed identically by two or more subjects, only one[8] of them gets randomly selected to make a punishment choice that will be actually implemented.[9] Like in the previously illustrated two treatments, all subjects receive information about their peers' contribution behavior. The Top Contributor who gets drawn might decide to actually assign up to 10 points to each co-player, with punishment costs being charged according to the previously illustrated cost function. Each participant knows that she can go on with stage 2 in the experiment only if she is the Top Contributor at stage 1 and (in case two or more individuals contribute the highest amount) gets drawn, that is only if she obtains the right to sanction others.

Real-life examples of groups appointing only one peer at a time as a potential punisher contingent on her prior behavior and changing her over time include horizontal relationships in the workplace and peers' interaction in web communities. In labor environments, the highest ability worker in a team often has the authority (and sometimes even the formal legitimacy – e.g. when he becomes the 'team leader') to sanction the coworkers who exert low effort and jeopardize the achievement of the team's goals.[10] Next, as shown by Mas and Moretti's (2009) field study, low productivity workers are sensitive to the presence of a higher productivity coworker who can observe them. In commenting on their findings, the authors point to the role of *social pressure*, viewed as a force that can partially internalize free riding externalities. It is also important to note that, in firms and other organizations, powerful persons who can sanction others at a given time might easily lose their power, as it is the case with short-term contracts or performance-dependent project leadership (Dorrough et al., forthcoming).

[8] We chose not to have *all* the top contributors deciding about punishment (and then select randomly the choice to be implemented) because we wanted to make the single punisher decision more salient. The single punisher, when she is appointed, knows for sure that she can punish an that she is immune from punishment.

[9] The other highest contributors are asked to answer a hypothetical question asking what they would do in terms of punishment, knowing that their decision won't produce real consequences in the game.

[10] Many online initiatives provide further examples along these lines. While a very large community of users voluntarily develops and maintains the pages of Wikipedia (the most world famous internet encyclopedia), only the top contributors (that is the volunteers who are active and regular Wikipedia contributors for at least several months, have considerable experience and, therefore, are expected to have the trust of the community) are likely to become administrators, i.e. editors who hold sanctioning power as they have been granted the ability to block user accounts and IP addresses from editing pages and other actions. The Wikipedia website clearly states that administrators are not external subjects but members of the same community of users.

3.1.4. Restricted single punisher for four periods (FOUR)

Like TOP, the FOUR treatment is characterized by the presence of a Top Contributor that is the only subject who is allowed to sanction her group co-players. What specifically characterizes this treatment is that here the Top Contributor obtains this right not only for the current period, but also for the following three periods. In other words, subjects on the whole face five 'key' periods (namely periods 1, 5, 9, 13, and 17) – out of the 20 periods of interaction – where they become the Top Contributor and, when they do, they earn the right to be the sole punishers for four consecutive periods. As a consequence, for those four periods they are also *immune* from others' punishments. As in the TOP treatment, when in key periods the highest amount is contributed identically by two or more subjects, only one of them gets randomly selected. In developing countries, developmental programs increasingly rely on local participation, with local leaders getting power over groups for a certain period of time. Kosfeld and Rustagi (2015) illustrate a large forest commons management program launched in 2000 in Ethiopia. The program involved 56 different forest user groups that were given complete jurisdiction to manage their forests as a common property resource. In these groups, group members elect a leader – who typically is experienced, has a dominant personality and is in charge for a period of at least five years. Once in charge, the leader, though he is not materially incentivized, is expected to sanction rule violators to enforce cooperation.

3.1.5. Randomly selected restricted single punisher (Random TOP)

This treatment has been introduced to shed further light on the search for immunity as a motivation that potentially underlies high levels of contribution. As in both the SR and the TOP treatments, punishment power is attributed to one subject only and lasts for one period only. However, the potential punisher is selected at random (as in SR and differently from TOP) but cannot punish the top contributor (as in TOP and differently from SR). Subjects who wish to be immune from peers' punishment should choose high levels of contribution in order to reach the position of top contributor, since the top contributor is the only peer that the randomly selected punisher cannot hit.

Figure 1 reports a synthetic description of each treatment and the links between the five treatments, highlighting the specific differences across them. Only treatments connected by an arrow are directly comparable. The numbers on the arrows refer to the variables whose effect is measured by comparing two connected treatments: 1) centralization vs. decentralization of punishment activity; 2) random selection of punisher vs. selection based on being (one of) the top contributor(s); 3) punisher turnover in each round vs. punisher turnover every four rounds; 4) immunity for the top contributors vs. immunity for the punisher only.

[FIGURE 1]

3.2. Procedures

A total of 276 subjects participated voluntarily in the experiment at the CEEL Lab of the University of Trento. Fifteen sessions were conducted between November 2009 and November 2016. We run three sessions for each treatment: for Random TOP we run two sessions with 16 subjects and one with 20 subjects; for all the other treatments, we run two sessions with 20 subjects and one with 16 subjects.[11]

The experiment was programmed by using the z-tree platform (Fischbacher, 2007). Subjects were undergraduate students (55% are students of Economics and or Management, 45% females, 85% Italians). We employed a between subjects design: no individual participated in more than one session. In each session, participants were paid a 5 euros show-up fee, plus their earnings from the experiment. The average payment per participant was 14.55 euros (including the show-up fee) and the sessions averaged approximately 90 minutes. At the beginning of each session, participants were welcomed and asked to draw lots, so that they were randomly assigned to terminals. Once all of them were seated, the instructions were handed to them in written form before being read aloud by the experimenter.[12] Participants had to answer several control questions and we did not proceed with the actual experiment until all participants had answered all questions correctly.

For each treatment, participants in each session were randomly assigned to groups of size N=4, so that they did not know the identities of the other members of their group. Like other experimental studies (see e.g. Cinyabuguma et al., 2006; Denant-Boemont et al., 2007), we used a partner protocol that kept the composition of each group constant over rounds, so that, at the

[11] UP data are the same used in Faillo et al. (2013).

[12] A translation of the instruction sheet for treatment TOP is provided in Appendix A. Original instructions were written in Italian. They are available upon request from the authors.

end of each period, individuals remained in the same group. The reason why we used a partner design is that repeated interaction is a typical feature of many real world settings (e.g., businesses or web-based communities) in which sanctioning often takes place (Xiao and Houser, 2011). However, individuals' labels were randomly reassigned in each period. For example, the same player could be labeled as player *32* in period *t*, as player *5* in period *t* + 1, and as player *43* in period *t* + 2. Therefore, our partner protocol was also characterized by anonymity of the members of the group and change of participants' labels across rounds, to prevent individual reputation formation throughout the game. The parametric structure of the experiment is based on Fehr and Gächter (2000).

4. Results

This section presents our experimental evidence on contribution levels in the five treatments and organizes it through three propositions summarizing our results on (1) centralization of punishment power, (2) turnover of punishment power, and (3) role of immunity, respectively. Then, we illustrate our findings on punishment activity and earnings.

4.1. Contribution levels

Figure 2 displays the pattern of average contributions by period in the five treatments, while Table 2 reports average contributions of groups across treatments (see also Figure 1B in Appendix B).[13]

[FIGURE 2]

[TABLE 2]

We start the analysis by focusing on the effects of centralization of the punishment mechanism and of the criterion used to identify the punisher (arrows 1 and 2 in Figure 1). This implies that we are interested in comparing the levels of cooperation across UP, SR and TOP.

[13] We observe a decrease in the level of contributions in the very last rounds. The same pattern has been observed in several experiments on linear public goods with punishment conducted in the last years. See for example Fehr and Gächter (2000), Masclet et al. (2003), Bochet et al. (2006), Denant-Boemont et al. (2007), Carpenter and Matthews (2009), Ertan et al. (2009), Gächter et al. (2012), Herrmann et al. (2008) and Nikiforakis et al. (2012).

We find that there is no significant difference between the level of contributions of UP and SR (Wilcoxon-Mann-Whitney two-tailed test taking average contributions of groups as independent observations:[14] UP vs. SR: z=0.55, p=0.58; see also Table 2)[15], while mean contributions in TOP are *significantly higher* than those in both UP and SR (TOP vs. UP: z=2.38, p=0.017; TOP vs. SR: z=2.27, p=0.02)

Result 1. The effects of centralization of punishment activity. *Taking the unrestricted decentralized sanctioning mechanism (treatment UP) as benchmark, appointing a single punisher randomly in each round (treatment SR) does not affect the level of cooperation, but, when the single punisher is selected among the top contributors of the group (treatment TOP), the level of cooperation increases significantly.*

The next comparison is between the level of contribution in the TOP treatment and that in FOUR (arrow 3 in Figure 1). This allows us to test whether the high level of contribution in the TOP treatment can be explained by the presence of subjects who are willing to contribute and sustain cooperation of the group or by motives like the desire to acquire the right to punish others and/or gain immunity from punishment. We find that the average contribution in TOP is significantly higher than in FOUR (TOP vs. FOUR: z=2.34, p=0.02).

As shown in Figure 2, treatment FOUR displays a peculiar "up-and-down" pattern reflecting the fact that subjects concentrate their contribution efforts in the five 'key periods' and – on average – *substantially drop* their level of contribution in the subsequent three periods: the average contribution in key periods is 13.09 vs. 7.39 in non-key periods (z=3.27, p=0.001). In Tables 3 and 4 we provide a detailed investigation of the dynamic of contributions of subjects selected as punishers in the three treatments with centralized peer punishment.

[TABLE 3]

[14] In the remainder of the analysis, as we will always use this type of tests, we will report only the value of z and p.

[15] The level of contribution of our UP treatment is not particularly high, and in general punishment seems not to be very effective in this treatment. To check for this, we have run also a standard finitely repeated linear public good without the punishment phase (No Punishment) with the same length, endowment and MPCR of the other treatments. The level of contribution observed in this treatment is not significantly different from that in UP and in SR (the two treatments that can be directly compared with it). Mann-Whitney-Wilcoxon two-tailed test with group averages as independent observations: No Punishment vs UP: z = 1.019, p = 0.308; No Punishment vs SR: z = 1.413, p = 0.157).

[TABLE 4]

Looking at the behavior of subjects selected as punishers in the FOUR treatment, we observe that their average contribution in the key periods is not significantly different from the average contribution of the punishers in the TOP treatment (z=1.51, p=0.13); in contrast, in the remaining (non-key) periods, when they are immune from punishment and can assign points independently of their contribution, their average contribution drops to 6.02, which is not significantly higher than the average contribution of the subjects not entitled to be selected as punishers (4.21 tokens; z=0.02, p=0.92).

Both in the TOP and in the FOUR treatment, when there is more than one subject entitled to become the punisher – i.e. more than one member who contributes the maximum amount in the group –, only one of them gets randomly selected as the actual punisher.[16] We observe that the average contribution of non-drawn but entitled subjects drops from 20 in key periods to 11.33 in non-key periods, whereas the average contribution of selected punishers drops from 18.1 to 4.61: both categories of subjects reduce their contribution levels, but the reduction of non-drawn subjects is smaller.[17] A possible explanation is that, unlike selected punishers, non-selected ones know that they are not immune from punishment.

Result 2. The effect of the reduction of punishers' turnover. *Average contributions in the top contributor as single punisher (TOP) is significantly higher than that in the top contributor as single punisher for four periods (FOUR). The average contribution of subjects who have gained the right to punish by becoming the top contributors in key periods decreases significantly in non-key periods.*

[16] In the TOP treatment, considering the 20 periods and the 14 groups, and excluding the 27 cases in which all the members of the group gave the same contribution, the cases with more than one top contributor were 90. The average number of entitled but non-selected subjects in TOP was 6.75 (0.48 per group) per period.

[17] In treatment FOUR, considering the key periods and the 14 groups, the cases with more than one and less than four top contributors were 35. In all these cases the contribution of the entitled subjects was equal to 20. The cases in which all the members gave the same contribution were 10. The average number of entitled but non-selected subjects in FOUR was 10.6 (0.75 per group) per period (considering the five key periods).

Result 2 then suggests that the high level of contribution observed in TOP seems to be due a race to become the top contributor in order to gain the right to punish others and/or to become immune from others' punishment.

The final comparison is between the average contribution in the Random TOP treatment and that of both the SR and the TOP treatment (arrows 2 and 4 in Figure 1). As we have explained in section 3.1, this comparison allows us to investigate the role of search for immunity in the explanation of the level of cooperation observed in the TOP treatment.

We observe that the level of contribution in Random TOP is not significantly different from that of SR (z=0.77, p=0.43) while it is significantly lower than that of TOP (z=2.76, p<0.01).

Result 3. The role of immunity as driver of high contributions in the TOP treatment. *Average contributions in the Random TOP treatment are significantly lower than in the TOP treatment while they are not significantly different from those of the SR treatment.*

Result 3 suggests that we can rule out immunity as an explanation of the high level of contribution observed in the TOP treatment.

All these results are supported also by a random effect GLS estimation (Table 5).

[TABLE 5]

4.2. Punishment behavior

With regard to the distribution of punishment points (Figure 3; see also Figure 2B in Appendix B), in all treatments we observe the decreasing pattern already detected in prior work.[18] In SR, the average amount of points assigned within groups is significantly lower than both in UP and in TOP (UP vs. SR: z=2.50, p=0.01; TOP vs. SR: z=3.31, p=0.00). There is no significant difference in the amount of points distributed neither between UP and TOP (z=1.35, p=0.17) nor between TOP and FOUR (z=0.82, p=0.41). The amount of points assigned in

[18] The same pattern has been observed, for example, by Denant-Boemont et al. (2007), Carpenter (2007b) and Nikiforakis et al. (2012).

Random TOP is not significantly different from that assigned in TOP (z=0.48, p=.62), while it is higher than that assigned in SR (z=3.44, 0, p< 0.01).

[FIGURE 3]

In case of equally ranked top contributions, in TOP the Top Contributors who are drawn and get the concrete chance of punishing do assign an average of 2.59 punishment points, whereas the average number of hypothetical punishment points virtually assigned by non-drawn Top Contributors is 1.99. In FOUR, the Top Contributors who are drawn and get the concrete chance of punishing do assign an average of 3.44 punishment points, whereas the average number of hypothetical punishment points virtually assigned by non-drawn Top Contributors is 1.44. This indicates that, not surprisingly, in both treatments Top Contributors' propensity to punish is higher when they know that their assigned punishment points have a *real*, rather than virtual, nature.

4.2.1. Antisocial punishment

When punishment activity is not restricted a non-negligible share of punishment points can be classified as 'antisocial', since they are directed to subjects who contribute *more than the punisher or the same amount as the punisher* (see on this e.g. Herrmann et al., 2008). Antisocial punishment turns out to be a widespread, quantitatively relevant phenomenon in our experiment: considering the total amount of points distributed in each treatment, the percentage of antisocial points is 19,5% in UP, 30% is SR and 45.2% in FOUR.

In order to compare the Random TOP treatment with SR, we must exclude from the computation the top contributors of the former, since we must consider only the members of the group who can actually assign antisocial points. Differently from what happens in SR, in this treatment the top contributor, when she is selected, cannot assign points to other top contributors. The percentage of antisocial punishment points assigned in the Random TOP treatment is 48.5% of the total amount of points distributed within the groups in the twenty periods (see Table 1B in Appendix B for detailed data).

These findings interestingly reveal that a potential ‘enemy’ of cooperation such as misdirected, antisocial punishment manifests itself as an important phenomenon not only when discretionary punishment is decentralized (UP), but also when one (randomly selected) participant at a time is free to sanction others (SR and Random TOP) and even more so when the right to punish *entails being the highest contributor* in stage 1 in key periods and punishment (including antisocial punishment) can freely occur in subsequent periods (FOUR). When they are given the possibility to punish antisocially, as it is the case in SR and FOUR, a significant number of single punishers do actually exploit this opportunity.

Available experimental evidence indicates that antisocial punishment is quantitatively significant. Anderson and Putterman (2006) show that a surprisingly large share of punishment events (30%) involve low contributors (i.e. subjects who contributed less than the average) punishing high contributors (i.e. subjects who contributed more than the average). Herrmann et al. (2008) make clear that antisocial punishment behavior strongly differs among societies, ranging from very little antisocial punishment in some participant pools (e.g. US, Australia and UK) to a level of antisocial punishment close to the level of punishment targeted at lower contributors (Greece and Oman). As we noted in Section 3, Kosfeld and Rustagi (2015) provide experimental field evidence that also within large forest commons management programs, like the one they analyze in Ethiopia, *antisocial leaders* emerge, in the sense that they punish individuals who contributed their full endowment to the public good.

But why do people engage in antisocial punishment? Herrmann et al. (2008) suggest that revenge is a plausible explanation: low contributors, who tend to be punished by high contributors, may decide to sanction the latter as a form of (blind) revenge. However, our data lead us to rule out that antisocial punishment in our experiment is due to blind revenge, as the regression summarized in Table 6 reveals that the percentage of antisocial punishment points that a subject assigns in a specific period (computed as the ratio between antisocial punishment points and the total amount of assigned punishment points she assigns in that period) does not significantly depend on the number of punishment points she has received in the previous period.[19]

[TABLE 6]

[19] An estimation with the absolute number of antisocial punishment points as dependent variable, with the inclusion of the absolute number of points received as regressor, provides very similar results, that are available upon request.

4.3. Average earnings

We conclude the analysis of the experimental results by comparing the average earnings across treatments. Taking group cumulative average earnings as independent observations, we observe that average earnings in TOP are significantly higher than in UP (TOP vs. UP z=2.11, p=0.03) and in FOUR (TOP vs. FOUR: z=2.20, p =0.02). Average earnings in SR are slightly greater than those in UP (z=1.65, p=0.09). Earnings in TOP are not significantly different from those in SR (z=0.82, p=0.41) Earnings in Random TOP are smaller than both the ones in TOP (z=2.47, p=0.01) and in SR (z=1.89, p=0.06) (see Figure 3, see also Figure 3B in Appendix B).

[FIGURE 3]

The better performance – in terms of earnings – of the SR treatment can be explained by recalling the significantly lower use of costly punishment that characterizes this treatment.

5. Discussion and conclusion

Our findings indicate that the 'Top Contributors as Punishers' institution with frequent turnover (TOP) works well in terms of cooperation and punishment of low contributors. Next, we show that, in FOUR, a decay phenomenon occurs, with top contributors *suddenly* and *significantly* decreasing their contribution levels in non-key rounds.[20] Moreover, they punish a lot and, despite being the highest contributors in key periods, surprisingly display a very large amount of antisocial punishment in non-key periods. Therefore, our results from FOUR interestingly reveal that the 'Top Contributors as Punishers' mechanism in TOP is successful *despite* the widespread presence of players who suddenly and significantly reduce their contribution levels (in non-key periods) and mete out a large amount of antisocial punishment.

In an attempt to disentangle the search for immunity motive from willingness to sanction others, we have analyzed the Random TOP treatment. By comparing our findings from this treatment with the results from SR and TOP, we showed that while search for immunity does not appear to be an important driver of individuals' decisions in the game (as contribution levels in

[20] This finding is similar to the one obtained by Orzen (2008) within a social dilemma setting in which a first-price all-pay auction is at work: his data indicate that, after winning the prize, subjects' average contributions *significantly drop* in the subsequent round.

Random TOP are not significantly different from the ones in SR and significantly lower than the ones observed in TOP), the willingness to punish motive seems to play a role.

Therefore, on the whole, our results reinforce and extend to centralized sanctioning institutions based on single punishers the discovery, advanced by previous experimental studies, that individual behavior crucially depends on the specific features of the sanctioning system at work (see e.g. Ambrus and Greiner, 2012, Xiao, 2013 and Faillo et al., 2013) and that punishment opportunities are not always socially beneficial (Herrmann et al., 2008; Nikiforakis, 2008).

In particular, our central treatment reveals that a centralized punishment mechanism assigning punishment power to top contributors only from round to round successfully raises cooperation, compared to single punisher institutions where antisocial punishment is permitted. This interestingly occurs despite the fact that the 'Top Contributors as Punishers' mechanism is a demanding institution that entails relevant monetary costs with regard to both the first-order and the second-order public good to be provided (i.e. contributions and punishment). Previous experimental work indicates that, within the public goods game framework, a design that makes top contributions salient to the players is not sufficient *per se* to foster cooperation. Dale and Morgan (2010) document a perverse *negative* effect of asking individuals to contribute the socially optimal amount, as this tool turns out to be, at best, ineffective. Similarly, Samek and Sheremeta (2014) show that recognizing only the highest contributors, by displaying their identities, does not increase contributions. Even our FOUR treatment confirms that it is not the salience of the top contributor role by itself that raises cooperation, as here many top contributors both suddenly decrease their contributions and display a significant amount of antisocial punishment in non-key rounds.

More generally, we view our results as consistent with Ostrom et al.'s (1992) claim that policy-makers responsible for the governance and management of common pool resources should not presume that individuals facing social dilemma situations are caught in inexorable tragedies from which there is no escape. Our findings yield implications for the design of mechanisms intended to foster cooperation and challenge the Hobbesian view that an external enforcer is always necessary to grant cooperation and make self-governance possible.

Acknowledgements: We thank Elisa Cavatorta, Antonio Fasano, Enrique Fatas, Natalie Gold, Shaun Hargreaves-Heap, George Loewenstein, Arnout van de Rijt, Alessandro Tavoni, Mirco Tonin, the London Behavioral and Experimental Economics Group, seminar participants at LUISS University in Rome, King's College London and at the Free University of Bozen, participants in the 2014 IMEBESS Conference in Oxford and in the 2015 Experiment a BIT Workshop in Trento for useful suggestions and comments. The usual caveats apply. We gratefully acknowledge the University of Verona (2011 Joint Projects on "Punishment and Decision-making: Neuroeconomic Foundations, Behavioural Experiments and Implications for Law and Economics"), the Department of Economics and Management of the University of Trento and the Department of Economics of the University of Verona for financial support.

References

Ambrus, A., Greiner, B. (2012). Imperfect public monitoring with costly punishment: an experimental study. American Economic Review, 102 (7), 3317-3332.

Anderson, C.M., Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior, 54 (1), 1-24.

Andreoni, J., Gee, L.K. (2012). Gun for hire: Delegated enforcement and peer punishment in public goods provision. Journal of Public Economics, 96, 1036-1046.

Baldassarri, D., Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. Proceedings of the National Academy of Science, 108, 11023-11027.

Bochet, O., Page, T., Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior and Organization, 60 (1), 11-26.

Carpenter, J. (2007a). Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. Games and Economic Behavior, 60, 31-51.

Carpenter, J. P. (2007b). The demand for punishment. Journal of Economic Behavior and Organization, 62 (4), 522-542

Carpenter, J., Holmes, J., Matthews, P.H. (2008). Charity auctions: a field experiment, Economic Journal, 118 (525), 92-11.

Carpenter, J., Matthews, P. H. (2009). What norms trigger punishment? Experimental Economics, 12(3), 272–288. Carpenter, J., Kariv, S., Schtter, A. (2012). Network architecture, cooperation and punishment in public good experiments. Review of Economic Design, 16 (2), 93-118.

Casari, M., Plott, C. (2003). Decentralized management of common property resources: experiments with a centuries-old institution. Journal of Economic Behavior and Organization, 51, 217-247.

Cinyabuguma, M., Page, T., Putterman, L. (2006). Can second-order punishment deter perverse punishment? Experimental Economics, 9 (3), 265-279.

Dale, D.J., Morgan, J. (2010). Silence is golden. Suggested donations in voluntary contribution games. Mimeo.

Denant-Boemont, L., Masclet, D., Noussair, C.N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic Theory, 33, 145-167.

Dorrough, A., Glockner, A., Lee, B. Race for power in public good games with unequal, unstable punishment power. Journal of Behavioral Decision Making, forthcoming.

Dreber, A., Rand, D.G., Fudenberg, D., Nowak, M.A. (2008). Winners don't punish. Nature, 452, 348-351.

Eriksson K., Strimling P., Ehn M. (2013). Ubiquity and efficiency of restrictions on informal punishment rights. Journal of Evolutionary Psychology, 11, 17-34.

Ertan, A., Page, T., Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. European Economic Review, 53 (5), 495-511.

Faillo, M., Grieco, D., Zarri, L. (2013). Legitimate punishment, feedback, and the enforcement of cooperation. Games and Economic Behavior, 77, 271-283.

Falk, A., Kosfeld, M. (2006). The hidden costs of control. American Economic Review, 96 (5), 1611-1630.

Farjam, M.D., Faillo, M., Haselager, W.F.G. and Sprinkhuizen-Kuyper, I.G. (2015) Punishment Mechanisms and their Effect on Cooperation - A Simulation Study, Journal of Artificial Societies and Social Simulation, 18 (1).

Fehr, E., Gächter, S. (2000). Cooperation and punishment in public goods experiments. American Economic Review, 90 (4), 980-994.

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. Nature, 415, 137-140**.**

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics, 10, 171-178.

Gächter, S., Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. European Economic Review, 55 (2), 193-210.

Gächter, S., Renner, E., Sefton, M. (2008). The long-run benefits of punishment. Science, 322, 5907, 1510.

Hauser, O.P., Nowak, M.A., Rand, D.G. (2014). Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible. Journal of Theoretical Biology, 360, 163-171

Herrmann, B., Thoeni, C., Gächter, S. (2008). Antisocial punishment across societies. Science, 319, 1362-1367.

Konrad, K.A. (2009). Strategy and dynamics in contests, Oxford, Oxford University Press.

Kosfeld, M., Rustagi, D. (2015). Leader punishment and cooperation in groups: experimental field evidence from commons management in Ethiopia. American Economic Review, 105 (2), 747-783.

Lazear, E.P., Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. Journal of Political Economy, 89 (5), 841-864.

Mas, A., Moretti, E. (2009). Peers at work. American Economic Review, 99 (1), 112-145.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.C., 2003. Monetary and non-monetary punishment in the voluntary contributions mechanism. American Economic Review, 93 (1), 366-380.

Nikiforakis, N. (2008). Punishment and counter-punishment in public goods games: Can we really govern ourselves? Journal of Public Economics, 92, 91-112.

Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. Journal of Public Economics, 96(9-10), 797-807.

Nosenzo, D., Sefton, M. (2014). Promoting cooperation: the distribution of reward and punishment power, in Van Lange, P.A.M., Rockenbach, B., Yamagishi, T. (eds.), Social dilemmas: New perspectives on reward and punishment, Oxford, Oxford University Press.

O'Gorman, R., Henrich, J., Van Vugt, M. (2009). Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. Proceedings of the Royal Society B, 276, 323-329.

Onderstal, S., Schram, A.J., Soetevent, A.R. (2013). Bidding to give in the field. Journal of Public Economics, 105, 72-85.

Orzen, H. (2008). Fundraising through competition: evidence from the lab. Discussion Paper N. 11, Centre for Decision Research and Experimental Economics, University of Nottingham.

Ostrom, E., Walker, J., Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. American Political Science Review, 86, 404-417.

Samek, A.S., Sheremeta, R. (2014). Recognizing contributors: an experiment on public goods. Experimental Economics, 17 (4), 673-690.

Samek, A.S., Sheremeta, R. Selectiv recognition: how to recognize donors to increase charitable giving, Economic Inquiry, forthcoming.

Schram, A.J., Onderstal, S. (2009). Bidding to give: an experimental comparison of auctions for charity, International Economic Review, 50 (2), 431-457.

Xiao, E. (2013). Profit seeking punishment corrupts norm obedience. Games and Economic Behavior, 77, 321-344.

Xiao, E., Houser, D. (2011). Punish in public. Journal of Public Economics, 95, 1006-1017.

Xiao, E., Kunreuther, H. (2016), Punishment and cooperation in stochastic prisoner's dilemma game. Journal of Conflict Resolution, 60 (4), 670-693.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology, 51, 110-116.

Zizzo, D.J. (2010). Experimenter demand effects in economic experiments. Experimental Economics, 13 (1), 75-98.

Figures and Tables

Figure 1. Treatments

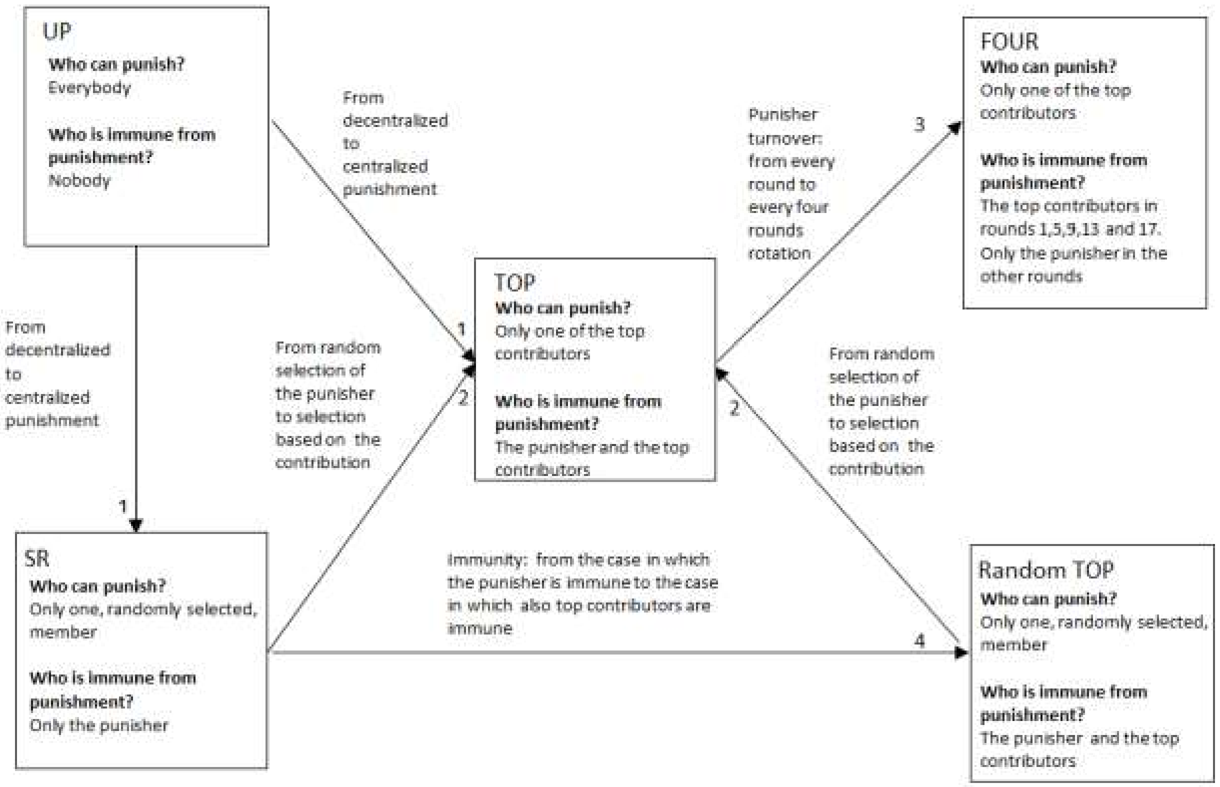

UP
Who can punish?
Everybody
Who is immune from punishment?
Nobody
From decentralized to centralized punishment
1
From decentralized to centralized punishment
1
SR
Who can punish?
Only one, randomly selected, member
Who is immune from punishment?
Only the punisher
From random selection of the punisher to selection based on the contribution
2
TOP
Who can punish?
Only one of the top contributors
Who is immune from punishment?
The punisher and the top contributors
Punisher turnover: from every round to every four rounds rotation
3
FOUR
Who can punish?
Only one of the top contributors
Who is immune from punishment?
The top contributors in rounds 1,5,9,13 and 17. Only the punisher in the other rounds
From random selection of the punisher to selection based on the contribution
2
Immunity: from the case in which the punisher is immune to the case in which also top contributors are immune
4
Random TOP
Who can punish?
Only one, randomly selected, member
Who is immune from punishment?
The punisher and the top contributors

Figure 2. Average contributions

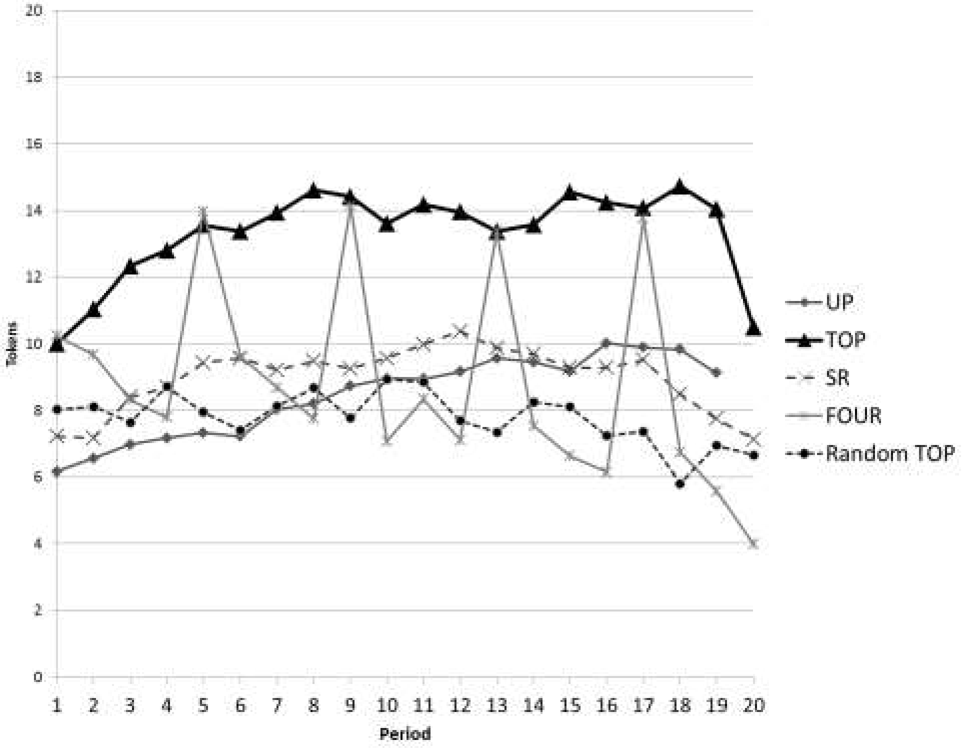


Figure 3. Average quantity of points assigned within groups

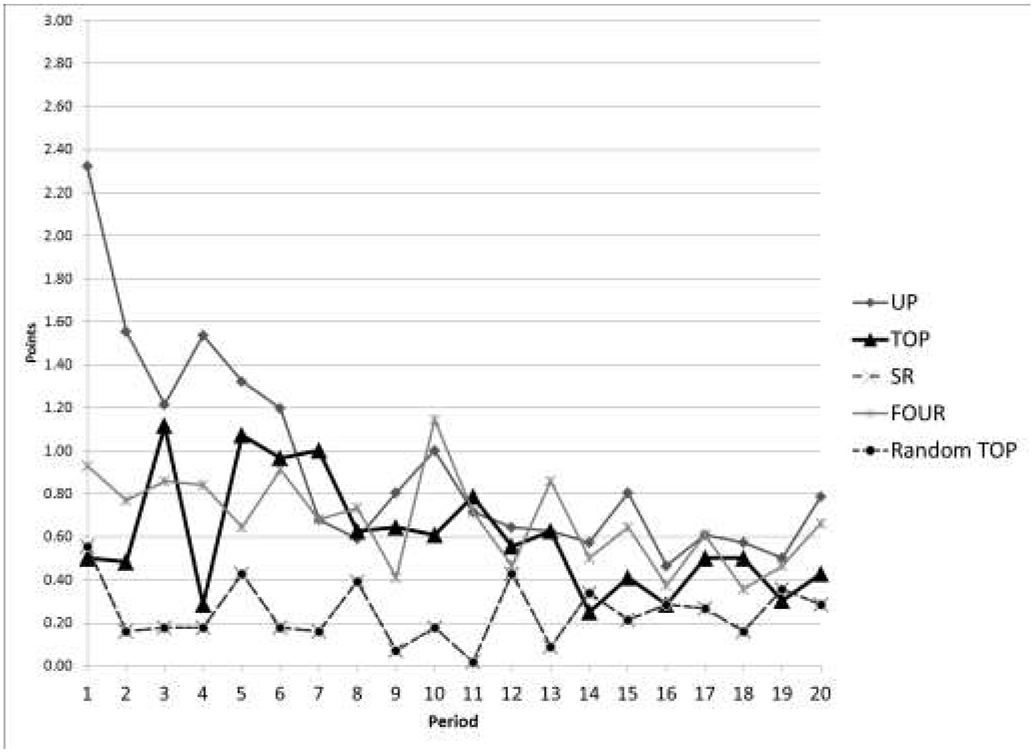


Figure 4. Average earnings

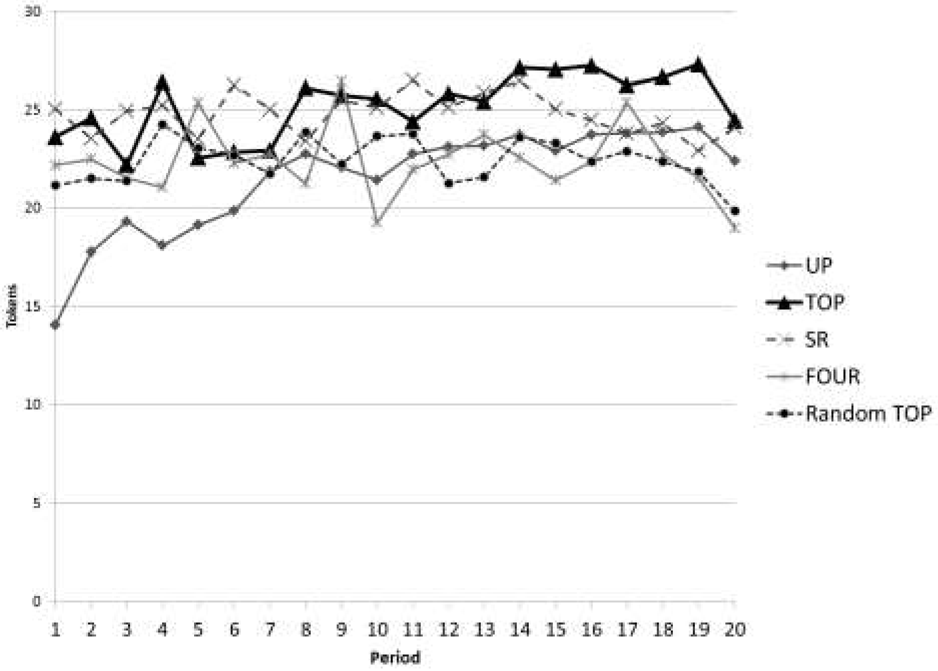


Table 1. Cost function

Points	0	1	2	3	4	5	6	7	8	9	10
Cost (tokens)	0	1	2	4	6	9	12	16	20	25	30

Table 2. Mean contributions across treatments (standard deviation in parentheses)

Group	UP	TOP	SR	FOUR	Random TOP
1	13.76	19.06	16.61	14.81	2.27
	(7.34)	(2.46)	(3.77)	(3.61)	(1.48)
2	18.40	11.86	11.55	9.53	18.42
	(3.31)	(3.90)	(2.70)	(6.06)	(2.73)
3	4.94	14.19	4.55	4.41	9.82
	(1.32)	(1.45)	(1.53)	(3.67)	(1.37)
4	11.30	11.29	5.48	12.51	4.58
	(4.03)	(1.58)	(3.35)	(3.87)	(2.50)
5	12.85	9.33	15.96	7.88	9.68
	(4.04)	(4.22)	(4.90)	(4.02)	(2.60)
6	4.58	17.95	2.45	4.11	14.53
	(2.95)	(2.71)	(1.86)	(4.50)	(3.63)
7	6.46	11.70	6.96	10.66	9.85
	(0.85)	(3.94)	(0.75)	(3.23)	(3.37)
8	2.18	16.25	4.84	5.43	8.4
	(0.72)	(2.05)	(0.68)	(3.36)	(3.18)
9	4.39	14.89	3.04	5.82	4.76
	(2.37)	(4.45)	(3.16)	(4.84)	(2.06)
10	1.64	17.85	6.51	18.25	4.52
	(0.94)	(2.93)	(1.83)	(2.93)	(1.59)
11	2.84	5.39	4.90	4.67	2.83
	(2.11)	(2.22)	(2.70)	(2.49)	(1.43)
12	15.13	17.31	16.44	4.41	3.18
	(4.77)	(3.61)	(3.10)	(3.93)	(3.05)
13	7.05	6.96	11.64	5.63	8.17
	(1.63)	(2.45)	(5.32)	(5.77)	(2.64)
14	11.11	12.75	14.63	15.31	
	(2.95)	(2.43)	(3.35)	(3.26)	
15					
16					
Mean	8.33	13.34	8.96	8.82	7.77

Table 3 Average contributions of subjects selected as punishers

Treatment	Average contribution (Standard deviation in parentheses)
TOP	17.23 (2.66)
SR	9.05 (5.53)
FOUR (key periods)	18.27 (2.50)
FOUR (non-key periods)	6.02 (6.41)
FOUR (all periods)	9.13 (5.12)

Table 4. Determinants of contributions in treatment FOUR

Contribution at t	All periods	Key periods only	Non-key periods only
Entitled not selected	7.35***	13.59***	5.67***
	(1.20)	(0.99)	(1.30)
Selected	2.97***	12,29***	0.07
	(0.81)	(1.36)	(0.88)
Constant	-8.18	1.70	-10.05
	(6.52)	(4.15)	(7.82)
R.sq.overall	0.30	0.74	0.29
Wald Chi(2)	323.28	311.47	273.63
N. Of obs.	1060	265	795

Random effects GLS. Standard error adjusted for clusters in group in parentheses.

The dependent variable takes values from 0 to 20. The omitted category is that not entitled subjects.

The variable Entitled not selected *is a dummy variable that is equal to 1 when the subject is entitled to punish since contribution in the group but has not been drawn to become the punisher, and 0 when the subject's contribution is not th* Selected *is a dummy variable that is equal to 1 if the subject is one of the top contributors of her group and she has also b punisher.*

Key periods are 1, 5, 9, 13, and 17, in which selection of the actual punisher takes place.

Controls: gender, age, major, nationality, and previous experiments in which the subject took part.

**** significant at 1%; ** significant at 5%; * significant at 10%.*

Table 5. Treatment effects

Contribution at t	Random Effect GLS
TOP (β_{TOP})	4.40*** (1.82)
UP	-0.67 (1.94)
FOUR (β_{FOUR})	-0.30 (1.85)
Random TOP (β_{RT})	0.44 (2.17)
Constant	6.92* (4.18)
$\beta_{TOP} - \beta_{RT}$	3.96** (1.96)
$\beta_{TOP} - \beta_{FOUR}$	4.71*** (1.67)
$\beta_{TOP} - \beta_{UP}$	5.07*** (1.68)
R.sq.overall	0.07
Wald Chi(2)	22.66
N. Of obs.	5060

Random effects GLS. Standard error adjusted for clusters in group in parentheses.
The dependent variable takes values from 0 to 20. The omitted treatment is the SR
"TOP" is a dummy variable that takes value 1 in the TOP treatment, and 0 elsewhere. "SR" is a dummy variable that takes value 1 in the SR treatment, and 0 elsewhere. "FOUR" is a dummy variable that takes value 1 in the FOUR treatment, and 0 elsewhere. "Random TOP" is a dummy variable that takes value 1 in the Random TOP treatment, and 0 elsewhere.
Controls: gender, age, major, nationality, and number of previous experiments in which the subject took part.
**** significant at 1%; ** significant at 5%; * significant at 10%.*

Table 6. Determinants of antisocial punishment

Percentage of anti-social punishment points at t	**Random Effect GLS**
Punishment points received at t-1	0.09 (0.74)
Contribution	-0.038*** (0.06)
Contribution at t-1	-0.005 (0.06)
Distance from average contribution	-0.012 (0.09)
Distance from average contribution at t-1	0.005 (.09)
Constant	0.57 (0.36)
R.sq.overall	0.68
Wald Chi(2)	309.73
N. Of obs.	160

Random effects GLS. Standard error adjusted for clusters in group in parentheses
The dependent variable takes values from 0 to 1.
Controls: gender, age, major, nationality, and previous experiments in which the subject took part.

**** significant at 1%; ** significant at 5%; * significant at 10%.*

Appendix A: Instructions (for the TOP treatment)

Good morning, thank you for participating in this experiment. You are taking part into a study on economic decisions. During the experiment, you can, depending on your decisions and on other participants' decisions, earn a considerable amount of money in addition to the 5 euros you will receive anyway. The answers you give and the choices you make will be totally anonymous. The experimenters will not be able to associate your choices and your answers to your name.

During the experiment you cannot communicate with other participants (otherwise you would be excluded from the experiment) and you should be very careful in reading the instruction that will appear on your screen and will be read out by one of the experimenters. If you have any questions, please ask the experimenters.

Your earnings will be calculated in tokens; each token will be converted in euros at the following ratio: 1 token = 0,02 euros.

At the end of the experiment, you will be asked to fill a short questionnaire; afterwards, we will proceed with the payment, that will occur in cash.

THE PARTICIPANTS

In this experiment there are in total 20 participants, which are divided into 5 groups with 4 members each (*in the sessions with 16 participants, we have 4 groups with 4 members each*).

The group composition will be the same for the whole experiment. Therefore you will always interact with the same three people, but you do not know their identity, and they do not know your own identity.

The experiment is composed by 20 rounds. In each round, the other 3 persons in your group will be randomly identified by means of numbers ("labels"). Note that the labels will be changing across rounds, and you will not be able to associate the choices made by a specific participant to a specific label. For example: in the first round the labels will be 1, 2 and 3. In the second round the labels will be 7, 5 and 11, in the third they could be 45, 2, 23 or 22, 32 and 11. But there is no relation among the different labels. The participant that is labeled with 1 in the first round, in the second could be indicated with 32, 54, or 33.

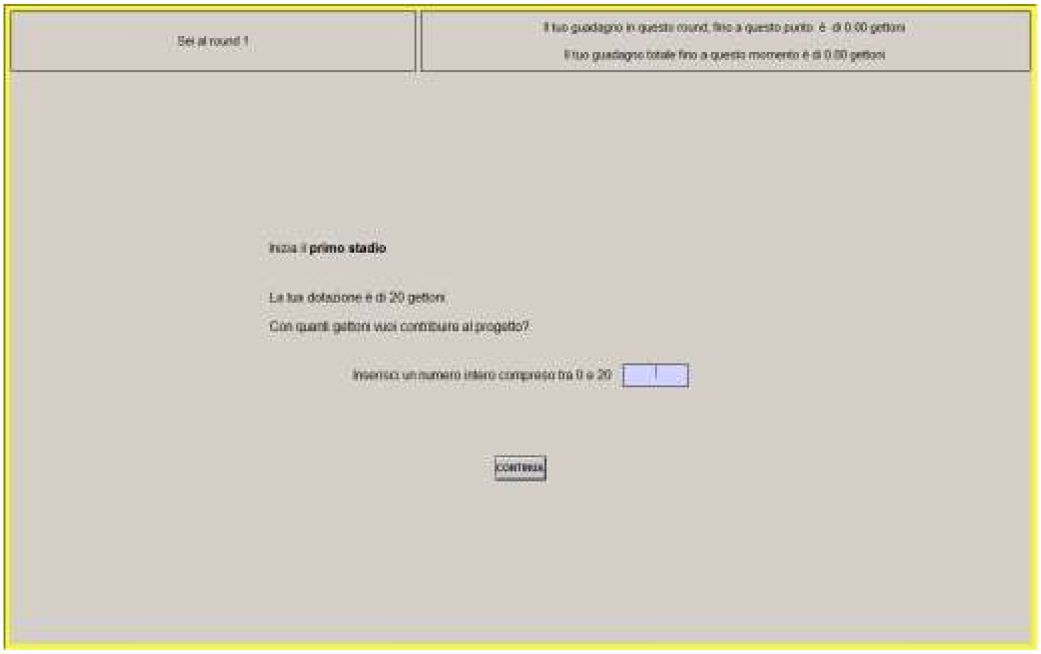
THE STAGES

Each round is composed by 2 stages.

During the first stage, you will decide how many tokens you will contribute to a "project". In the second stage, you will receive information on the number of tokens that the other 3 members of the group have decided to contribute to the project and therefore you will be able to reduce or not the earnings of any member of the group according to your and their levels of contribution at stage 1. This could be done by assigning points using your endowment of tokens.
The following paragraphs will describe the experiment in detail.

STAGE 1

At the beginning of each round each participant will receive 20 tokens. We call this amount "endowment". Your task is deciding how to use your endowment. You have to decide how many tokens you want to use to contribute to the project and how many tokens you will keep for yourself.
The number of the round appears in the left-top corner of the screen, whereas in the right-top corner you will see the amount of tokens you earned in the round and your total earnings up to that moment.
You have to decide how many tokens to contribute to the project by typing a number between 0 and 20 in the specific area (Figure 1). You can access to that area by clicking with the mouse. After typing the number of tokens you want to contribute to the project, you should press [CONTINUE]. Once you have taken your decision, you cannot modify it.



Sei al round 1

Il tuo guadagno in questo round, fino a questo punto è di 0.00 gettoni

Il tuo guadagno totale fino a questo momento è di 0.00 gettoni

Inizia il **primo stadio**

La tua dotazione è di 20 gettoni

Con quanti gettoni vuoi contribuire al progetto?

Inserisci un numero intero compreso tra 0 e 20

At the end of stage 1, you will be informed individually, on the screen of your computer, about your earnings, that consist of two elements:

a. The amount of the 20 initial tokens you kept for yourself (i.e. 20 tokens minus your contribution to the project);

b. Your payment deriving from the project, that is equal to the 40% of the sum of all the individual contributions to the project in your group (your contribution is included).

YOUR EARNINGS AT STAGE 1

Therefore, your earnings at the end of Stage 1 are calculated from the computer in the following way:

Your earnings after Stage 1 = (20 tokens – your contribution to the project) + 40% * (total contribution to the project)

Each group member's earnings are calculated in the same way; moreover each individual receives the same payment from the project. Assume, for instance, that in your group Member 1 will contribute 4 tokens, Member 2 will contribute 2 tokens, Member 3 will contribute 3 tokens and you will contribute 1 token. As a consequence, the total amount that the group will contribute is 10 tokens. Therefore, each member of the group will receive an amount equal to the 40% of 10 tokens = 4 tokens. The earnings of the 4 members of the group will be:

- Member 1: 20-4+4=20
- Member 2: 20-2+4=22
- Member 3: 20-3+4=21
- You: 20-1+4=23

As a further example, if Member 3 contributes 20 tokens, while you and Members 5 and 6 contribute 0 tokens, the whole amount of tokens that the group contributes is 20 tokens and each member of the group receives a payment from the project that is equal to the 40% of 20 tokens = 8 tokens. In this case, the earnings of the 4 members of the group will be:

- Member 3: 20-20+8=8
- Member 5: 20-0+8=28
- Member 6: 20-0+8=28
- You: 20-0+8=28

As a final example, if Member 13 contributes 0 tokens and you and Member 11 and 15 contribute 10 tokens each, the whole amount that the group contributes is 30 tokens and each member receives a payment from the project that equals 40% of 30 tokens = 12 tokens. In this case, the earnings of the 4 members of the group will be:

- Member 11: 20-10+12=22
- Member 13: 20-0+12=32
- Member 15: 20-10+12=22
- You: 20-10+12=22

STAGE 2

At the beginning of Stage 2, you can see how much the other members of the group have contributed to the project and which is the average level of contribution in the group.
Only the participant whose contribution has been **the highest in the group** will go on with the experiment and take part into Stage 2. The other participants have to wait the next round and will

be updated on the amount of their earnings as soon as the other members will have taken their decisions.

In this stage, if your contribution is the highest in the group, you can reduce or let unchanged the earnings of the members of the group (you can assign up to 10 points). Any other member of the group can, if she like, reduce your earnings if her contribution is the highest in the group. Each point reduces the earnings at stage 1 of the participants who receive them by 10%. You have to type the number of points you want to assign to each member of the group whose contribution has been lower than yours. If you choose to assign 0 points to a specific member of the group, you do not modify her earnings. If you choose to assign 1 point to a specific member of the group, you reduce her earnings of 10%. The amount of points you assign to each member of the group determines the amount you reduce their earnings at stage 1. If an individual receives 4 points, her earnings will be reduced by 40%, and if she receives 10 or more points her earnings will be reduced by 100%.

If you assign points, you face a cost that depends on the number of points you distribute in total. The table below shows the relation between the number of points you assign to a participant and the cost you incur:

# points assigned	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Cost you incur	**0**	**1**	**2**	**4**	**6**	**9**	**12**	**16**	**20**	**25**	**30**

For example, in your group you contributed 18 tokens. The other members (Members 1, 2 e 3) have contributed 10, 15 and 5 tokens respectively. Therefore, your contribution is the highest in the group and you can decide to assign points to all other members incurring the cost indicated in the table above, or keeping their earnings constant by assigning 0 points (and this will cost 0). If you assign, for instance, 0 point to Member 1 and 3 points to Member 3, the earnings of Member 1 at Stage 1 will be constant whereas the earnings of Member 3 will be reduced by 30%. As the number of points you assigned at Member 3 is 3 and you assigned points only to her, the cost you incur in total is 4 tokens.

When you have taken your decision, click on [CONTINUE].

YOUR EARNINGS AT STAGE 2

Your earnings at the end of Stage 2 will be calculated by the computer in the following way:

If your contribution has been the highest of the contributions in the group:
Your earnings at the end of Stage 2 = Your earnings at the end of Stage 1 - cost of the points assigned at Stage 2

If your contribution has not been the highest in the group:

Your earnings at the end of Stage 2 = Your earnings at the end of Stage 1 - points you received *10%* (Your earnings at the end of Stage 1)
Please note that at the end of the second stage your earnings could be negative. This occurs when the cost of the points you decided to assign is higher than your earnings at the first stage. In general, you can avoid this by paying a little attention.

APPENDIX B: Additional figures

Figure 1B: Contributions across treatments

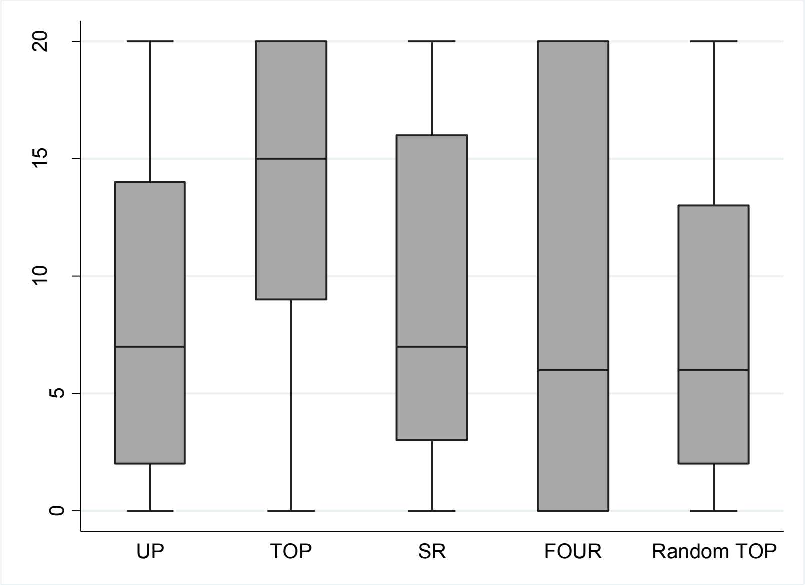


Figure 2B: Average number of punishment points distributed across treatments

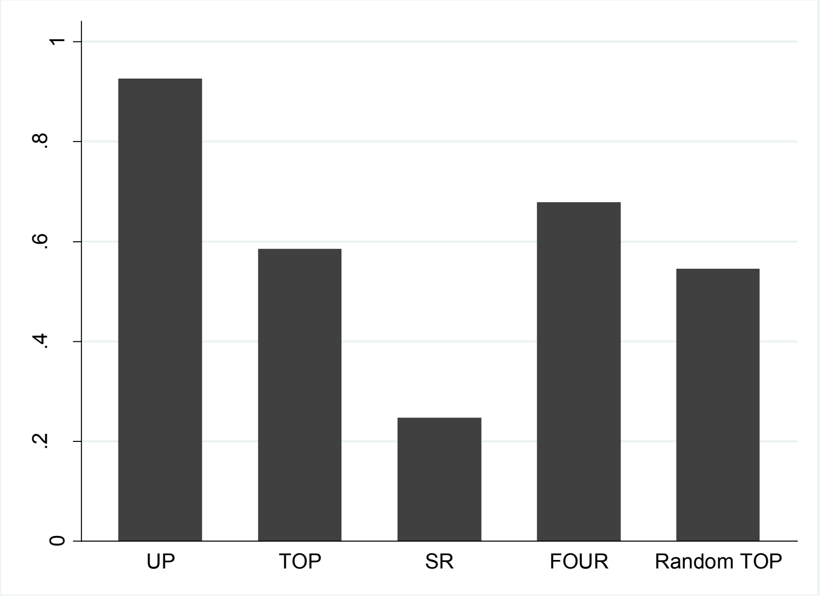


Figure 3B: Average earnings across treatments

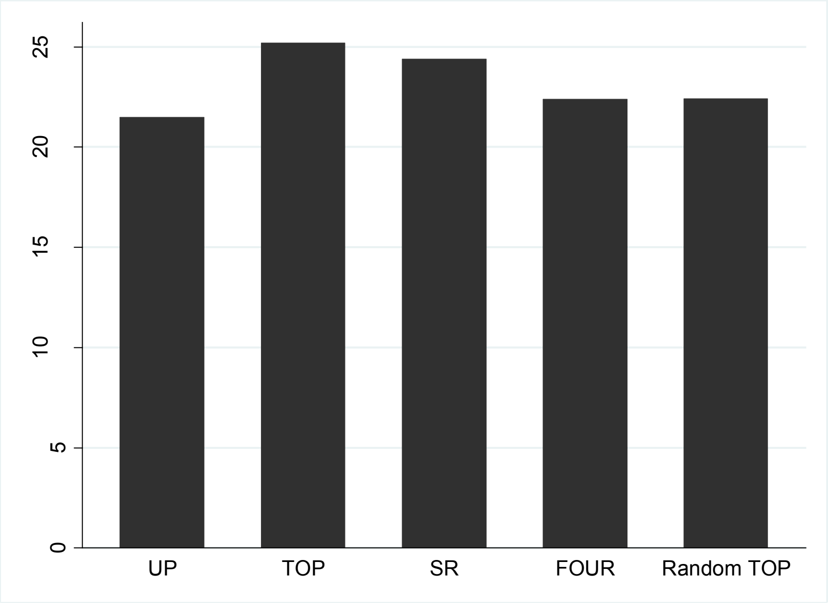


Table 1B. Punishment points and percentage of antisocial punishment points assigned across treatments

UP

	(1)	(2)	(3)
Group	Points given by *i* to *j*	Antisocial Contri<=Contrj	% of antisocial points
1	31	5	16.1%
2	42	0	0.0%
3	2	0	0.0%
4	45	12	26.7%
5	60	12	20.0%
6	72	29	40.3%
7	25	0	0.0%
8	45	1	2.2%
9	133	9	6.8%
10	181	51	28.2%
11	132	22	16.7%
12	15	0	0.0%
13	130	62	47.7%
14	123	8	6.5%
Total	1036	211 (20.3%)	
Mean	74.00	15.07	

SR

	(1)	(2)	(3)
Group	Points given by *i* to *j*	Antisocial Contri<=Contrj	% of antisocial points
1	1	1	100.0%
2	14	13	92.9%
3	31	8	25.8%
4	40	15	37.5%
5	2	0	0.0%
6	25	2	8.0%
7	16	1	6.3%
8	17	4	23.5%
9	5	5	100.0%
10	37	7	18.9%
11	26	0	0.0%
12	36	19	52.8%
13	14	2	14.3%
14	12	6	50.0%
Total	276	83(30.1%)	
Mean	19.71	5.93	

FOUR

	(1)	(2)	(3)
Group	Points given by *i* to *j*	Antisocial Contri<=Contrj	% of antisocial points
1	13	0	0.0%
2	55	26	47.3%
3	73	41	56.2%
4	48	17	35.4%
5	58	35	60.3%
6	123	43	35.0%
7	43	17	39.5%
8	42	22	52.4%
9	84	74	88.1%
10	0	0	0.0%
11	11	2	18.2%
12	39	16	41.0%
13	80	48	60.0%
14	90	1	1.1%
Total	759	342 (45.2%)	
Mean	54.21	24.43	

Random TOP (excluding the top contributor)

	(1)	(2)	(3)
Group	Points given by i to j	Antisocial Contri<=Contrj	% of antisocial points
1	17	0	0.0%
2	8	4	50.0%
3	52	15	28.8%
4	21	9	42.9%
5	29	7	24.1%
6	23	12	52.2%
7	24	13	54.2%
8	17	11	64.7%
9	32	21	65.6%
10	17	6	35.3%
11	27	25	92.6%
12	12	4	33.3%
13	40	28	70.0%
Total	319	155(48.5%)	
Mean	24.54	11.92	