Journal of Public Economics 144 (2016) 1–13

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Information-sensitive Leviathans ☆

Andreas Nicklisch [a], Kristoffel Grechenig [b], Christian Thöni [c,*]

[a] HTW Chur and German Research Foundation, Research Unit "Needs Based Justice and Distributive Procedures", Switzerland
[b] Max Planck Institute for Research on Collective Goods, Bonn, Germany
[c] University of Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 12 July 2015
Received in revised form 10 September 2016
Accepted 15 September 2016
Available online 19 September 2016

JEL classification:
C92
D02
H41

Keywords:
Centralized sanctions
Cooperation
Experiment
Endogenous institutions

ABSTRACT

We study information conditions under which individuals are willing to delegate their sanctioning power to a central authority. We design a public goods game in which players can move between institutional environments, and we vary the observability of others' contributions. We find that the relative popularity of centralized sanctioning crucially depends on the interaction between the observability of the cooperation of others and the absence of punishment targeted at cooperative individuals. While central institutions do not outperform decentralized sanctions under perfect information, large parts of the population are attracted by central institutions that rarely punish cooperative individuals in environments with limited observability.

1. Introduction

Human life in Thomas Hobbes' natural state is lonely, short, and brutal, "a time of war where every man is enemy to every man" (Hobbes, 1651). To redress this grim fate of violence and distrust, people appoint a central authority—a *Leviathan*—to enforce cooperative behavior. People voluntarily delegate their sanctioning power to the Leviathan, in the hope for a more efficient outcome.

In contrast to Hobbes' bleak view, contemporary research suggests that people successfully use *decentralized sanctions* (peer-to-peer punishment) to enforce cooperation (Ostrom et al., 1992; Fehr and Gäter, 2000) and reach efficient outcomes in the long run (Gächter et al., 2008). If human societies are able to organize themselves in a decentralized fashion, one would expect to find many self-governed societies. However, the opposite is the case: We live in a world where centralized sanctions play are very important role, on the national and even on the supranational level.[1] Why did modern societies develop centralized institutions to enforce norms? Under which conditions are people willing to renounce their sanctioning power in favor of a central authority?

We use an experimental approach to these questions, and we analyze a voting by feet mechanism in favor of or against central authorities. We introduce an environment where players ('citizens') participate in a social dilemma. Prior to this, they can vote by feet for one of three institutions: centralized punishment (*CenPun*), decentralized punishment (*DecPun*), and a sanction-free institution (*NoPun*). In *CenPun*, an additional (randomly drawn) subject (the 'authority') can punish the citizens in his institution, while citizens are not allowed to punish each other. The authority's payoff is increasing in the citizens' contributions, and the authority does not have to bear the costs of punishment. In *DecPun* citizens can punish other citizens in the same institution, at their own expenses.

☆ For helpful comments and discussion, we thank the Editor, two anonymous referees, Berno Büchel, Ernst Fehr, Urs Fischbacher, Guillaume Frechette, Simon Gächter, Sonja Köke, Manfred Milinski, Nikos Nikiforakis, Louis Putterman, Arno Riedl, Bettina Rockenbach, Roberto Weber, and the seminar participants at the University of Pennsylvania, the Max Planck Institute (Collective Goods, Bonn), the University of Nottingham, the University of Lausanne, and the University of Cologne. We would also like to thank the Max Planck Society and the German Research Foundation (Ni 1610/1-1) for financial support. An earlier version of this paper circulated under the title "Information-sensitive Leviathans: The Emergence of Centralized Punishment" (Nicklisch et al., 2015).

* Corresponding author.
E-mail address: christian.thoeni@unil.ch (C. Thöni).

[1] Examples are institutions like the European Union, the International Military Tribunal in Nuremberg in 1945/46, and the United Nations Security Council.

http://dx.doi.org/10.1016/j.jpubeco.2016.09.008
0047-2727/

Our analysis builds on three major challenges for self governance which have been identified in the literature: antisocial punishment, revenge, and incomplete information. Antisocial punishment (or perverse punishment) refers to the observation that some subjects target their punishment at cooperative subjects. There is ample evidence that the strength and frequency of antisocial punishment negatively relates to contributions.[2] Related is the problem of retaliation for received punishment. Some studies find that retaliation weakens decentralized punishment institutions because cooperative individuals are less willing to punish free riders (Denant-Boemont et al., 2007; Nikiforakis, 2008; Nikiforakis et al., 2012), while others do not find such a general effect (Kamei and Putterman, 2015). Finally, decentralized punishment can become inefficient in increasing contributions when subjects receive only imperfect information about the contributions of others. Contrary to intuition (but in accordance to the theoretical argument we develop below) subjects tend to punish more when information becomes more noisy (Grechenig et al., 2010; Ambrus and Greiner, 2012).[3]

All three problems are closely related. For instance, less information leads potentially to more punishment of cooperative subjects, which might in turn trigger retaliatory punishment. While in principle one could exogenously vary multiple dimensions of this complex interaction we restrict our design to a manipulation of the informational quality. In terms of the underlying phenomenon (establishing cooperation in groups) we think that it makes sense to see informational conditions as an exogenous characteristic of the environment, while the individual propensity to engage in antisocial punishment or revenge seems endogenous in its nature. Thus, our approach is to vary the quality of information exogenously and study its impact on the relative popularity of the three institutions. More specifically, we introduce three environments differing with respect to the accuracy of information citizens and the authority receive about the contributions of others. In treatment condition ONE, they receive accurate signals about the contributions; in POINT-NINE, they receive signals which are correct in 90% of the cases, while in POINT-FIVE, the signals are correct in 50% of the cases. We measure the popularity of an institution by the fraction of citizens it attracts.

We find that the treatment variation significantly influences institutional choices. In particular, imperfect information lowers the popularity of *DecPun*. We show that the punishment of cooperative citizens significantly influences institutional choices. Finally, with regard to our main research question we find that *CenPun* becomes the most popular institution only when there is imperfect information and at the same time the central authority (the Leviathan) applies a punishment strategy which minimizes the punishment of cooperative citizens. At the same time, revenge motives seem to be less important in our design and cannot explain differences across treatments.

Our study complements and expands recent discussions on the formation of centralized institutions. Dal Bó et al. (2010) compare the effect of a democratically chosen and an exogenously imposed policy intervention aimed at eliminating the attractiveness of free-riding. They find that democratically installed interventions increase cooperation significantly compared to exogenously imposed interventions. Moreover, endogenously introduced regimes with centralized sanctions perform well, even when sanctions are non-deterrent (Tyran and Feld, 2006), or, in some cases, outperform decentralized sanctions (O'Gorman et al., 2009).

Another important aspect of centralized institutions is the way how sanctions are implemented. In contrast to our approach, the majority of articles focus on sanctions that are executed automatically. If both, decentralized and automatically executed centralized punishment are available, the latter seems to crowd out the use of the former (Kube and Traxler, 2011). Markussen et al. (2014) investigate the choice of centralized sanctions through voting, when centralization is costly (and executed automatically). They find that people are particularly responsive to the fixed costs of having a centralized sanctioning scheme in place, more so than they respond to whether or not the sanctioning scheme is fully deterrent. Putterman et al. (2011) allow participants to vote on the rules of an automatically executed sanctioning scheme. The results show that many groups quickly implement sanctions that induce efficient outcomes.

Kosfeld et al. (2009) analyze the choice for automatically executed punishment mechanism which may govern only a subset of players. They show that participants are unwilling to implement equilibrium punishment which allows some players to free-ride. Andreoni and Gee (2012) investigate the formation of centralized sanctions through voting for a sanctioning scheme that punishes only the lowest contributor and find that full contributions are quickly achieved at very low punishment costs. Importantly, these articles focus on sanctions that are executed automatically; that is, once an implemented rule is violated, players are punished with a certain probability while contribution decisions are perfectly observable.[4]

Closer to our approach is Fehr and Williams (2013). They offer citizens the choice between uncoordinated decentralized, coordinated decentralized, or centralized punishment, which is executed by a democratically elected leader. They show that centralization of sanctions leads to high cooperation along with the selection of pro-social leaders who refrain from punishing high contributors. Similarly, Gross et al. (2016) explore the emergence of central punishment authorities under perfect information. They demonstrate that if individuals can transfer their punishment power to others, cooperators empower subjects who have previously indicated their willingness to sanction free-riders. As a consequence, groups with centralized punishment and high cooperation emerge.

Summarizing the previous literature, both centralized as well as decentralized sanctioning sustain cooperation if chosen endogenously. If available, evidence suggests that citizens choose very selectively centralized institutions. That is, effective centralized sanctioning is demanded, but citizens are unwilling to accept centralized punishment that violates their fairness considerations (e.g., allowing some players to free-ride, or punishment targeted at contributors).

In our setting, it is up to the authorities to deliver effective sanctioning. Like in Fehr and Williams (2013), we introduce the authority as a player, who may use punishment in a similar, potentially erroneous or malevolent fashion as his citizens.[5] We do so as we believe that the feature is of particular importance to explain the choice of authorities in earlier societies. That is, we compare centralized and decentralized sanctioning when authorities are not equipped with better mechanisms to guide behavior than citizens (e.g., our authorities are not better informed than citizens, nor do they have more efficient punishment technologies than citizens). Rather, our authorities are autocratic leaders, holding absolute punishment power. Furthermore, like in a feudal society the authority is not appointed by a competitive procedure, but he is merely born into his position.[6]

[2] See e.g. Gächter et al. (2005), Bochet et al. (2006), and Herrmann et al. (2008).

[3] On the other hand, Leibbrandt et al. (2015) provide evidence that antisocial punishment increases when *more* information is provided, i.e., when subjects can identify individual punishers in the group.

[4] See also Sutter et al. (2010), who study endogenous choices between positive and negative sanctioning systems.

[5] This is similar to Carpenter and Matthews (2012), who analyze the effect of third-party punishment for contributions in public good games.

[6] For the effect of democratically appointed leaders see also Hamman et al. (2011) and Corazzini et al. (2014).

Following previous works showing that decentralized sanctions prevail over a sanction-free environment (Gürek et al., 2006) and over a pure reputation-building environment (Rockenbach and Milinski, 2006), we let our players choose their institution by leaving societies (exit), but not by vote (voice).[7] Consequently, each citizen is free to migrate to his most preferred institution. In addition, due to the third alternative *NoPun*, our setting requires citizens to choose actively in favor of one punishment institution, which allows us to interpret citizens' institutional choice predominantly as a choice in favor of centralized or decentralized punishment rather than a decision against the alternative sanctioning institution which is not chosen.

Our article is structured as follows. Section 2 describes our basic game and derives an expression for deterrent punishment; in Section 3 we introduce the experimental setting and discuss behavioral predictions. Section 4 presents the results, and Section 5 concludes the paper with a discussion.

2. Model

2.1. The game

We set up a game which embeds competition between centralized punishment, decentralized punishment, and a punishment-free institution in a public goods game. We combine a *voting by feet* mechanism between different sanctioning regimes (Gürek et al., 2006) with *imperfect information* about individual contributions (Grechenig et al., 2010). There are ten citizens and one authority. The game consists of three stages. In stage one, each citizen i independently chooses an institution. There are three institutions, each associated with a specific punishment rule: centralized punishment (*CenPun*), decentralized punishment (*DecPun*), and no punishment (*NoPun*). We denote by $\boldsymbol{C}$, $\boldsymbol{D}$, and $\boldsymbol{N}$ the set of citizens in the three institutions. Citizens in a given institution play a public goods game as long as at least two citizens are present.

In stage two, each citizen receives an endowment of 20 experimental currency units (ECU). Citizens simultaneously choose a contribution $g_i \in \{0, 2, 4, \ldots, 20\}$ to the public good. Each ECU contributed to the public good is multiplied by 1.6 and the resulting amount is divided equally among the citizens in the respective institution. This payoff function keeps the marginal social return from the public good constant for different group sizes, so that there are no productivity advantages for larger groups. Consequently the marginal per capita return decreases in group size.[8] At the end of stage two a citizen i in the institution *CenPun* earns a profit of

$$\hat{\pi}_i = 20 - g_i + \frac{1.6}{c}\sum_{k\in\boldsymbol{C}} g_k, \tag{1}$$

where $c \equiv |\boldsymbol{C}|$ denotes the number of citizens in *CenPun*. For citizens in the other two institutions the same payoff function holds with respect to the sets $\boldsymbol{D}$ and $\boldsymbol{N}$.

In stage three, players receive signals about the contribution of the other citizens in their institution. For each citizen i a signal is produced, such that

$$s_i = \begin{cases} g_i & \text{with} \quad prob = \lambda, \\ \tilde{g}_i & \text{with} \quad prob = 1-\lambda, \end{cases} \tag{2}$$

where $\tilde{g}_i$ is randomly drawn from the set $\{0, 2, 4, \ldots, 20\}\backslash\{g_i\}$ with uniform probabilities. Thus, for each citizen, there is an independent random draw determining whether the signal corresponds to the true contribution or not. If not, another independent draw selects a different contribution. The signal s_i is communicated to all other citizens in i's institution, and, in case of *CenPun*, also to the authority. Citizen i does not know whether the other participants receive a true or false signal about his contribution.

In addition, all citizens receive an extra endowment of three units. Depending on their institution, players assign punishment points (that is, citizens in *DecPun* and the authority in *CenPun*), and the final payoffs are realized. The three institutions differ only in stage three. For a citizen in *NoPun* the payoff equals the profit after stage two plus the extra endowment:

$$\pi_i = \hat{\pi}_i + 3 \qquad \forall\, i \in \boldsymbol{N}. \tag{3}$$

In *DecPun* all citizens decide simultaneously over punishment $p_{i\to k}$ with $k \in \boldsymbol{D}\backslash\{i\}$. Each punishment point assigned to another citizen leads to a deduction of three units from the punished citizen's payoff and reduces the punisher's payoff by one unit. Each citizen can spend up to her extra endowment for punishment, that is, $\sum_k p_{i\to k} \le 3$.[9] Units not spent on punishment are credited to the citizens' payoff. For a citizen i in *DecPun*, the payoff equals

$$\pi_i = \hat{\pi}_i + \left(3 - \sum_{k\in\boldsymbol{D}\backslash\{i\}} p_{i\to k}\right) - 3\sum_{k\in\boldsymbol{D}\backslash\{i\}} p_{k\to i} \qquad \forall\, i \in \boldsymbol{D}. \tag{4}$$

In *CenPun* all punishment decisions are delegated to the authority. The authority decides over punishment $p_{\to k}$ with $k \in \boldsymbol{C}$. Like in *DecPun* each punishment point assigned to a citizen leads to a deduction of three units from the punished citizen's payoff and costs one unit. In *CenPun* these costs have to be borne equally by all other citizens in the institution. In sum, the authority can spend up to the extra endowment of all its citizens for punishment, i.e., $\sum_k p_{\to k} \le 3c$. In addition, maximum punishment targeted at a single citizen is restricted to $3(c-1)$. Units not spent on punishment are credited to the particular citizen's account. Hence, *DecPun* and *CenPun* are identical with regard to the feasible set as well as the financial consequences of punishment. The only difference is that punishment *decisions* are taken by the authority. For citizen i in *CenPun*, the payoff equals

$$\pi_i = \hat{\pi}_i + \left(3 - \frac{\sum_{k\in\boldsymbol{C}\backslash\{i\}} p_{\to k}}{c-1}\right) - 3p_{\to i} \qquad \forall\, i \in \boldsymbol{C}. \tag{5}$$

[7] Historically, the importance of exit mechanisms for the organization of tribes, or even the fall of entire nations (e.g., East Germany), is well documented (Hirschman, 1970, 1978). Contemporary exit mechanisms capture competition between jurisdictions for corporations or tax payers.

[8] We designed the game to be neutral with regard to the optimal group size. Our payoff function ensures that any given average contribution results in the same average profit for all group sizes. It is of course still possible that the change in the marginal per capita return introduces group size effects, as suggested in the literature for public goods games without punishment (Nosenzo et al., 2015).

[9] This design of the punishment stage has the property that larger groups have more resources for punishment. It is, however, unclear whether this means that punishment is more severe in larger groups, because larger groups might be faced with more deviators, or subjects might be more likely to act as a bystander on the punishment stage. For a discussion of the adaptation of the punishment mechanism to various group sizes see also Roux and Thöni (2015).

The authority's payoff equals the average profit after stage two of all citizens in institution *CenPun*

$$\pi_A = \frac{\sum_{i \in \boldsymbol{C}} \hat{\pi}_i}{c} \quad \text{if } c \geq 2. \tag{6}$$

If there is only one citizen in an institution, there is no public good and no punishment. In this case, the citizen receives a payoff of 20. If there are less than two citizens in *CenPun* the authority receives a payoff of 20.[10] All parameters, the signal technology (λ), and payoff functions are public information.

We vary the information environment λ across treatment conditions. In treatment ONE citizens and the authority receive perfect information regarding the contributions of members of their institution ($\lambda = 1$). In treatment POINT-NINE we set $\lambda = .9$, such that citizens and the authority receive a signal about the others' contributions that displays the accurate information in nine out of ten cases (and a different contribution in the remaining case). Finally, in POINT-FIVE players receive accurate information in five out of ten cases ($\lambda = .5$).

2.2. Deterrent punishment

In the main text we restrict our theoretical analysis of the game to the punishment stage. In particular, we derive an expression for *deterrent* punishment, that is, the strength of punishment required to render unilateral deviation from a situation with mutual cooperation unprofitable. At the end of this section, we sketch the equilibria of the entire game, but characterize it fully in the online appendix.

Assume a central punishment institution seeks to enforce a contribution of $\bar{g} > 0$ in a group of c citizens. Let us assume that a citizen i is risk neutral and has selfish preferences, and all other citizens contribute $g_{-i} = \bar{g}$. As we show in the Appendix, minimal punishment required to make i indifferent between contributing $g_i = \bar{g}$ and $g_i = 0$ is

$$p^*_{\rightarrow i}(s_i, \lambda, c, \bar{g}) = \frac{(10c - 16)}{3c(11\lambda - 1)} \max\{\bar{g} - s_i, 0\} \quad \text{for } c \geq 2, \lambda > \frac{1}{11}. \tag{7}$$

Signals equal to $\bar{g}$ or above trigger no punishment. For signals lower than $\bar{g}$ punishment is linearly decreasing in the signal. The absolute slope of the punishment function in $s_i\left(\left|\frac{\partial p}{\partial s_i}\right|\right)$ is increasing in λ and approaching infinity as $\lambda \to \frac{1}{11}$, which refers to the case of an uninformative signal. Thus, the lower the accuracy of the signals the more punishment is required at a given signal to achieve deterrence. Enforcing higher contributions ($\bar{g}$), as well as increasing the group size (c) requires more punishment for a given signal.

The punishment necessary to deter free-riding is independent of the punishment institution. In case of *CenPun* the authority uses Eq. (7) to punish all the $i \in \boldsymbol{C}$ citizens in the group. In case of *DecPun* the expression would be the same with the exception that we have to replace c by d, the number of players in *DecPun*. However, it lies in the very nature of this institution that players face a coordination problem in the punishment stage. Eq. (7) just specifies the total punishment that should be assigned to player i, but does not specify the allocation to punishers. A natural benchmark would be that each citizen bears the same share of the punishment costs for citizen i, i.e., all players j punish player i by $p_{j \rightarrow i} = \frac{p^*_{\rightarrow i}}{d-1}, \forall j \in \boldsymbol{D} \setminus \{i\}$ punishment points. If all players follow this punishment strategy, then *DecPun* and *CenPun* would be equivalent in terms of payoffs for the citizens.

In the Introduction 1, we stressed the role of antisocial punishment as a determinant for contributions. We understand this term in the sense of describing an act with the *intent* to punish cooperative subjects. Under imperfect information there is a potential misalignment between the intended action and the realized action, either due to false signals, or due to signals suspected to be false. This makes it difficult to qualify a punishment act as antisocial punishment. To account for this we introduce two terms, both relating to antisocial punishment: For punishment which—independent of the signal—hits cooperative citizens (with $g_i > \bar{g}$), we use the term *contributor punishment*; for punishment acts targeted at citizens for which the punisher receives a high signal ($s_i \geq \bar{g}$) we will use the term *misguided punishment* (independent of g_i). In both cases we will refer to the opposite punishment (either $g_i < \bar{g}$, or $s_i < \bar{g}$) as free-rider punishment. Punishment according to Eq. (7) therefore rules out misguided punishment, while contributor punishment becomes stronger the lower the quality of the signals.[11]

Given the payoff functions it is in the interest of all players in *DecPun* and *CenPun* to enforce maximum contributions ($\bar{g} = 20$). However, depending on λ this might not be feasible. In particular, high levels of noise in the signals require amounts of punishment which are outside of the feasible set of the punisher(s). In Appendix A.1 we show that enforcing maximum contributions is feasible in ONE and POINT-NINE, but typically not in POINT-FIVE. Since the punishment endowment and technology are identical in *DecPun* and *CenPun* this holds equally for both institutions. Under standard assumptions (selfish preferences and subgame perfection) punishment in *DecPun* is a non-credible threat and should not occur. Consequently, the central authority should be able to attract all citizens in ONE and POINT-NINE. Countless experiments suggest, however, that this is not an accurate description of punishment behavior in decentralized sanctioning institutions. In the next section we specify the experimental setup and use theoretical arguments and stylized facts on punishment to formulate behavioral predictions.

3. Implementation

3.1. Experimental setup

The experiment is played in matching groups of eleven subjects. Prior to the start of the game we randomly allocate one subject in each matching group to the role of the authority and ten subjects to the role of the citizen.[12] Roles remain the same throughout the experiment.

Because the game is fairly complicated, and because we think that interesting things might unfold with time we implement a repeated game of 32 periods. Participants know that they play the game for the finite number of periods.[13] Since we want to provide the three institutions with some time to establish cooperation before they are put into competition with other institutions, the citizens in our experiment choose their institution every fourth period only. Thus we implement a game with eight phases consisting of four periods each. At the beginning of each phase all subjects allocated to the role of citizens choose one of the three institutions and remain there during the phase.

Each period consists of three steps, a contribution step, a punishment step, and an information step. Appendix A.3 shows the information provided on the screens during the experiment. In the punishment step, all citizens and the authority receive the signals from the citizens in their institution. If applicable, citizens or the

[10] These payoffs ensure that the authority has an incentive attract at least two citizens, and citizens have an incentive to form groups.

[11] When analyzing the data we cannot observe $\bar{g}$ and we will use the mean contribution (or the mean signal) instead.

[12] Grosse et al. (2011) use the same technique to introduce a observer in their public good game.

[13] English translations of the instructions are reported in Appendix A.2. Before the experiment starts, subjects have to solve a set of control questions on the computer screen.

authority choose their punishment points. The identification number of citizens are randomly reassigned between periods. In the information step, citizens learn their period payoff including the total amount of punishment received. Citizens do not receive information about their own signal, that is, they do not know whether other subjects were correctly informed about their contribution or not. Citizens learn only the total amount of punishment received, and not the number nor the identifier of the citizens who punished them.

At the beginning of each phase all citizens are informed about the outcome in all institutions (see screenshot in Fig. A2). In particular, when choosing an institution citizens know (i) the number of citizens (ii) the average contribution, and (iii) the average profit in all three institutions and for all previous periods. At this point all information is undistorted. In the light of this information citizens choose their institution for the next phase. There is no cost attached to switching an institution.[14]

3.2. Behavioral predictions

In this section, we develop predictions about the effect of the treatment variation on the main outcome variable, the institutional choice. Our null hypothesis is that the amount of noise in the signals does not affect the number of citizens attracted by the three institutions. The alternative hypothesis is that the popularity of the three institutions systematically varies with the amount of noise.

While our theoretical analysis of the institutional choice briefly mentioned at the end of Section 2.2 does not offer a compelling prediction, we still want to be more specific as to what we expect from the three treatments. In the following we use theoretical arguments as well as stylized facts from previous experimental research to develop three conjectures about the direction of the effect. We start with ONE, the treatment with perfect information.

A large body of evidence on public goods games with punishment shows that the majority of individuals is willing to use costly punishment to sanction free-riders in games with decentralized punishment institutions (Chaudhuri, 2011). While it is difficult to explain costly punishment under standard assumptions, theories of social preferences explain punishment either by assuming inequality averse preferences or reciprocal preferences.[15] In the former cooperative citizens are willing to punish free-riders to eradicate their payoff differences; in the latter free-riding is perceived as an unkind act, which motivates retaliatory punishment.

While we do not offer a theoretical treatment of our game with social preferences, it seems intuitive that both flavors of social preferences can rationalize that citizens contribute if (and only if) others contribute as well, and that unequal contributions trigger punishment to deter free-riding.

In line with this perspective, earlier experimental studies find that under perfect information a majority of the subjects end up in the punishment institution when given the choice between decentralized punishment and no punishment (Gürek et al., 2006). Subjects' willingness to punish free-riders creates a credible threat and coordinates behavior on high contributions, and only little actual punishment is required to enforce this outcome. Based on this stylized fact, we expect that citizens manage to reach and maintain high contributions in *DecPun* in ONE.[16] Furthermore, previous experimental evidence suggests that subjects have a preference for retaining authority (Fehr et al., 2013), and they might gain satisfaction from punishing defectors themselves (De Quervain et al., 2004). In addition, citizens may fear that the central authorities punish excessively due to the fact that they do not bear the marginal cost of punishment. For treatment ONE we thus expect *DecPun* to be the prevailing institution:

Conjecture 1. *Under perfect information (treatment* ONE*) the majority of citizens choose DecPun.*

What changes under imperfect information? We begin with treatment POINT-NINE, where signals are accurate in 90% of the cases. The results of Grechenig et al. (2010) suggest that a small amount of noise does not hamper the enforcement of high contributions in decentralized punishment. While they find more punishment in POINT-NINE than under perfect information, profits in later periods are still higher than in the treatments without punishment. Consequently, as in treatment ONE, institutions with punishment should have a competitive advantage.

For three reasons we think that in this treatment citizens might find it attractive to delegate the sanctioning decisions to the central authority. First, unlike in ONE, there is the risk that punishment acts do not hit the right citizen, and thereby do not serve as retaliation (in case the punishment was motivated by reciprocal preferences), or increase instead of decrease the inequality in the group (for inequality averse agents). Since the authority does not have more information than the citizens, erroneous punishment acts are just as likely in *CenPun*. While an inequality averse player does not care about who pulled the trigger, reciprocally motivated agents do. Thus, the latter type of agent might prefer to shift the responsibility for punishment to the authority.

A second reason is that punishment which mistakenly hits high contributors might motivate revengeful reactions in the form of misguided punishment (see Herrmann et al., 2008; Leibbrandt et al., 2015). While retaliatory punishment may also play a role in ONE, we think that the levels of misguided punishment in our subject pool (Univ. of Bonn) are too low to trigger such vicious cycles. The treatment POINT-NINE, on the other hand, might introduce the right amount of ambiguity in order to mess things up in *DecPun*. *CenPun* could then be an attractive alternative, because this institution delegates all responsibility for punishment to the authority and rules out retaliatory punishments among the citizens.

Third, the results of Ertan et al. (2009) show that subjects prefer institutional environments which do not allow for punishment of high contributors. While these results stem from experiments with perfect information, we interpret this as evidence that the more punishment of cooperative citizens is perceived a problem, the more citizens are willing to tolerate restrictions in their punishment authority. For the treatment with low noise levels we therefore expect:

Conjecture 2. *Under low noise levels (treatment* POINT-NINE*) the majority of citizens choose CenPun.*

In treatment POINT-FIVE signals are accurate only half of the time. In experiments with exogenous institutional environments,

[14] It is certainly an extreme assumption that moving from one institution to another is costless. However, we decided against introducing an arbitrary switching cost because we want to measure preferences for institutions unaffected by other considerations such as the sunk cost fallacy.

[15] For inequality aversion there are fairly specific predictions for free-rider punishment (Fehr and Schmidt, 1999), as well as antisocial punishment (Thöni, 2014). We are not aware of an application of contemporary models of reciprocity (such as Dufwenberg and Kirchsteiger, 2004 or Falk and Fischbacher, 2006) to the punishment decision in public goods games. Ambrus and Pathak (2011) analyze public good games with reciprocal preferences.

[16] Herrmann et al. (2008) show that there is large cross-societal variation in the public goods game with punishment. We conducted the experiments in Bonn, Germany, where previous evidence points towards high contributions and little antisocial punishment in *DecPun*. Consequently, the conjectures we develop here are sensitive to the societal background in which the experiment is conducted.

Grechenig et al. (2010) find that contributions in the treatment with decentralized punishment are similar as in the treatment without punishment, despite the fact that the subjects use punishment no less than in the treatments with accurate signals. Because punishment is costly, profits are lower in the treatment with punishment. In a similar setting, Ambrus and Greiner (2012) also find lower profits when punishment is available. This should give *NoPun* a competitive advantage over *DecPun* in POINT-FIVE. Furthermore, our theoretical analysis in Appendix A.1 shows that enforcing full contributions under high noise is often impossible, because the deterrent punishment levels are outside the feasible set. Unlike under perfect information, where the threat of punishment suffices, enforcing contributions in POINT-FIVE requires high levels of punishment even if the group fully cooperates. For this reason we expect that in such a noisy environment *CenPun* cannot offer substantial advantages over *DecPun*, and thus neither of the punishment institutions will prevail:

Conjecture 3. *Under high noise (treatment* POINT-FIVE*) the majority of citizens choose NoPun.*

4. Results

We ran 15 experimental sessions with 30 independent populations (330 participants, 110 per treatment). Each subject participated in only one population. The experiments were conducted at the laboratory for economic experiments (EconLab) at the University of Bonn with mostly undergraduate students from various fields. Six percent of participants were non-students, 56% of participants were females, and age ranged between 18 and 64 (median 22). The experiment was programmed in z-Tree (Fischbacher, 2007); we used ORSEE (Greiner, 2015) for recruiting. A session lasted for about 120 min. Payoffs were converted at an exchange rate of 1 Euro per 75 ECUs; payoffs accrue over all periods. Subjects earned on average 15.64 Euros, including a show-up fee of 4 Euros.

The results section is structured as follows: First we show that noise influences institutional choices in a systematic way: Citizens opt predominantly for *DecPun* in ONE, for *NoPun* in POINT-FIVE, while all three institutions attract similar shares in POINT-NINE (Result 1). Then we relate institutional choices to punishment behavior and show that punishment towards cooperative citizens predicts exit in *DecPun* and *CenPun* (Result 2). In the next step, we analyze contributions and profits and show that both decrease when signals become noisy (Result 3). We then demonstrate for the final stage of the game that central authorities choose punishments close to the deterrent levels, while punishment in *DecPun* is typically stronger (Result 4). Finally, we show that, under imperfect information, authorities who avoid misguided punishment gain a competitive advantage and are able to attract the majority of citizens (Result 5).

4.1. Choice of institution

For the choice of institution in the first phase, *NoPun* attracts the majority of the population in all treatments. About two thirds of the subjects choose this institution in POINT-NINE and even more so in the other two treatments. This is in line with the results of Gürek et al. (2006), who also find that their punishment institution is not popular early in the game. Centralized punishment initially attracts 21% of the citizens in POINT-NINE, compared to 13 and 7% in ONE and POINT-FIVE, respectively. These differences in the initial choice of institutions are significant across treatments ($p = .027$, Fisher's exact test). Over time, most citizens move to the two punishment institutions. Comparing the three institutional choices across treatments we can reject our null hypothesis: The allocation of citizens into *DecPun*, *CenPun*, and *NoPun* is significantly different across treatments ($F(2.49, 72.3) = 3.41$, $p = .029$ for all phases; $F(2.85, 82.7) = 3.14$, $p = .032$ for the final phase, Pearson χ^2 statistic with correction for dependence within matching group, see Rao and Scott, 1984).

The top panels of Fig. 1 show the average choice of institutions for each treatment. Across all phases we find evidence for our three conjectures: In ONE, the modal choice is clearly *DecPun*, while in POINT-FIVE the modal choice is *NoPun*. In POINT-NINE the modal choice is *CenPun*, although only by a small margin over the two other institutions.[17]

The bottom panels of Fig. 1 show the relative share of the institutions over time. In all treatments, *NoPun* loses a lot of citizens during the first three phases. Most of the adjustments happen through the first half of the 32 periods and we observe relatively stable shares of institutions in the second half of the experiment in ONE and POINT-FIVE. In POINT-NINE, the share of *CenPun* is stable, but *NoPun* loses in favor of *DecPun* throughout the 32 periods. Thus, while the evidence supports our Conjectures 1 and 3, the results are less clear with regard to Conjecture 2, which postulated the dominance of *CenPun* in treatment POINT-NINE. Summarizing our results on the choice of institutions we find:

Result 1. Institutional choices are significantly affected by the level of noise in the signals. After some early adjustments, citizens choose predominantly *DecPun* in ONE, while *NoPun* retains highest shares in POINT-FIVE. In POINT-NINE all three institutions attract similar shares of the population.

In the next step we want to take a closer look at the determinants for the choice of an institution. Recall that when citizens can move between institutions, they are informed about the outcomes in the three institutions. In particular, citizens learn (i) the number of citizens, (ii) the average contribution, and (iii) the average profits earned in each of the three institutions in all previous periods. We use multinomial probit models to explain the choice of institution between phases. For each citizen we observe seven institution choices with information about the outcome of the prior phase. In Model (1) of Table 1 we explain the choice of institution by the average profit of the citizens in each institution in the previous phase.[18] We use two dummies for the treatments ONE and POINT-NINE, with POINT-FIVE being the omitted case. We also add two dummies for the institution in which the subject is currently in, with *NoPun* as the omitted case, and we add a linear time trend (variable *Phase*). The treatment dummies indicate that citizens are less likely to choose *NoPun* over *DecPun* in the two treatments with relatively accurate or perfect information.

We find evidence for inertia in the choice of institution. Having been in *NoPun* before significantly increases the chance of choosing *NoPun* relative to *DecPun*, as shown by the significant negative effects of both institution dummies. The coefficients of the three profit variables show that this information is indeed a strong determinant for the institutional choice. Observing high profits in *NoPun* significantly increases the probability of choosing *NoPun* over *DecPun* for the next phase, while the opposite is true for high profits in *DecPun*. The profits in *CenPun* do not seem to affect the choice between *NoPun* and *DecPun*. The estimates for choosing *CenPun* (the second set of

[17] One-sample Pearson χ^2 tests for the null hypothesis of equal probabilities for all three institutions (corrected for dependence within matching group) are insignificant for ONE ($p = .196$) and POINT-NINE ($p = .892$), and significant for POINT-FIVE ($p = .003$). In the final phase we have ONE: $p = .051$, POINT-NINE: $p = .064$, and POINT-FIVE: $p = .538$.

[18] In case there were no citizens in a given institution we cannot observe a profit. In the estimates we use the same profit as in the case when there is only one citizen in a given institution.

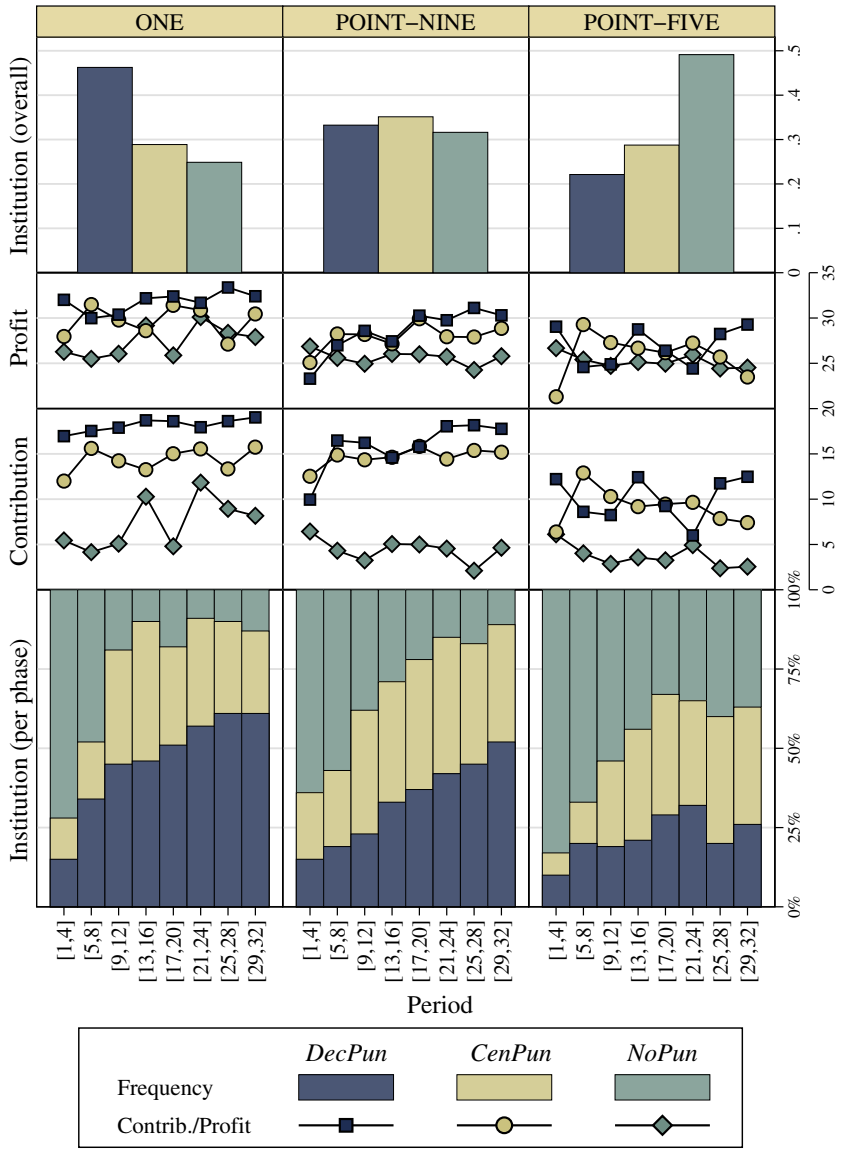

Fig. 1. Top panel: Average choice of institution over all periods and by treatment. Middle panels: Average profits and contributions in *NoPun*, *DecPun*, and *CenPun* across time. Dots show averages in a phase of four periods. Bottom panel: Choice of institution during the eight phases.

covariates in Table 1) show a very similar pattern. High profits in *CenPun* increase the probability of choosing that institution for the next phase over *DecPun*, while the opposite is true for high profits in *DecPun*.

Although the relation between relative profits and institution choice is strong, it is not informative with regard to the ultimate causes of the relative attractiveness of the institutions, because profits are merely a result of the activities in a given phase. The profits are mainly linked to contributions (for *NoPun* they are linearly dependent). If we replace the profits by contributions in Model (1) of Table 1 we get very similar results (not shown in the table), that is, high contributions in an institution increase the probability of choosing the respective institution. However, the main source of the relative popularity of the two punishment institutions should be determined in the way the citizens and the authority use the punishment option.

In Model (2) of Table 1, we investigate the use of the punishment option as a determinant of institution choice. Just adding the frequency or strength of punishment used in a given institution is, however, not an adequate measure of how well cooperation norms are enforced. Punishment can not only hit low contributors, but also at high contributors. We classify received punishment into free-rider punishment (if the punished citizen contributed less than the group average) or contributor punishment (otherwise). We replace the covariates for the profits by variables measuring free-rider, and contributor punishment, interacted with the dummy for the two institutions allowing for punishment. The results in the upper half of Table 1 show that punishment in *DecPun* increases the probability of leaving the institution in favor of *NoPun*. Interestingly this holds both for free-rider punishment and contributor punishment, although the latter effect seems to be stronger (the coefficients are not significantly different). In the lower half of the table we find clear evidence that the occurrence of contributor punishment is decisive for the choice between the two institutions allowing for punishment. High contributor punishment in *DecPun* significantly increases the probability of choosing *CenPun*, and vice versa. The strength of free-rider punishment, on the other hand, does not significantly affect the choice between these two institutions.

Table 1
Choice of institution.

	Dependent variable: Institution in $t+1$			
	(1)		(2)	
Choose NoPun				
ONE	−0.417***	(0.139)	−1.001***	(0.240)
POINT-NINE	−0.298**	(0.131)	−0.644***	(0.205)
DecPun	−1.740***	(0.181)	−2.137***	(0.196)
CenPun	−0.824***	(0.163)	−0.894***	(0.188)
Phase	−0.003	(0.026)	−0.113***	(0.031)
Profit *NoPun*	0.100***	(0.018)		
Profit *DecPun*	−0.124***	(0.014)		
Profit *CenPun*	−0.007	(0.011)		
Free-rider pun × *DecPun*			0.045**	(0.020)
Contributor pun × *DecPun*			0.117***	(0.045)
Free-rider pun × *CenPun*			0.074**	(0.035)
Contributor pun × *CenPun*			0.040	(0.066)
Constant	1.753***	(0.586)	1.607***	(0.162)
Choose CenPun				
ONE	−0.208	(0.165)	−0.473*	(0.287)
POINT-NINE	−0.072	(0.162)	−0.183	(0.285)
DecPun	−0.778***	(0.190)	−1.286***	(0.180)
CenPun	0.787***	(0.162)	1.219***	(0.168)
Phase	−0.006	(0.028)	−0.034	(0.032)
Profit *NoPun*	0.035**	(0.016)		
Profit *DecPun*	−0.123***	(0.015)		
Profit *CenPun*	0.133***	(0.015)		
Free-rider pun × *DecPun*			0.028	(0.021)
Contributor pun × *DecPun*			0.114**	(0.046)
Free-rider pun × *CenPun*			−0.010	(0.028)
Contributor pun × *CenPun*			−0.125**	(0.057)
Constant	−1.005*	(0.577)	0.313	(0.202)
Wald χ^2-test	1724.2		606.2	
p	0.000		0.000	
N	2100		2100	

Notes: Multinomial probit estimates. Dependent variable: Chosen institution for the next phase (*DecPun* is the omitted case). Independent variables are treatment dummies (POINT-FIVE as omitted case), dummies for the institution in the previous phase (*NoPun* as omitted case), Phase, average profits in the actual phase in the respective institution, and free-rider and contributor punishment in the respective institutions during the previous phase. Robust standard errors, clustered on matching group, in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Result 2. The choice of institutions is importantly influenced by the punishment behavior in the previous periods. In particular the amount of punishment towards cooperative citizens (contributor punishment) significantly predicts exit, both for *DecPun* and *CenPun*.

When exiting an institution where punishment is possible citizens can either opt for *NoPun* or for the alternative punishment institution. In ONE and POINT-NINE the majority of citizens leaving *DecPun* opt for *CenPun* in the next phase (71.4 and 65.5%), while in POINT-FIVE we observe a majority of moves towards *NoPun* (57.7% of the cases). For citizens deciding to leave *CenPun* the results are less clear. In all treatments we observe a move to *DecPun* in slightly more than half of the cases (ONE: 57.4%, POINT-NINE: 56.4%, POINT-FIVE: 51.3%).

Given the crucial role of punishment of cooperative citizens in the choice of institution we will analyze the punishment behavior in response to the treatment variation in more detail in Section 4.3. Before we do that, we focus on two other outcome variables of interest, contributions and profits.

4.2. Contributions and profits

Varying the noise in the contribution signals does not only affect institutional choices, but also the degree to which citizens manage to mitigate free-rider problems within their population. Table 2 shows the average contributions in the three treatments. Averages over all institutions are highest in ONE, followed by POINT-NINE, and POINT-FIVE. These treatment differences are significant at $p = .000$ (Kruskal–Wallis test on matching group averages). The same holds for average profits across institutions ($p = .002$).

The averages per institutions show that both contributions and profits are typically highest in *DecPun*, followed by *CenPun* and *NoPun*.[19] The middle panels of Fig. 1 show the average profits and contributions in the three institutions over time. In most of the phases profits are higher for the two institutions allowing for punishment, but overall differences are not pronounced. This is not surprising, given that there is free movement between institutions every fourth round.

The profits of *DecPun* are comparable to the results of Grechenig et al. (2010, p. 861) (GNT hereafter), who use the same signal technology and marginal per capita return, but randomly assign subjects to an institution and use a partner matching with groups of four subjects. In ONE (POINT-FIVE) we observe somewhat higher average profits of 31.6 relative to 29.7 in GNT (25.9 vs. 24.2 in GNT).[20] In POINT-NINE the profits are very similar (28.8 vs. 28.5 in GNT). On the other hand, profits in *NoPun* are lower than in the corresponding treatments in GNT (26.1 vs. 28.5 in GNT for ONE, and 25.4 vs. 28.3 in GNT for POINT-FIVE, POINT-NINE not available).

[19] We calculate the averages based on all individual decisions or outcomes. The overall average is therefore not equal to the average of the values for the three institutions.

[20] The values of GNT are corrected for the fact that their punishment endowment was higher (10) than in the present experiment (3).

Table 2
Contributions and profits.

		ONE	POINT-NINE	POINT-FIVE
Contribution	Overall	14.2	12.3	6.6
	NoPun	5.8	4.7	4.0
	DecPun	18.4	16.6	9.8
	CenPun	14.4	14.8	9.0
Profit	Overall	29.2	27.3	25.4
	NoPun	26.1	25.5	25.4
	DecPun	31.6	28.8	25.9
	CenPun	29.2	27.9	25.7

Notes. Average contributions and profits for the three treatments, both overall and for each institution separately. Averages are calculated based on individual observations.

Fig. 1 suggests a positive correlation between the share of the population and average profits in an institution. This correlation is strong and significant for ONE ($\rho = .711$, $p = .000$, correlations based on matching group averages for the three institutions) and POINT-NINE ($\rho = .700$, $p = .000$), while the two are virtually uncorrelated in POINT-FIVE ($\rho = -.016$).

Table 3 shows the results from OLS estimates explaining contributions. In Model (1) we include treatment dummies and two controls for time effects. The first identifies the period within each phase of four periods, the second controls for the time trend over the course of the eight phases. The results confirm that the treatments POINT-NINE and ONE result in significantly higher contributions. Interestingly contributions significantly drop within a phase, but at the same time significantly increase over the eight phases.

In Model (2) we add dummies for the institutions to the model (with *NoPun* as omitted case), and we control for the number of citizens in the institution. We allow for non-linear effects of group size by adding a quadratic term as well (n and n^2). While the controls in Model (1) are clearly exogenous, this is no longer the case for Model (2). The results should therefore not be interpreted in a causal way. Both institutions with a punishment are related to higher contributions compared to *NoPun*. Furthermore, the coefficient for *DecPun* is significantly larger than for *CenPun* ($p = .014$). These results are similar when using profits instead of contributions as dependent variable (models not shown in the table): Profits are highest in *DecPun*, followed by *CenPun*, and *NoPun*. The differences are significant in the overall sample. The result that *DecPun* leads to higher profits that *CenPun* is, however, mainly driven by treatment ONE. Both in contributions and in profits the differences become insignificant if we estimate the models based on the treatments POINT-NINE and POINT-FIVE only ($p > .2$).[21]

With respect to group size it seems that either large or small groups are conducive to high contributions, whereas groups of around five citizens tend to have lowest contributions.[22] In the remaining models (3)–(5), we repeat the estimates for the three treatments separately. In all three treatments contributions are highest in *DecPun*, followed by *CenPun*. The differences are, however, only significant in ONE. It seems that higher levels of noise close the gap in contributions between *DecPun* and *CenPun*. For the group size we observe a significant u-shaped effect with a minimum around five to six participants for ONE and POINT-FIVE, while the coefficients are insignificant for POINT-NINE.

Result 3. Contributions and profits decrease as the information about other players' behavior becomes noisy. The two institutions with punishment result in higher contributions than *NoPun*. Contributions in *DecPun* tend to be higher than in *CenPun*, but the difference becomes small and insignificant under noise.

4.3. Punishment strategies

Our results above suggest that the use of the punishment option importantly influences the choice of institutions. In Section 2.2 we derived an expression for minimal deterrent punishment (see Eq. (7)). In the following we compare the punishment decisions observed in the experiment to this theoretical benchmark. Recall that the expression requires to specify a contribution level $\bar{g}$ to be enforced. To calculate the benchmark we set $\bar{g}$ to a 'typical' contribution level. More precisely, we set $\bar{g}$ equal to the median of the signals an authority in *CenPun*, or a citizen in *DecPun* receives in a given period.[23]

The left panel of Fig. 2 shows the results for the punishment of authorities in the three treatments. On the horizontal axis we depict the difference between the signal and the median signal in the group ($s_i - \bar{g}$). For example, a value of −20 refers to the case where the signals indicate that the citizen in question contributed zero and the majority of the other citizens contributed fully. The bars indicate the average number of punishment points meted out for the respective deviation. The horizontal lines show the minimal deterrent punishment, which is decreasing in the signal for free riders (negative deviations) and zero thereafter.[24] The top left panel of Fig. 2 shows the results for ONE. Punishment clearly follows the theoretical pattern, but tends to be somewhat lower than predicted, with the exception of moderate negative deviations of four to two units. Zero or positive deviations trigger almost no punishment. For POINT-NINE we observe that punishment for negative deviations tends to be very close to the predicted level, while again misguided punishment seems to be negligible. We will, however, argue below that misguided punishment still plays an important role for the popularity of the central authority. In POINT-FIVE, we observe a different pattern. For negative deviations punishment is much lower than deterrent punishment. In addition, punishment seems almost invariant across the deviation classes.[25]

In a next step, we want to contrast the punishment in *CenPun* to the punishment of the citizens in *DecPun*. For the benchmark we assume that each citizen calculates $\bar{g}$ on the basis of the signals she receives, including her own contribution. In addition, for groups with more than two citizens we assume that each citizen punishes other citizens by $\frac{1}{d-1}$ (with d denoting the number of citizens in *DecPun*) of the minimal deterrent punishment according to Eq. (7). The right panel of Fig. 2 shows the results for *DecPun*. In ONE we observe that

[21] Alternatively we estimated the model with all observations, but allowed for an interaction between *DecPun* and ONE. For both profit and contributions the interaction is positive and significant, while the difference between the dummies for *CenPun* and *DecPun* becomes insignificant ($p > .27$).

[22] Thus, unlike what could have been expected from the literature discussed in footnote 8, the fact that the marginal per capita return decreases in group size does not produce a similar monotonic pattern in the contributions. In case of decentralized punishment this is in line with previous research (Carpenter, 2007; Roux and Thöni, 2015).

[23] For even numbers of signals we slightly deviate from the usual calculation of the median and take the higher of the two middle values, that is, in case of the signals {20, 18, 12, 0} we set $\bar{g} = 18$. Alternatively one could argue that punishers should always try to enforce full contributions and thus set $\bar{g} = 20$. This would result in higher predicted levels of punishment.

[24] The expression in Eq. (7) is linearly decreasing in s_i for $\bar{g} > s_i$, while the horizontal lines in Fig. 2 are not. The reason for this is that we combine all cases with various values for $s_i - \bar{g}$ and c into a single average per bar.

[25] We use exact Wilcoxon signed rank tests for the difference between prediction and data (matching group averages). In case of ONE none of the differences are significant ($p > .125$). For POINT-NINE we observe a significant difference for the bars $[-8, -6]$ ($p = .039$), and $[-4, -2]$ ($p = .006$), all others are insignificant ($p > .4$). In case of POINT-FIVE all but one differences are highly significant.

Table 3
Contributions.

	Dependent variable: Contribution				
	All observations		ONE	POINT-NINE	POINT-FIVE
	(1)	(2)	(3)	(4)	(5)
POINT-NINE	5.719***	3.745***			
	(1.262)	(0.908)			
ONE	7.591***	4.587***			
	(1.216)	(0.799)			
Period in phase	−0.592***	−0.592***	−0.395*	−0.462**	−0.921***
	(0.097)	(0.097)	(0.203)	(0.158)	(0.077)
Phase	0.816***	0.092	0.239**	0.172	−0.260***
	(0.138)	(0.067)	(0.078)	(0.135)	(0.071)
DecPun		9.572***	11.707***	11.066***	5.417***
		(0.685)	(0.557)	(0.630)	(1.293)
CenPun		7.352***	7.860***	9.528***	5.197***
		(0.789)	(1.798)	(1.199)	(0.677)
n		−1.616***	−1.680*	−0.172	−2.976***
		(0.459)	(0.772)	(0.576)	(0.630)
n^2		0.152***	0.136**	0.042	0.233***
		(0.035)	(0.059)	(0.041)	(0.049)
Constant	4.363***	7.130***	10.625***	4.840**	15.564***
	(0.979)	(1.731)	(2.701)	(1.519)	(1.812)
F-test	84.6	237.9	184.8	350.1	49.0
Prob > F	0.000	0.000	0.000	0.000	0.000
R^2	0.215	0.445	0.451	0.446	0.217
N	9160	9160	3064	3052	3044

Notes: OLS estimates. Dependent variable: contribution. Independent variables: treatment dummies (with POINT-FIVE as omitted case), period within a phase (1 − 4), and phase (1 − 8), dummies for the institution (*NoPun* as omitted case), and two measures for the number of citizens in the institution, n, n^2. Robust standard errors, clustered on matching group, in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

punishment for negative deviations is substantially higher than predicted. This holds also for POINT-NINE, where in addition we observe a clear increase in misguided punishment relative to *CenPun*. Finally, POINT-FIVE leads again to punishments far from deterrent and largely invariant in the deviation.[26]

Result 4. In the treatments ONE and POINT-NINE average central authorities' punishments match the predicted minimal deterrent punishment patterns surprisingly well. In contrast, decentralized punishment for low signals is substantially higher than the theoretical benchmark. For POINT-FIVE punishments are largely invariant to the signal and lower than the deterrent level.

Despite the fact that citizens face a second order free-rider problem in the punishment stage and have to bear the marginal cost of punishment, it seems that decentralized norm enforcement is *stronger* than centralized norm enforcement. Consequentially, as shown in Fig. 1 decentralized punishment institutions tend to achieve higher contributions but lose some of the efficiency gains for stronger punishment. From the estimates shown in Model (2) of Table 1 we learned that free-rider punishment increases the likelihood of choosing *NoPun*, but does not seem to affect the choice between the two punishment institutions. On the other hand, contributor punishment significantly affects the relative popularity of the two punishment institutions. Misguided punishment is (highly) likely to result in contributor punishment in POINT-FIVE (POINT-NINE). Furthermore, when the signal is wrong, then a fraction of the punishments for signals below $\bar{g}$ result in contributor punishment. Taken together these observations suggest a rationale for the shift of the competitive advantage from *DecPun* towards *CenPun* once we move from perfect information to low noise. Central authorities tend to be more moderate in their punishment relative to the citizens in *DecPun*, both for free-rider and misguided punishment. Under noise, this leads to substantial differences of punishment for high contributors. For example, in POINT-NINE average punishment received by a full contributor is 0.48 units in *DecPun* compared to 0.27 *CenPun* ($p = .050$ exact Wilcoxon signed rank test based on matching group averages).[27]

When formulating Conjecture 2, we stressed the role of retaliatory motives for misguided punishment in *DecPun*. Note that our experimental design makes targeted revenge difficult, because citizens do not know who is responsible for the punishment (unless there are only two citizens in *DecPun*). Nevertheless citizens might punish others in response to punishment received in the previous period. To investigate the role of retaliation we ran regressions explaining punishment decisions. In a model controlling for the deviation in signals, time, and group size effects we find a small but significant positive effect of received punishment in the previous round on misguided punishment ($\beta = 0.0052$, $p = .006$, clustered standard errors). If we interact the term with the treatment POINT-NINE we find the main coefficient unchanged and the interaction small and insignificant ($\beta = -0.0007$, $p = .813$). From this we conclude that revenge is a motive for misguided punishment in *DecPun*, but we do not see a strong indication that the effect varies across treatment.

Fig. 2 suggests that average levels of misguided punishment are low, in particular in *CenPun*. This conclusion might, however, be premature. Because institutions are endogenous, popular authorities are strongly overrepresented in this analysis. In the next subsection we show that the authorities and the populations of citizens differ in their use of misguided punishment, which is a crucial determinant for the choice of institution under imperfect information.

[26] All differences between prediction and data are at least weakly significant in ONE. In POINT-NINE the bars [−16, −14] and [−12, −10] are slightly above 10%, while all other comparisons are significant at $p < .031$ (same test as in Footnote 25).

[27] In ONE, we observe also a significant difference ($p = .040$), but in both cases average punishment is very small (0.05 and 0.03) such that the difference presumably does not matter anymore.

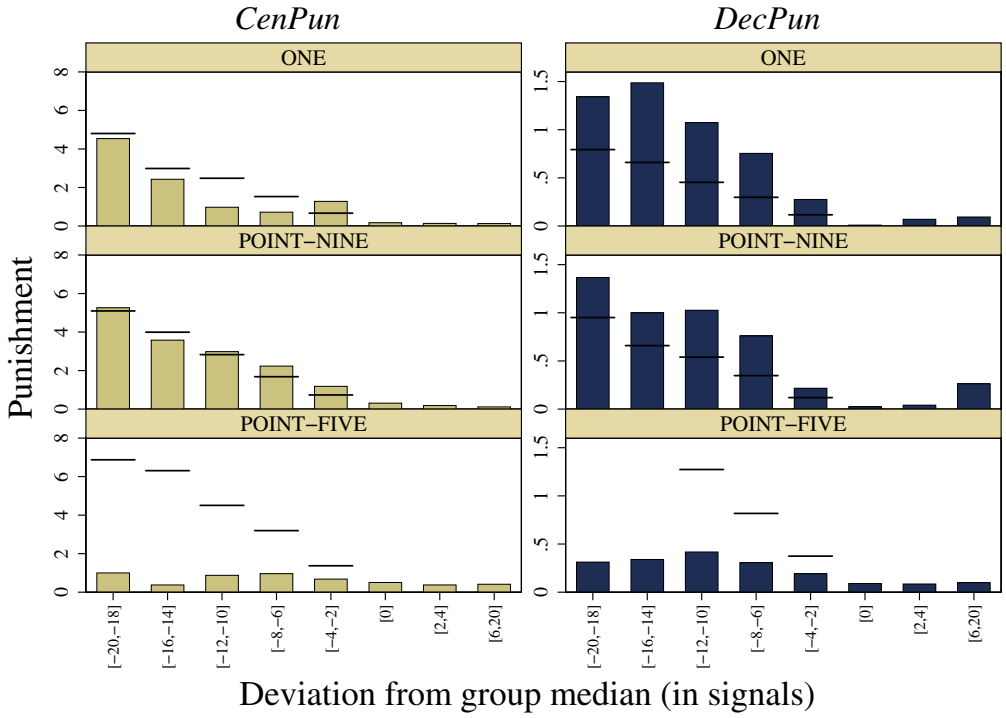

Fig. 2. Predicted and actual punishment of authorities in *CenPun* (left panel) and citizens in *DecPun* (right panel) for the three treatments. Bars show average punishment targeted at a citizen dependent on the difference between the citizen's signal and the median signal in the group. Horizontal lines show the average of all deterrent punishments according to the theoretical prediction (values for the two first bars in the bottom right panel are outside the plotted range).

4.4. Misguided punishment and the popularity of an institution

In the previous sections we provided evidence that contributor punishment is a crucial determinant of entry into and exit out of an institution. At the same time, misguided punishment occurs in *DecPun* and—to a lesser extent—also in *CenPun*. We now investigate the role of misguided punishment for the relative popularity of a punishment institution.

In the following we derive a measure for the relative strength of misguided punishment in *DecPun* and *CenPun*. In particular, we calculate for each population the frequency of punishment targeted at citizens with above average signals. For *CenPun* we use the punishment data from the subject in the role of the authority, while for *DecPun* we calculate the average over all the punishment decisions of citizens in *DecPun*. Populations in which we observe less frequent misguided punishment in *CenPun* than in *DecPun* are classified as populations with a 'good' authority. Conversely, if the authority metes out more misguided punishment than the citizens we speak of a population with a 'bad' authority.[28] Our classification is based on the data of phases 1–7 and we explain the institutional choices in the final phase.[29] In ONE and POINT-FIVE this criterion leads to an equal split of the matching groups, while in POINT-NINE we classify 40% of the matching groups as populations with a good authority.

Panel A of Fig. 3 shows that authorities attract only a small fraction of the citizens in ONE and POINT-NINE when they mete out a lot of misguided punishment relative to the citizens in *DecPun*. Panel B shows that good authorities manage to attract a larger share than bad authorities in all treatments. However, only under imperfect information *CenPun* is clearly the modal choice. Under perfect information not even good authorities are able to gain the support of the majority of the population.[30]

Panels C and D of Fig. 3 provide information about the stability of the population in *CenPun* over time. Bars show the fraction of citizens in this institution, divided into incumbents (darker part) and immigrants (lighter part). Incumbents are citizens who were already in *CenPun* in the preceding phase; immigrants are citizens who were previously in *DecPun* or *NoPun*. The graph shows that bad authorities have a high turnover. Most of the time, more than half of their population are immigrants. Populations of good authorities are much more stable, with a large fraction of the citizens remaining in the institution.

Panel D also shows that, unlike in POINT-NINE, good authorities continuously lose support in the second half of the experiment in ONE. Presumably the low differences in misguided punishment between *CenPun* and *DecPun* are not sufficiently important for many citizens to be willing to subordinate themselves to the Leviathan.

Instead of dividing the observations in two groups we can also use the difference between the frequency of misguided punishment in *DecPun* and *CenPun* as a continuous measure of an authority's relative performance in punishing. If we use OLS to regress the share of the population in *CenPun* in the final phase (the middle bars in Fig. 3) on

[28] The results in this analysis are robust with respect to a number of alternative criteria for good and bad authorities. In particular, the result that *CenPun* is clearly the modal institution in POINT-NINE does also hold if we (i) define good authorities by a median split according to misguided punishment in *CenPun* (ignoring punishment in *DecPun*), or (ii) if we define good authority by the average strength of contributor punishment in *CenPun* (instead of the frequency).

[29] In ONE we have three matching groups for which the punishment data is missing because the citizens never chose the respective institution. In these cases we use the average of the corresponding figures in the other matching groups in the same treatment as an estimate for misguided punishment. Qualitatively the results are the same if we drop the three matching groups from the sample: *DecPun* remains clearly the modal institution for good (67.5%) as well as bad authorities (46.7%).

[30] Testing for differences in the institutional choices between good and bad authorities (Panel A and B of Fig. 3) results in $F(1.66, 48.1) = 3.07, p = .064$ (Pearson χ^2 statistic with correction for dependence within matching group).

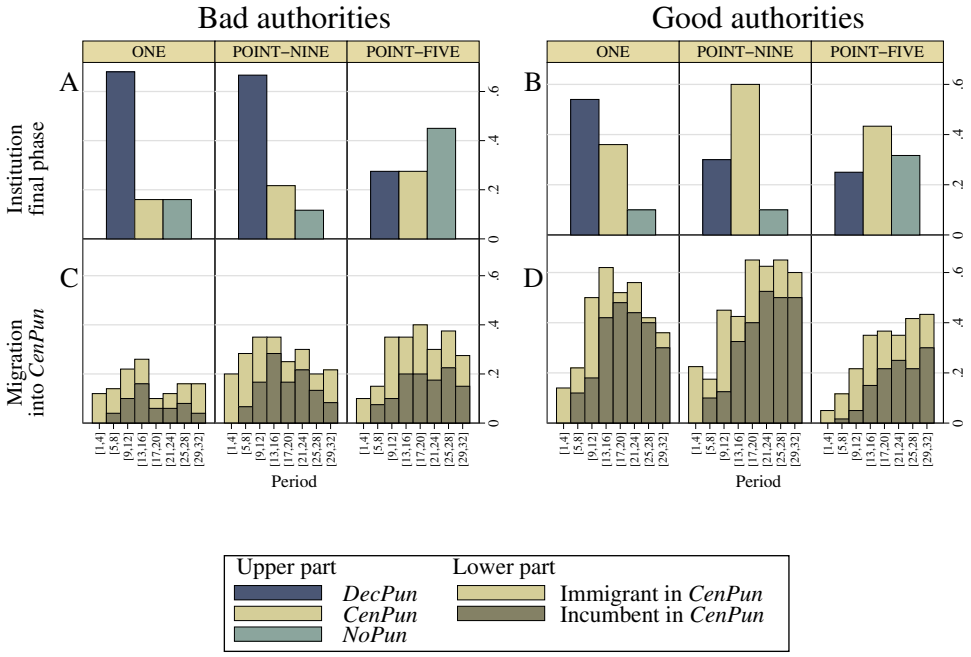

Fig. 3. Choice of institution in the final phase of the game for matching groups with bad (panel A) and good authorities (panel B). Bars show the fraction of participants choosing *DecPun, CenPun, NoPun*, separated by the treatments with perfect information ONE and imperfect information POINT-NINE, POINT-FIVE. Panels C and D: Migration patterns in *CenPun*. The dark part of the bars shows the citizens who were in *CenPun* already in the previous phase (incumbents); the light part shows the immigrants.

this measure we observe a highly significant positive effect for POINT-NINE ($\beta = .721$, $p = .001$, robust standard errors, group averages as observations), but not for the other two treatments.

Result 5. Authorities who assign less misguided punishment than the citizens attract a larger share of the population in all three treatments. In the two treatments with imperfect information good authorities attract more citizens than the other two institutions, while in ONE decentralized punishment remains the modal institution.

5. Discussion

Our study analyzes information conditions which lead subjects to voluntarily subordinate themselves to a central authority. We vary the accuracy of the information concerning the contributions of the other participants in the group. We show that an environment with perfect information tends to favor the decentralized punishment institution, while a high level of noise favors an institution with no punishment. Under low levels of noise we observe the highest support for centralized punishment.

In line with the literature on the importance of antisocial punishment we observe that the punishment of cooperative subjects plays a crucial role for the popularity of an institution. Institutions that punish cooperative citizens tend to lose citizens in favor of other institutions. In the treatments ONE and POINT-NINE we observe that punishment is stronger under decentralized punishment than under centralized punishment, which is remarkable given that centralized punishers do not have to bear the marginal cost of punishment, while decentralized punishers do. In ONE, citizens can easily avoid punishment by contributing (nearly) fully. In POINT-NINE, however, a fraction of the punishments targeted at citizens with low signals ends up with high contributors. While this holds for both punishment institutions equally, the fact that punishment is stronger in the decentralized institution leads to more contributor punishment in this institution. In addition we observe higher levels of punishment targeted at citizens with a high signal in decentralized punishment. These two observations explain why decentralized punishment loses support when we introduce imperfect information.

Taking into account differences between the central authorities we show that the interaction between imperfect monitoring and the availability of a central authority who prevents punishment of cooperative citizens boosts the choice of centralized punishment institutions. This is not the case for the treatment with perfect information. In this treatment there is very little punishment of cooperative citizens and citizens prefer not to delegate their punishment power.

According to the data reported in Herrmann et al. (2008) we conducted the experiments in a society with very low levels of antisocial punishment. In a subject pool with higher levels of antisocial punishment centralizing punishment might be more attractive, even under perfect information. Furthermore, the Leviathans in our experimental design have full discretion in the use of their power: There are no legal constraints, no noblesse oblige which limit the actions of authorities. Real world authorities typically face institutional and moral constraints. Moreover, there are arguably better selection mechanisms for authorities in place. Presumably, these societies come close to the outcome of good authorities in POINT-NINE, where *CenPun* is clearly the dominant institution. Therefore, we might underestimate the attractiveness of centralized punishment in our experiment.

We consider our treatment variations as prototypical for various epochs of the evolution of social structures in humans. Early societies allowing for nearly perfect observation of others tend to apply decentralized punishment regimes. In maturing societies with increasing agglomeration and complexity, it becomes difficult to monitor others' behavior. These are the circumstances, in which people are willing to sacrifice some of their autonomy and delegate the sanctioning power to a Leviathan. In times of social unrest and destabilized law enforcement systems, however, punishment by authorities becomes more erratic. Under these circumstances centralized sanctions lose their competitive advantage and, if possible, citizens migrate to other institutional arrangements.

Recently, the appearance of new media like social networks and mobile communication technologies give rise to another interesting development, as they increase transparency of actions among group members. As a consequence, we might expect a decentralization of the societal structures. The latest developments on the administration of mass protests during the Arab Spring via social networks are an example for this development (Hussain and Howard, 2013). Whether this is a first indication for a general shift towards more decentralized organizational structures is too early to tell.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jpubeco.2016.09.008.

References

Ambrus, A., Greiner, B., 2012. Imperfect public monitoring with costly punishment: an experimental study. Am. Econ. Rev. 102 (7), 3317–3332. http://dx.doi.org/10.1257/aer.102.7.3317.

Ambrus, A., Pathak, P.A., 2011. Cooperation over finite horizons: a theory and experiments. J. Public Econ. 95 (7-8), 500–512. http://dx.doi.org/10.1016/j.jpubeco.2010.11.016.

Andreoni, J., Gee, L.K., 2012. Gun for hire: delegated enforcement and peer punishment in public goods provision. J. Public Econ. 96 (11–12), 1036–1046. http://dx.doi.org/10.1016/j.jpubeco.2012.08.003.

Bochet, O., Page, T., Putterman, L., 2006. Communication and punishment in voluntary contribution experiments. J. Econ. Behav. Organ. 60 (1), 11–26. http://dx.doi.org/10.1016/j.jebo.2003.06.006.

Carpenter, J.P., 2007. Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. Games Econ. Behav. 60 (1), 31–51. http://dx.doi.org/10.1016/j.geb.2006.08.011.

Carpenter, J.P., Matthews, P.H., 2012. Norm enforcement: anger, indignation, or reciprocity? J. Eur. Econ. Assoc. 10 (3), 555–572. http://dx.doi.org/10.1111/j.1542-4774.2011.01059.x.

Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. Exp. Econ. 14 (1), 47–83. http://dx.doi.org/10.1007/s10683-010-9257-1.

Nicolò, Corazzini, L., Kube, S., Marécal, M.A., 2014. Elections and deceptions: an experimental study on the behavioral effects of democracy. Am. J. Polit. Sci. 58 (3), 579–592. http://dx.doi.org/10.1111/ajps.12078.

Dal Bò, P., Foster, A., Putterman, L., 2010. Institutions and behavior: experimental evidence on the effects of democracy. Am. Econ. Rev. 100 (5), 2205–2229. http://dx.doi.org/10.1257/aer.100.5.2205.

De Quervain, D.J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., 2004. The neural basis of altruistic punishment. Science 305 (5688), 1254–1258. http://dx.doi.org/10.1126/science.1100735.

Denant-Boemont, L., Masclet, D., Noussair, C.N., 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Econ. Theory 33 (1), 145–167. http://dx.doi.org/10.1007/s00199-007-0212-0.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298. http://dx.doi.org/10.1016/j.geb.2003.06.003.

Ertan, A., Page, T., Putterman, L., 2009. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. Eur. Econ. Rev. 53 (5), 495–511. http://dx.doi.org/10.1016/j.euroecorev.2008.09.007.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315. http://dx.doi.org/10.1016/j.geb.2005.03.001.

Fehr, E., Gäter, S., 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90 (4), 980–994. http://dx.doi.org/10.1257/aer.90.4.980.

Fehr, E., Herz, H., Wilkening, T., 2013. The lure of authority: motivation and incentive effects of power. Am. Econ. Rev. 103 (4), 1325–1359. http://dx.doi.org/10.1257/aer.103.4.1325.

Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114 (3), 817–868. http://dx.doi.org/10.1162/003355399556151.

Fehr, E., Williams, T., 2013. Endogenous emergence of institutions to sustain cooperation.

Fischbacher, U., 2007. Z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178. http://dx.doi.org/10.1007/s10683-006-9159-4.

Gächter, S., Herrmann, B., Thöni, C., 2005. Cross-cultural differences in norm enforcement. Behav. Brain Sci. 28 (6), 822–823. http://dx.doi.org/10.1017/S0140525X05290143.

Gächter, S., Renner, E., Sefton, M., 2008. The long-run benefits of punishment. Science 322 (5907), 2008. http://dx.doi.org/10.1126/science.1164744.

Grechenig, K., Nicklisch, A., & Thöni, C. Punishment despite reasonable doubt—a public goods experiment with sanctions under uncertainty. J. Empir. Leg. Stud. 7 (4), 847–867. 10.1111/j.1740-1461.2010.01197.x.

Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125. http://dx.doi.org/10.1007/s40881-015-0004-4.

Gross, J., Méder, Z.Z., Okamoto-Barth, S., Riedl, A., 2016. Building the Leviathan — voluntary centralisation of punishment power sustains cooperation in humans. Sci. Rep. 6, 20767. http://dx.doi.org/10.1038/srep20767.

Grosse, S., Putterman, L., Rockenbach, B., 2011. Monitoring in teams: using laboratory experiments to study a theory of the firm. J. Eur. Econ. Assoc. 9 (4), 785–816. http://dx.doi.org/10.1111/j.1542-4774.2011.01026.x.

Gürek, Ö., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312 (5770), 108–111. http://dx.doi.org/10.1126/science.1123633.

Hamman, J.R., Weber, R.A., Woon, J., 2011. An experimental investigation of electoral delegation and the provision of public goods. Am. J. Polit. Sci. 55 (4), 738–752. http://dx.doi.org/10.1111/j.1540-5907.2011.00531.x.

Herrmann, B., Thöi, C., Gächter, S., 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367. http://dx.doi.org/10.1126/science.1153808.

Hirschman, A.O., 1970. Exit, Voice and Loyalty. Responses to Decline in Firms, Organizations and States. Harvard University Press, Cambridge, MA.

Hirschman, A.O., 1978. Exit, voice, and the state. World Polit. 31 (1), 90–107. http://dx.doi.org/10.2307/2009968.

Hobbes, T., 1651. The Leviathan. Oxford World's Classics SeriesOxford University Press, Incorporated, 1996.

Hussain, M.M., Howard, P.N., 2013. What best explains successful protest cascades? ICTs and the fuzzy causes of the Arab Spring. Int. Stud. Rev. 15 (1), 48–66. http://dx.doi.org/10.1111/misr.12020.

Kamei, K., Putterman, L., 2015. In broad daylight: fuller information and higher-order punishment opportunities can promote cooperation. J. Econ. Behav. Organ. 120, 145–159. http://dx.doi.org/10.1016/j.jebo.2015.09.020.

Kosfeld, M., Okada, A., Riedl, A., 2009. Institution formation in public goods games. Am. Econ. Rev. 99 (4), 1335–1355. http://dx.doi.org/10.1257/aer.99.4.1335.

Kube, S., Traxler, C., 2011. The interaction of legal and social norm enforcement. J. Public Econ. Theory 13 (2006), 639–660. http://dx.doi.org/10.1111/j.1467-9779.2011.01515.x.

Leibbrandt, A., Ramalingam, A., Sksvuori, L., Walker, J.M., 2015. Incomplete punishment networks in public goods games: experimental evidence. Exp. Econ. 18 (1), 15–37. http://dx.doi.org/10.1007/s10683-014-9402-3.

Markussen, T., Putterman, L., Tyran, J.-R., 2014. Self-organization for collective action: an experimental study of voting on sanction regimes. Rev. Econ. Stud. 81 (1), 301–324. http://dx.doi.org/10.1093/restud/rdt022.

Nicklisch, A., Grechenig, K., Thöni, C., 2015. Information-sensitive Leviathans: the emergence of centralized punishment. WiSo-HHWorking Paper Series, Working Paper No. 24.

Nikiforakis, N., 2008. Punishment and counter-punishment in public good games: can we really govern ourselves? J. Public Econ. 92, 91–112. http://dx.doi.org/10.1016/j.jpubeco.2007.04.008.

Nikiforakis, N., Noussair, C.N., Wilkening, T., 2012. Normative conflict and feuds: the limits of self-enforcement. J. Public Econ. 96 (9-10), 797–807. http://dx.doi.org/10.1016/j.jpubeco.2012.05.014.

Nosenzo, D., Quercia, S., Sefton, M., 2015. Cooperation in small groups: the effect of group size. Exp. Econ. 18 (1), 4–14. http://dx.doi.org/10.1007/s10683-013-9382-8.

O'Gorman, R., Henrich, J., Van Vugt, M., 2009. Constraining free riding in public goods games: designated solitary punishers can sustain human cooperation. Proc. R. Soc. B 276 (1655), 323–329. http://dx.doi.org/10.1098/rspb.2008.1082.

Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without a sword: self-governance is possible. Am. Polit. Sci. Rev. 86 (2), 404–417. http://dx.doi.org/10.2307/1964229.

Putterman, L., Tyran, J.-R., Kamei, K., 2011. Public goods and voting on formal sanction schemes. J. Public Econ. 95 (9-10), 1213–1222. http://dx.doi.org/10.1016/j.jpubeco.2011.05.001.

Rao, J.N.K., Scott, A.J., 1984. On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. Ann. Stat. 12 (1), 46–60. http://dx.doi.org/10.1214/aos/1176346391.

Rockenbach, B., Milinski, M., 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444 (7120), 718–723. http://dx.doi.org/10.1038/nature05229.

Roux, C., Thöni, C., 2015. Collusion among many firms: the disciplinary power of targeted punishment. J. Econ. Behav. Organ. 116, 83–93. http://dx.doi.org/10.1016/j.jebo.2015.03.018.

Sutter, M., Haigner, S., Kocher, M.G., 2010. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. Rev. Econ. Stud. 77 (4), 1540–1566. http://dx.doi.org/10.1111/j.1467-937X.2010.00608.x.

Thöni, C., 2014. Inequality aversion and antisocial punishment. Theor. Decis. 76 (4), 529–545. http://dx.doi.org/10.1007/s11238-013-9382-3.

Tyran, J.-R., Feld, L.P., 2006. Achieving compliance when legal sanctions are non-deterrent. Scand. J. Econ. 108 (1), 135–156. http://dx.doi.org/10.1111/j.1467-9442.2006.00444.x.