Discussion Papers in Economic Behaviour

DPEB
04/17

On the emergence of a sanctioning institution

>Adriana Alventosa
Universitat de València, ERI-CES, Spain
>Gonzalo Olcina
Universitat de València, ERI-CES, Spain

www.erices.es

On the emergence of a sanctioning institution

Adriana Alventosa*, Gonzalo Olcina*

September 5, 2017

Abstract

This paper theoretically studies the emergence of a sanctioning institution in a selfish and wealth-diverse group where the provision of a public good is realized only once. In particular, we present a public goods game where players are given the opportunity to implement a sanctioning institution by hiring an external enforcer which sanctions free-riding behavior. However, the enforcer's effectiveness will not be guaranteed and will depend on the level of effort he exerts to chase these opportunistic attitudes. Whether the sanctioning institution is implemented or not is a task delegated to a government concerned in its persistence, who will represent the interests of a social class with a particular level of wealth. The emergence of the sanctioning institution will depend on a set of institutional and technological parameters, the wealth distribution in the society and the identity of the social class whose interests are represented by the government. Given these exogenous variables, the sanctioning institution will emerge more easily if the government represents the social class with the lowest opportunity cost in the provision of a public good. If implemented, the sanctioning institution can achieve a positive provision of such good if the society counts with a relatively high quality in its sanctioning institutions and high social return of the public good. The case of heterogeneous valuations of the public good will also be proved to show symmetric results.

Keywords: public goods game, cooperation, wealth inequality, pool punishment, moral hazard.

JEL classification: C72, D02, H41.

*ERI-CES, University of Valencia.

The authors thank the comments of participants in Public Economic Theory PET 2016 (Rio de Janeiro, 2016), the 27th International Conference on Game Theory (Stony Brook, New York, 2016) and the Workshop in Industrial Organization (Valencia, 2017). The authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness and the project EC02014-58297-R.

1 Introduction

Public goods provision has been extensively discussed as a social dilemma. In one-shot interactions among selfish individuals, any of them will prefer to free ride in the contribution and keep their endowments as private investment. To tackle this problem, the consideration of social preferences jointly with the introduction of peer punishment has been the most standard way of implementing a mechanism to address the coordination issue. Peer punishment consists on the opportunity for each individual to penalize, at the end of the game, those participants who have been free riders at a cost. However, in real life, formal individual punishment is hard to implement as individuals themselves have no choice of sanctioning their peers.

In today's societies, sanctioning is mostly not individually decided but delegated to a monitoring figure or institution. Moreover, the implementation of this agent is not chosen once the outcomes are observed, but must be agreed before the game even starts. The novel concept of *pool punishment* was introduced by Sigmund, Hauert and Traulsen (2011) as a mechanism that captured how investments in monitoring and sanctioning institutions are made. Several studies have introduced a punishing figure that automatically applies sanctions at the end of the game (Kosfeld, Okada and Riedl, 2009; Gürerk, Irlenbusch and Rockenbach, 2006; Okada, 1993). This way, the punishment is outsourced and put aside from players' interests. However, beyond automatic punishment in case of defection, which is no longer a realistic approach to the mechanisms operating in these contexts, institutions are run by individuals with their own goals and their own interests.

Furthermore, the standard literature has mainly considered homogeneity in the contributors' endowments. Undoubtedly, this is not realistic. However, more importantly it can provide misleading results as it hides the different incentives individuals with different levels of wealth have when it comes to contributing. In real life, when contributing to a public good individuals benefit equally from the generated return. However, *will everybody be interested in contributing to the same extent?* Experimental evidence has already shown that wealth heterogeneity affects the contribution decision (Cherry, Kroll and Schrogen, 2005; Burlando and Guala, 2005; Buckley and Crosson, 2006; Reuben and Riedl, 2013). As we will see in this paper, at an individual basis and given an expected fine, individuals with lower wealth have greater incentives to contribute to the public good provided the net gains of free riding increase with wealth. Furthermore, wealth heterogeneity also affects the implementation of the sanctioning institution. The decision of whether to implement such an institution or not is going to be made by a government representing the interests of a social class with particular incentives for contributing.

The objective of this paper is to analyze whether the implementation of a sanctioning institution of these characteristics will be able to attain cooperation among unequally rich individuals and under which conditions will this institution emerge.

To do so, we theoretically model a society conformed by selfish individuals[1] with different levels of wealth, i.e. different individual endowments from which to contribute to a public good.

[1]The selfishness assumption clearly holds in relevant cases such as international agreements among countries or bargaining among companies. Beyond these cases, the introduction of social preferences will encourage cooperation making the positive provision of a public good occur more easily. Therefore, we are considering the worst possible case, where selfishness prevails.

Without loss of generality, we will assume individuals either belong to a poor class, a middle class or a rich class. Being their selfish preferences known by everybody, they have the opportunity of implementing a sanctioning institution before contributions are made, which will take action at the end of the game. This sanctioning institution can be reasonably seen as a county sheriff, who, being a strategic agent, is subject to moral hazard issues. In real life, subjects don't make every single social decision, but will delegate this task to an elected government on their behalf. Moreover, governments are rarely altruistic bodies, but will also have their own interests. In this case, we consider that the government's interest is to be reelected in future elections and will, therefore, maximize the utility of that citizen that guarantees its persistence. We call this individual the political decisive agent. As we will see, who the political decisive agent is and what incentives for contributing does he have, turns out to be a crucial issue in the implementation of the sanctioning institution.

The objective can therefore be reformulated as the analysis on how does wealth inequality and the identity of the political decisive agent affect the provision of a public good among selfish individuals when punishment is implemented by an external punishing institution subject to moral hazard problems.

In our context, if the sanctioning institution is not implemented by the government or it is done with inappropriate incentives, there would be no contribution. Our first result claims that if the institution emerges, the public good provision can be positive. At an individual basis, the higher an individual's wealth is, the lower his incentives to contribute under the threat of a sanctioning institution will be[2]. The total provision, will depend on the interplay between the quality of institutional and technological variables (efficiency in detecting free riders, fine paid in case of punishment and social return of the public good) and the severity of wealth inequality (society's wealth levels and its proportions in the population). If the institutional and technological variables reach sufficiently high levels, the institution could achieve full contribution, which would maximize social welfare. Otherwise, only partial contribution could be achieved with a resultant welfare loss proportional to the number of free riders.

In second place, we characterize under which conditions will the institution emerge. The sanctioning institution will be implemented as long as the return of the public good exceeds the individual cost of the institution (salary paid to the sheriff) plus the opportunity cost of the political decisive agent. This opportunity cost can either be the contribution in case he is a contributor or the expected fine in case he is a free rider. Hence, who the political decisive agent is becomes decisive, as the government representing the political decisive agent with the lowest opportunity cost will implement the institution in a greater range of cases. Furthermore, the incentives of the political decisive agent are determinant. For very low expected fines, a government representing a free-rider is better (will implement the institution in a greater range of cases) than a government representing a contributor. For sufficiently high expected fines, however, the government representing the poor class will always be better than any other.

Concerning efficiency, if the sanctioning institution achieves full cooperation, implementing it will always be social welfare maximizing. While a poor-class government will always implement

[2]This contrasts with the result obtained in the literature regarding step-level public good games, where wealthier individuals have stronger incentives to contribute (Rapaport, 1988).

the institution in this case, a middle or a rich-class government may not do it. This, perhaps puzzling, result is due to the wealth inequality among contributors. Given that all of them receive an even part of the public good, those who contribute with the lowest amount will always be better off, whereas it is not clear-cut for the others. If, however, the sanctioning institution achieves partial contribution, a government representing a political decisive agent with the lowest opportunity cost will implement the institution in situations where it is inefficient to do so and the government representing a political decisive agent with the highest opportunity cost will decide not to implement the institution in situations where it is efficient to do so.

Finally, we extend this model to heterogeneous valuations of the public good, commonly known as the marginal per capita return, and show that results are symmetric to those of heterogeneous wealth: individuals which assign a higher value to having public goods will indeed contribute to a larger extent. This agrees with experimental evidence shown in previous studies (Fellner, Iida, Kröger and Seki, 2011; Fisher, Isaac, Schatzberg and Walker, 1995; Reuben et al. 2013).

1.1 Related Literature

This paper is mainly related with previous literature dealing with pool punishment. Sigmund at al. (2011) modeled the comparison between peer and pool punishment in a public goods game with and without counter-punishment, i.e. the punishment of those who cooperate, but do not punish. Milinski,Traulsen and Röhl (2012) reproduce this model experimentally, using their same assumptions. Their main result is that when pool punishment is combined with counter-punishment, contributions increase. Furthermore, in their experiments, pool punishment clearly prevailed over peer punishment.

Beyond pool punishment, this study is related to previous literature dealing with sanctioning institutions. Kosfeld et al. (2009) portray the institution formation through a public goods game. In their model, which they also take to the laboratory, players must first decide whether or not to form a sanctioning institution at a cost. If the institution is formed, at the end of the game free riders will be automatically punished for their deviation. Both theoretical and experimental results show the endogenous formation of these institutions, which enhance cooperation and have positive effects on group cooperation. Okada (1993) studies this mechanism for a more general prisoners' dilemma. Furthermore, previous studies have also proved the superiority of centralized institutions with respect to decentralized ones in terms of efficiency. Tommasi and Weinschelbaum (2014) formally expose the advantages and disadvantages each one presents. Fehr and Williams (2013) experimentally show how effective centralized institutions emerge endogenously as dominant firms to which most finally migrate. Acemoglu and Wolitzky (2015) approach the problem by endogenizing specialized enforcement remarking the importance of incentives in fraud chasing. All but this last paper, present automatic institutions, ignoring the possibility of strategic institutions with private information, characteristics considered in this study.

Additionally, this paper is related to economic history literature concerning the historical emergence of institutions over time. Not only nowadays, but throughout history, individuals have tended to group themselves and develop centralized sanctioning institutions with the power of punishing defectors. These institutions haven't been automatic punishers, but have had their

own incentives in the maintenance of social and political order. In the case of Genoa, for example, in the period 1194-1339 a *poderestia* system was established after the failure of the genoese *commune*, incapable of adapting to socio-economical changes. The transition reflected local learning from past institutions introducing an additional strategic player (*podestà*) that needed to be appropriately motivated to implement the desired outcome. This figure, who had coercive power and decision-making ability, had to reinforce cooperation among clans but should necessarily be limited, having no incentives to become a dictator or to side with any genoese clan (Greif, 2006). Similarly, even previously in time (11^{th} century), merchants in Medieval Europe created guilds with implicit contractual relations and a specific communication-mechanism (Greif, 1993). These examples show that including an external enforcer with its own underlying preferences is not only realistic today, but also reflects the historical emergence of these kind of institutions.

The rest of the paper is organized as follows. Section 2 describes the model: a public goods game with an external enforcer. In the next place, Sections 3 and 4 will provide the solution for the model, presenting the paper's results. In the following section, a comparative statics analysis will be carried out, after which an extension to heterogeneity in the valuation of the public good will be made. Finally, the last section will sum up the main results obtained.

2 The model

Consider the following n-player public goods game (PGG, hereinafter). There are $n \geq 2$ risk neutral players belonging to z social classes of q^j individuals each one. Each player has a private endowment or level of wealth $\omega_i^j \in [\underline{\omega}, \overline{\omega}]$ (where $i = 1, ..., n$ identifies the individual and $j = 1, ..., z$ the social class) from which he can contribute $g_i^j \leq \omega_i^j$ to a public good. Each social class is characterized for being composed by individuals with the same endowment, thus, with slight abuse of notation, we will indistinctly use $\omega_i^j = \omega^j$ as the wealth of individual i in social class j.

Given the contribution of the n players captured by the vector of contributions g, the material payoff of player i from social class j is equal to:

$$\pi_i^j(g) = \omega_i^j - g_i^j + \frac{\lambda}{n}[\sum_{i=1}^{n} g_i] \tag{1}$$

where $1 < \lambda < n$ is the factor by which the public fund is multiplied, also known as the marginal social return of the public good. Assumption $\lambda < n$ implies that zero contribution is the dominant action for every player with standard selfish preferences, i.e. each player's payoff is maximized by contributing zero to the public good regardless of the other players' contributions[3]. In consequence, the strategy profile $g_i = 0 \ \forall_i$ is the unique Nash Equilibrium. Assumption $\lambda > 1$ implies that all players are better off if everybody contributes with his full wealth to the public good. In fact, the strategy profile $g_i = \omega_i \ \forall_i$ is welfare maximizing.

This game gives rise to a cooperation issue where the stage game Nash Equilibrium is inefficient. Punishment has formerly been introduced as a mechanism with the purpose of attaining

[3]This happens because the individual marginal return of the public good is less than 1, which is the marginal return of private investment.

this socially desired cooperation. However, given the characteristics of this game (one-shot with selfish preferences), if at the end players were given the opportunity to peer punish each other at a cost, nobody would do so. This occurs because selfish players who maximize their material payoff will never reduce their profit in order to detriment others when they're only interacting once. Hence, another type of sanctioning must be implemented in order to make punishment an effective mechanism for attaining cooperation among selfish individuals.

To this purpose, we introduce an external enforcer, let's say, a sheriff, which is in charge of monitoring and implementing the punishment under a pre-designed contract. Hence, even though punishment is also implemented at the end of the game, this is done by this employed external agent. Furthermore, citizens' representative must decide *beforehand* whether it is in their interest to have this enforcer in the game and formalize a contract previous to the contribution decisions. In this case, a government representing some particular interests designs a contract for the sheriff, which can be accepted or discarded by him. After observing whether there have been any free riders in the contribution stage, the sheriff chooses the level of non-verifiable effort to generate evidence for the courts of the criminal offense and eventually punish this fraud.

The precise sequence of actions is as follows:

Contract Design Stage- A government designs and offers a contract contingent on verifiable outcomes for the sheriff. This contract can be accepted or rejected by the sheriff. We assume that the government aims to remain in office (for instance, be reelected in future elections) and so maximizes the equilibrium utility of the political decisive agent (PDA, henceforth) that can guarantee a majority of votes or its permanence in power. Let's define ω^* as the wealth of this PDA.

Traditionally, the identity of the PDA has been attributed to the median voter. The median voter theory, first suggested by Hotelling (1929) and formally proposed by Black (1948) is extensively accepted in political economics theory as a fundamental property of democracy: an electioneer will try to maximize his number of votes by focusing on the median voter's welfare. A median player is either poor or middle class, but in no case rich. Societies worldwide are diverse, each one of them showing its own particularities, but what almost all of them have in common is a societal structure where rich conform a minority.

Nevertheless, we are specially concerned in the case where there is a corrupt or politically distorted government concerned in safeguarding the rich class interests. This can be approached in two different ways. On the one hand, democracy is not a global phenomenon yet and there are still situations with concealed plutocracies or even dictatorships. On the other hand, citizens are not politically identical when there is wealth inequality and political influence matters. In fact, it is common to see that political influence increases as wealth increases, making the PDA be above the median one (Przeworski, 2015). Given that these scenarios are not isolated examples, we have considered appealing analyzing this case as well. Thus, we will consider that any social class could potentially be the PDA.

However, notice that we will not analyze how the PDA is determined. We are interested in the comparison among the performance of different governments in terms of public good provision, where a government represents the interests of a decisive agent and has the power to create

and enforce the sanctioning institution.

If the sheriff accepts the contract offered by the government, he will have the opportunity of exerting two possible levels of costly effort to detect free riders: a low level of effort or a high level of effort. Let's define c_e as the cost of exerting effort e, where $e \in \{L, H\}$. We assume that the cost of exerting a high effort is greater than the cost of exerting a low effort. For the sake of simplicity, exerting a low effort can be interpreted as making no effort at all: $c_L = 0$; and $c_H = c$ where $c > 0$. If the sheriff exerts low effort, fraud will be detected and punished with probability p_L. If, however, the sheriff exerts high effort, free riders will be detected and punished with probability p_H, where $0 < p_L < p_H < 1$. Even though players are unable of observing the level of effort, they can observe, in the final stage, the game's outcomes. This general approach allows for a situation with moral hazard in a multiprincipal-agent context to exist, where the sheriff (agent) chooses a non-verifiable action, effort exerted in pursuing free riding, but consequences are taken over by the players (principals).

The contract will specify the sheriff's salary (s_k) for each possible outcome: (i) nobody has free ridden (s_0), (ii) some agent has free ridden and the sheriff has punished (s_p) or (iii) some agent has free ridden and the sheriff has not punished (s_{np}). We assume that these outcomes are perfectly observable and verifiable by everybody. Thus, a contract will be defined by the triplet $\{s\} = \{s_0, s_p, s_{np}\}$. The sheriff is risk-neutral with utility function given by $u = s_k - c_e$,where $k \in \{0, p, np\}$ represent the outcomes. For simplicity, we assume that the sheriff's reservation utility is zero, $\overline{u} = 0$. Moreover, we assume that the sheriff has limited liability.

Contribution Stage - Each player i will individually and simultaneously decide the level of contribution to the public good g_i^j. Those who do not contribute with their whole wealth to the public good ($0 \leq g_i^j < \omega_i^j$) will be considered free riders. On the other hand, if they contribute with their whole wealth, they will be considered contributors ($g_i^j = \omega_i^j$).

All citizens observe the size of the public fund, but cannot observe who has contributed with how much. If $\sum_{i=1}^{n} g_i^j = \sum_{i=1}^{n} \omega_i^j$, every player has behaved as a contributor and the sheriff's intervention is not necessary. In this case, he is paid the fixed salary s_0, the fund is equally divided among the players and the game ends at this point. Here, the case with information asymmetry on who has contributed and who has free ridden is not a matter of study. If the sheriff detects fraud, he will be able to detect its origin. The information asymmetry, however, lies on the enforcer's actions.

Punishment Stage- This last stage is only reached if the sheriff has accepted the contract and at least one of the players has free ridden. If free riding is indeed detected, information of who has been a contributor and who has been a free rider is perfectly observable by the sheriff. We assume the sheriff will never punish someone who has been a contributor. Notice that we are leaving out the chance of possible extortion from the sheriff to contributors by assuming a minimum institutional quality. The sheriff chooses the effort he will make in order to produce objective evidence on free-riding behaviour, with cost c_e. If fraud is detected it will automatically be punished with a fixed fine, $f > 0$.

Besides the case where no sheriff is contracted in the contract design stage, a player's final payoff is determined as follows:

$$\pi_i^j(g,\{s\},p_e) = \omega_i^j - g_i^j + \frac{\lambda}{n}[\sum_{i=1}^{n} g_i] - \gamma_e \tag{2}$$

where:

$$\gamma_e = \begin{cases} p_e\frac{s_p}{n} + (1-p_e)\frac{s_{np}}{n} & \text{if } \sum_{i=1}^n g_i < \sum_{i=1}^n \omega_i \text{ and } g_i^j = \omega_i^j \\ p_e\frac{s_p}{n} + (1-p_e)\frac{s_{np}}{n} + p_e f & \text{if } \sum_{i=1}^n g_i < \sum_{i=1}^n \omega_i \text{ and } g_i^j < \omega_i^j \\ \frac{s_0}{n} & \text{if } \sum_{i=1}^n g_i = \sum_{i=1}^n \omega_i \end{cases}$$

where g is the vector of contributions, $\{s\}$ is the contract $\{s_0, s_p, s_{np}\}$ and p_e are the conditional probabilities of fraud being detected, where $e \in \{L, H\}$.

Formally, this game is an n-player three-stage PGG where every player knows the course of the game in previous stages. In the following, we will characterize the set of Subgame Perfect Equilbria of the game allowing for moral hazard ($0 < p_L < p_H < 1$).

3 Effectiveness of the Sanctioning Institution: the Level of Public Good Provision

In this section we analyze the level of public good provision when the sanctioning institution has been formed, i.e. suppose the government has decided to hire the sheriff by offering him an acceptable contract in the contract design stage. For the sake of simplicity and without loss of generality, in the rest of the paper we assume that there are three social classes in this game, $z = 3$. Namely, a poor class ($j = P$), a middle class ($j = M$) and a rich class ($j = R$). Each social class has a total of q^j individuals with the same wealth level ω^j, which allows us to drop individual notation. Hence, a wealth distribution is characterized by a pair of vectors $\{(\omega^P, \omega^M, \omega^R), (q^P, q^M, q^R)\}$ where $q^P + q^M + q^R = n$. All results obtained for this number of groups can be generalized to any z.

To hire the sheriff with a contract such that he exerts low effort is not an interesting case so let's assume that if the sheriff has been offered a low-effort enhancing contract and indeed devotes little effort in fraud chasing, everybody's best response will be to free ride. Formally this will happen if the following assumption holds:

Assumption 1: $\omega^P \geq \frac{p_L f}{1-\lambda/n}$

Intuitively, the expected fine under a low-effort contract would be so small that for everybody, even the poorest individual, the net gains from free-riding would be larger. The unique Nash equilibrium in the continuation subgame after low effort will be that everybody free rides.

What we are going to analyze is with how much will each individual contribute to the public good under the threat of a sheriff exerting high effort. In order to do so, let's first introduce the following lemma, useful for the characterization of the individuals' best response function.

It shows that if any citizen belonging to a social class j free rides on the public good, he will do so with $g^j = 0$.

Lemma 1*: Given an initial wealth ω^j, free riding with $g^j = 0$ weakly dominates any other $g^j = \epsilon$, where $0 < \epsilon < \omega^j$.*

The intuition behind this lemma is as follows. If an individual decides to free ride he will have to pay the same expected salary of the sheriff plus the expected fine, no matter with how much he slopes off. In that case, it is rationally optimal for him to free ride with as much as possible.

The individual decision, therefore, sums up in whether to free ride with $g^j = 0$ or fully contribute with $g^j = \omega^j$. When deciding on this, citizens balance their individual net costs and net gains of free riding, and will contribute if the former ones are greater than the latter ones.

In a situation where everybody else contributes, the size of the fund before individual i's contribution is $\sum_{i=1}^{n} g_{-i} = \sum_{i=1}^{n} \omega_{-i}$, where $-i$ is the vector of players other than i, that is, (1,2 ...i-1, i+1, ...n). Individual i's decision is critical for the intervention of the sheriff. If everybody else has contributed, individual i from social class j will contribute as well as long as the net costs of free riding are greater than the net gains of doing so:

$$\frac{p_H s_p + (1 - p_H) s_{np} - s_0}{n} + p_H f \geq \omega_i^j (1 - \frac{\lambda}{n}) \tag{3}$$

Notice that the net costs of free-riding include both the expected fine and the per-capita increase in the salary of the sheriff. Let us denote by $\tilde{\omega}$ the critical value of ω_i^j such that this holds with equality. Notice that if individual i is endowed such that $\omega_i^j \leq \tilde{\omega}$, he will contribute as long as everybody else contributes as well.

In a situation where at least one other individual different from individual i from class j has free ridden, individual i's contribution is no longer critical in what concerns the sheriff's intervention. Now the net cost from free-riding is the possibility of being penalized (collected by the term $p_H f$).

Thus, in this case, individual i from social class j will contribute as long as:

$$p_H f \geq \omega_i^j (1 - \frac{\lambda}{n}) \tag{4}$$

Similarly, let's denote by $\hat{\omega}$ the critical value such that this holds with equality. If an individual has an initial wealth such that $\omega_i^j \leq \hat{\omega}$, he will contribute regardless of others' contributions. Notice that $\hat{\omega} \leq \tilde{\omega}$ always holds.

The following proposition summarizes individual i's best response function depending on the position of ω_i^j with respect to the obtained thresholds.

Proposition 1*: The best response function $BR^j(\cdot)$ of an individual from social class j is as follows:*

- If $\omega^j \leq \hat{\omega}$, then $BR^j(g) = \omega^j \ \forall g$
- If $\hat{\omega} < \omega^j \leq \tilde{\omega}$, then:

- $BR^j(g) = \omega^j$ when $\sum_{i=1}^n g_{-i} = \sum_{i=1}^n \omega_{-i}$
- $BR^j(g) = 0$ when $\sum_{i=1}^n g_{-i} < \sum_{i=1}^n \omega_{-i}$

- If $\tilde{\omega} < \omega^j$, then $BR^j(g) = 0 \ \forall g$

If an individual i is sufficiently poor ($\omega_i^j \leq \hat{\omega} \leq \tilde{\omega}$) he will always contribute to the public good, regardless of what others do. This type of individuals are unconditional contributors given that their net costs of free riding are always greater than their net gains. At the other end of the spectrum, if an individual i is sufficiently rich ($\hat{\omega} \leq \tilde{\omega} < \omega_i^j$) he will always free ride, given that his net gains of doing so are always greater than his net costs. These individuals are unconditional free riders. However, it could also happen that an individual had an intermediate wealth ($\hat{\omega} < \omega_i^j \leq \tilde{\omega}$) such that his best response is to contribute only if everybody else does so and to free ride if there is at least one free rider. Let's call these individuals conditional cooperators.

Now we are ready to compute the Nash equilibria of the contribution subgame when the sheriff exerts high effort. These equilibria will depend on the existing wealth distribution in the group. Although we leave the details of the proof for the appendix, let us provide some intuition before formally stating the result.

Notice that in many situations the equilibrium is going to be unique. For instance, it could be the case that everybody had a sufficiently low initial wealth such that they all had as a dominant action to contribute. In this case, where $\omega^R \leq \hat{\omega}$, everybody would contribute with their whole endowment, $g^j = \omega^j \ \forall_j$. Therefore, this is the unique equilibrium in this contribution subgame. On the other hand, it could happen that everybody had a sufficiently high wealth such that everybody preferred to free ride. In particular, if $\tilde{\omega} < \omega^P$, then the unique equilibrium would be $g^j = 0 \ \forall_j$. The interesting cases arise for wealth distributions such that we have different mix of individuals according to their best responding behaviour.

For instance, notice that if the population is composed by unconditional free riders and conditional cooperators ($\hat{\omega} \leq \omega^P \leq \tilde{\omega} \leq \omega^R$), then using successive elimination of dominated actions, conditional cooperators will also free ride. Thus, we obtain $g^j = 0 \ \forall_j$ as the unique equilibrium. However, it could also happen that a proportion of the individuals were unconditional contributors, but there were also conditional cooperators and unconditional free riders in the population, that is : $\omega^P \leq \hat{\omega} \leq \tilde{\omega} < \omega^R$. In this case the poorer individuals will contribute no matter what, the richer individuals will free ride no matter what, and the conditional cooperators will also free ride given that there are free riders. Thus, the unique equilibrium in this case is that everybody with a wealth below $\hat{\omega}$ contributes whereas everybody above this critical value fully free rides.

Nevertheless, in the cases where there are no unconditional free-riders in the population there will exist multiple equilibria in the contribution subgame. For example, it could be the case that everybody were conditional cooperators ($\hat{\omega} \leq \omega^P \leq \omega^R \leq \tilde{\omega}$) or that a proportion were unconditional contributors while the rest were conditional cooperators ($\omega^P \leq \hat{\omega} \leq \omega^R \leq \tilde{\omega}$). Therefore, everybody contributing will be a Nash equilibrium in the subgame. However, there will be another equilibrium where individuals from classes with a level of wealth above $\hat{\omega}$ do not contribute.

In these latter cases of multiplicity, an equilibrium selection has been made. We claim that the prediction in the subgame for the former wealth distributions ($\hat{\omega} \leq \omega^P \leq \omega^R \leq \tilde{\omega}$) will be $g_i = 0 \ \forall_i$ and for the latter one ($\omega^P \leq \hat{\omega} \leq \omega^R \leq \tilde{\omega}$) the equilibrium selection will be $g_i = \omega_i^j$ for

players with $\omega_i^j \leq \hat{\omega}$ and $g_i = 0$ for players with $\hat{\omega} < \omega_i^j$.

There are two reasons that explain why this equilibrium selection criterion has been applied. Firstly, our proposed solutions are more stable equilibria. In the previous situations, starting from the cooperative equilibrium of universal contribution it would be enough that one individual deviated to free riding for the rest to apply their best response to this deviation and switch to the non-cooperative equilibrium. However, from the non-cooperative equilibrium, an individual deviation towards contribution and allowing the rest of the group to apply their best responses, would not lead to the cooperative equilibrium. Thus, the cooperative equilibrium with full contribution is not robust to"small mistakes" or 'mutations" while the non-cooperative equilibrium is robust.

Additionally, we have tried to stay conservative, choosing the worst possible scenario, that is the inefficient equilibrium with a lower level of contribution. This selection makes $\hat{\omega}$ the unique critical value for the characterization of equilibria of the contribution subgame when the sheriff exerts high effort, summarized in Proposition 2.

Proposition 2: *Given a wealth distribution* $\{(\omega^P, \omega^M, \omega^R), (q^P, q^M, q^R)\}$ *and assuming the sanctioning institution has been hired and exerts high effort, the selected equilibria at the contribution subgame are:*

- If $\omega^R \leq \hat{\omega}$, then $g^j = \omega^j \ \forall j$
- If $\omega^P \leq \hat{\omega} \leq \omega^M$, then $g^j = \omega^j$ for all poor class citizens and $g^j = 0$ for middle class and rich class citizens.
- If $\omega^M \leq \hat{\omega} \leq \omega^R$, then $g^j = \omega^j$ for all poor and middle class citizens and $g^j = 0$ for rich class citizens.
- If $\hat{\omega} < \omega^P$, then $g^j = 0 \ \forall j$

where $\hat{\omega} = \frac{p_H f}{1-\lambda/n}$

See formal proof in the appendix.

Notice that according to this proposition, the higher an individual's wealth is, the less incentives he will have to contribute, provided that net gains of free riding increase with the level of wealth. Taking a different approach: if the quality of the pertinent sanctioning institution is high, either because the probability of fraud legal detection under high effort is high (p_H) or because the fine in case of punishment is large (f), or similarly, if a society assigns a high value to the public good (high social return of the public good λ), then more citizens will be incentivized to contribute. Given a fixed group size, societies with a relatively high quality sanctioning institutions and high social return of the public good will have higher levels of potential contribution.

Recall the introduction of a sanctioning institution aims to solve the full free-riding outcome in the provision of a public good among selfish individuals interacting once. Thus, the effectiveness of this sanctioning institution can be measured by the size the public good reaches, that is by the sum of the individual contributions: $\sum_{\omega^j \leq \hat{\omega}} q^j \omega^j$.

***Corollary 1**: Given a wealth distribution $\{(\omega^P, \omega^M, \omega^R), (q^P, q^M, q^R)\}$ and given that the sanctioning institution has been implemented and exerts high effort, the level of public good provision will be:*

$$\sum_{\omega^j \leq \hat{\omega}} q^j \omega^j$$

where: $\hat{\omega} = \frac{p_H f}{1-\lambda/n}$

Thus, given a fixed population size and a high effort sanctioning institution, the amount of pubic good provision will increase (or remain constant[4]) if the social return of the public good λ increases, if the probability of fraud detection under high effort (p_H) increases, or if the fine (f) increases. Intuitively, if a society assigns a greater value to the provision of this type of goods or the quality of sanctioning institutions improves, contributing would become more attractive (or free riding less appealing), so contributions would increase and, therefore, a more efficient outcome will be obtained. Additionally, an increase in the population size, captured by n will diminish the provision of the public good. This phenomenon is often referred as the *1/n problem*.

4 When Does a Centralized Sanctioning Institution Emerge?

After characterizing the provision of public good under a high-effort sanctioning institution, let's present our main result which concerns the condition that must be met for a sanctioning institution to emerge in this environment. Before players contribute, a contract $\{s\}$ characterizing the three possible salaries s_0, s_p, s_{np} must be designed for the external enforcer. Let's suppose there is a government who proposes this contract with the sheriff, as explained in the model. This government represents the interests and tries to maximize the utility of the political decisive agent, whose wealth will be denoted with ω^*. As previously mentioned, we are not concerned in this paper with the determination of the identity of such political decisive agent. We rather focus on the effects of different governments representing different social classes' interests in the likelihood of the emergence of a welfare-enhancing sanctioning institution.

4.1 Contracts

Up until now we have assumed that contribution can only occur if the sheriff exerts high effort. However, for this to happen the sheriff must have incentives to choose high instead of low effort. In other words, the incentive constraint must be satisfied, as summarized in Lemma 2.

***Lemma 2**: Given a contract scheme $\{s\}$, the sheriff will exert a high level of effort if and only if $(p_H - p_L)(s_p - s_{np}) \geq c$. Otherwise, he will exert a low level of effort.*

Let's now characterize the minimum-cost contracts offered to the sheriff with the purpose of encouraging high or low effort, which do not depend on the type of government. The formal proof is relegated to the appendix.

***Lemma 3**: Assuming the government has all the bargaining power, the contracts offered to implement the different levels of effort are:*

[4]Notice that for there to be an increase, variations should be sufficiently high such that free riders become contributors.

- *High-effort contract:* $\{s^H\} = \{s_0 = 0, s_p = \frac{c}{p_H - p_L}, s_{np} = 0\}$.
- *Low-effort contract:* $\{s^L\} = \{s_0 = 0, s_p = 0, s_{np} = 0\}$.

Notice that whilst the low-effort contract is an acceptable contract with no economic rents, in case the government wants to encourage the exertion of a high level of effort, he will have to pay economic rents of size $\frac{p_L c}{p_H - p_L}$ due to the existence of moral hazard and limited liability [5]. The economic rents captured by the institution depend on the relative cost of high effort and on the likelihood ratio which measures how important is the existing moral hazard problem in the punishment phase.

The next question to answer is when would the government prefer to offer the high-effort contract. He will do so when the expected utility of the PDA is higher under the high-effort contract than under any other contract.

If the sheriff exerts low effort, everybody free rides, according to Assumption 1.In this case the political decisive agent's utility would be:

$$\pi^*(\{s^L\}, g^* = \omega^*) = \omega^* - p_L f \tag{5}$$

Recall that without the sheriff, free riding is the unique Nash equilibrium, i.e. $\pi^* = \omega^*$. Consequently, offering a low-effort contract is always weakly dominated by offering a contract which is not acceptable at all, given that $p_L, f \geq 0$. Hence, in case the government does not find it profitable to offer the high-effort contract, he will offer an unacceptable contract with any $s_k < 0$. Therefore, the government must, in fact, decide whether to offer an acceptable contract, which additionally encourages high effort, or an unacceptable one and do not hire a sheriff. In other words, the government's decision is whether to implement a high-effort institution or not.

4.2 Main Result

If the sheriff is offered the high-effort contract, the position of the wealth level of the political decisive player becomes crucial. The government will compare the utility of the decisive player with and without sheriff and will hire the sheriff offering him a high-effort contract if the net gains of contributing to the public good are greater than the expected costs of having an external enforcer. The next proposition characterizes under which conditions will the sanctioning institution be formed.

Proposition 4*: Assume that $\omega^P \leq \hat{\omega}$ and the political decisive agent has wealth ω^*. The sanctioning institution will emerge if and only if:*

$$\frac{\lambda}{n}[\sum_{\omega^j \leq \hat{\omega}} q^j \omega^j] \geq \frac{p_H s_k}{n} + \begin{cases} \omega^* & \text{if } \omega^* \leq \hat{\omega} \\ p_H f & \text{if } \omega^* > \hat{\omega} \end{cases}$$

where $\hat{\omega} = \frac{p_H f}{1 - \frac{\lambda}{n}}$, $s_k = 0$ *if* $\omega^R < \hat{\omega}$ *and* $s_k = \frac{c}{(p_H - p_L)}$ *if* $\hat{\omega} < \omega^R$.

Obviously the case where $\omega^P > \hat{\omega}$ lacks of any interest because then nobody is going to contribute even in the presence of a sheriff exerting high effort. Therefore we focus on the most

[5] Notice that under automatic punishment, which is equivalent to verifiable effort, it would be enough to pay $s_p = c$ to implement high effort and that $\frac{p_H c}{p_H - p_L} = c + \frac{p_L c}{p_H - p_L}$. Thus, $\frac{p_L c}{p_H - p_L}$ are the economic rents.

appealing cases where the punishment technology and the social return of the public good are sufficiently high to make at least one social class be willing to contribute.

According to Proposition 4, the emergence of a sanctioning institution that permits positive levels of provision of public good depends on the interaction of three factors: a set of institutional and technological parameters (λ, p_H, p_L, c, f), the existing wealth distribution in the group and the identity of the PDA. The institutional parameters include the social return generated by the public good λ and the several parameters that characterize the monitoring technology of the punishing institution. In particular these latter ones determine the expected capacity of punishment and the severity of the moral hazard problem generated by the non-verifiability of the external enforcer's efforts. This agency cost is captured by the economic rents obtained in the expected payoff of the sheriff ($\frac{p_H c}{(p_H - p_L)}$).

Recall that the level of public good provision attained when the institution is formed depends exclusively on the interaction of the first two factors which determine the relation between the critical value $\hat{\omega}$ and the wealth distribution dividing the population in contributors and free riders. If the resulting critical value $\hat{\omega}$ is sufficiently high compared to the level of wealth of rich individuals, then a full contribution equilibrium will be reached under the sanctioning institution. In this equilibrium all social classes contribute and the wage paid to the sheriff s_0 equals his reservation utility (zero in our model) because there is no free riding in equilibrium. The high quality of the punishing institutions captured by high values of p_H and f and the high returns of the public good λ build a credible and strong threat of punishment by the sheriff. For lower levels of quality of the sanctioning institution and of the social return of the public good ($\omega^P \leq \hat{\omega} < \omega^R$) we obtain a partial contribution equilibrium under the institution where only some social classes contribute while the others free ride.

A natural question is which government will implement the punishing institution more frequently. We will say that a government representing a PDA from class j (a j government) is better than a government representing a PDA from class k when for all the situations for which the k government implements the institution, the j government also does it. The next proposition is derived from proposition 4.

Proposition 5*: Given a wealth distribution* $\{(\omega^P, \omega^M, \omega^R), (q^P, q^M, q^R)\}$ *and a set of institutional and technological parameters (*λ, p_H, p_L, c, f*) the government under which the sanctioning institution emerges in a greater range of cases is the one representing the political decisive agent with the lowest opportunity cost.*

Notice that the return of the public good is the same for everybody and is fully determined by $\hat{\omega}$. The costs have a common element, which is the expected wage of the sheriff, and an element that depends on the opportunity cost each PDA has. In particular, a contributor renounces to his wealth while a free rider pays the fine with probability p_H.

Then it is easy to see that with the conditions for a full contribution equilibrium ($\omega^R < \hat{\omega}$) where all social classes contribute, a government representing a poor individual is better than a government representing a middle-class individual which in turn is better than a government representing a rich-class individual.

This result does not necessarily hold for partial contribution equilibria. If, for example, $\omega^P \leq \hat{\omega} \leq \omega^M$ only poor individuals will contribute to the public fund, being their opportunity

cost ω^P, while both middle class and rich class would free ride being their opportunity cost $p_H f$. If $\omega^P \leq p_H f$ a government representing the poor would hire the sheriff in a greater range of occasions than a middle or a rich class government. Otherwise, a middle or rich-class government would do so. However, when both poor and middle class individuals contribute because $\omega^M \leq \hat{\omega} \leq \omega^R$, they both sacrifice their endowments ω^P and ω^M respectively. Notice that as, by definition, $\omega^P \leq \omega^M$ there would be a greater range of cases where the sheriff is hired if the PDA were poor than if it were middle class. However, whether it holds more easily under a government representing the poor class or the rich class depends, again, on the relationship between ω^P and $p_H f$. Summarizing, we can state the following result:

***Corollary 2**: For very low expected fines ($\omega^P > p_H f$), a government representing a free-rider political decisive agent is better than a government representing a contributor political decisive agent. For sufficiently high expected fines ($\omega^P \leq p_H f$) the government representing the poor class is better than the government of the middle or the rich class.*

4.3 Social Welfare

At this point, an imperative question to address is what is the level of Social Welfare (SW hereinafter) achieved by the implementation of a sanctioning institution and, specially, how does such level compare with the level of SW obtained without the institution. We already know that without the described sanctioning mechanism everybody would free ride, yielding a SW equivalent to the weighted sum of the endowments: $q^P\omega^P + q^M\omega^M + q^R\omega^R$, denoted by W from now onwards.

As already mentioned in Section 2, full contribution is the efficient outcome with $SW = \lambda W$ which is greater than the outcome of full free riding W, given that $\lambda > 1$. If the parameters of the game are such that everybody contributes because $\omega^R \leq \hat{\omega}$, then $SW = \lambda W - s_0$. Recall that for our model $s_0 = \overline{u} = 0$. For this case or even for other cases with $\overline{u} > 0$ sufficiently close to 0, social welfare will be very close to the one obtained in the efficient outcome, i.e.

$$SW \approx \lambda W \tag{6}$$

Thus, the implementation of a sanctioning institution that enhances the exertion of high effort will achieve the fully efficient outcome.

However, the following question to address is whether the different governments representing the different political decisive agents (PDAs) would implement it or not. Recall that for the case of full contribution, a government representing a PDA with wealth ω^* will implement the institution if and only if:

$$\frac{\lambda}{n} W \geq \omega^* \tag{7}$$

or equivalently

$$\lambda W \geq n\omega^* \tag{8}$$

The following proposition summarizes the results on SW maximization with full contribution:

***Proposition 5**: If $\omega^R \leq \hat{\omega}$, a poor-class government will always implement the institution, making the socially efficient decision. If $\omega^R \leq \hat{\omega}$, a middle-class or rich-class government may not implement the institution in situations where it is socially efficient to do so.*

Intuitively, even though it is fully efficient to implement the institution if everybody contributes, a middle or rich-class government may not be interested in doing so if wealth inequality is too pronounced and/or the social return of the public good is too low. Each citizen receives an n^{th} part of the PG return composed by contributions of the three social classes $\omega^P \leq \omega^M \leq \omega^R$. For certain, the poor class will be better off given that they are contributing with the lowest amount, but it is unclear whether it will pay off for the middle class and the rich class.[6] Ultimately, this will depend on the wealth distribution and on the social return of the public good λ.

If, instead, the parameters of the game are such that $\omega^M \leq \hat{\omega} < \omega^R$ only partial contribution of poor and middle class will occur. In this case, the level of SW achieved would be:

$$SW = \lambda(q^P\omega^P + q^M\omega^M) + q^R(\omega^R - p_H f) - \frac{p_H c}{p_H - p_L} \tag{9}$$

or equivalently,

$$SW = \lambda W - [q^R(\lambda - 1)\omega^R + q^R p_H f + \frac{p_H c}{p_H - p_L}] \tag{10}$$

Therefore, at a first glance we can see that there is a welfare loss captured by the second term of the equation which entails, on the one hand, the net loss of not contributing plus the expected fines for all the rich class and, on the on the other hand, the institutional costs (salary). Notice that these losses could be so high such that not implementing the institution became the SW maximizing decision. Comparing SW with and without the institution we can derive the following condition for the institution implementation to be SW maximizing:

$$(\lambda - 1)(q^P\omega^P + q^M\omega^M) \geq q^R p_H f + \frac{p_H c}{p_H - p_L} \tag{11}$$

Intuitively, the sanctioning institution should be implemented, from a social point of view, if the net aggregate gains from the public good provision (LHS) were greater than the aggregate expected fines plus the sheriff's rent (RHS).

A similar reasoning and interpretation could be followed for the other case of partial contribution where $\omega^P \leq \hat{\omega} < \omega^M$. Now, SW would be:

$$SW = \lambda W - [q^M((\lambda - 1)\omega^M + p_H f) + q^R((\lambda - 1)\omega^R + p_H f) + \frac{p_H f}{p_H - p_L}] \tag{12}$$

Notice that the SW loss captured in equation 12 is greater than the one obtained in the previous case of partial contribution with the poor and middle class, from what we can conclude that SW is increasing with contributions.

Correspondingly, the condition for it to be SW maximizing to implement the institution when only the poor class contributes is as follows:

$$(\lambda - 1)q^P\omega^P \geq (q^M + q^R)p_H f + \frac{p_H c}{p_H - p_L} \tag{13}$$

For the sake of briefness and given these analogous results, to tackle the question on whether the different governments would implement the sanctioning institution or not, we will focus on the case of partial contribution of both poor and middle-class individuals ($\omega^P \leq \hat{\omega} < \omega^R$). To

[6] Notice that the average of a variable is always found between the minimum and the maximum value such variable can take: $\underline{x} \leq \sum x_i/n \leq \overline{x}$

do so, let's take the condition for the implementation to be SW maximizing (equation 13) and rewrite it in the following way:

$$\frac{\lambda}{n}[q^P\omega^P + q^M\omega^M] - \frac{p_H c}{n(p_H - p_L)} \geq \frac{q^P\omega^P + q^M\omega^M + q^R p_H f}{n} \tag{14}$$

Alternatively, a government will implement the sanctioning institution if and only if:

$$\frac{\lambda}{n}[q^P\omega^P + q^M\omega^M] - \frac{p_H c}{n(p_H - p_L)} \geq \begin{cases} \omega^P & \text{if } \omega^* = \omega^P \\ \omega^M & \text{if } \omega^* = \omega^M \\ p_H f & \text{if } \omega^* = \omega^R \end{cases}$$

Notice these two conditions are equivalent on the LHS and differ o the RHS. While governments consider the opportunity cost of the individual they are representing, from the point of view of a social planner it is the average of everybody's opportunity cost what is being taken into account. In other words, the social criterium is different to the one followed by the government. Only if, coincidentally, the PDA's opportunity cost were equal to the average of everybody's opportunity cost, the decision of this government will always be efficient. From the comparison of these two conditions we can assert the following results:

***Proposition 6**: If $\omega^P \leq \hat{\omega} < \omega^R$, a government representing a political decisive agent with the lowest opportunity cost will implement the institution in situations where it is inefficient to do so. If $\omega^P \leq \hat{\omega} < \omega^R$, a government representing a political decisive agent with the highest opportunity cost will decide not to implement the institution in situations where it is efficient to do so.*

5 The determinants of the emergence of a sanctioning institution

Once the equilibria have been characterized, let's understand the effects of changes in the different determinants of the emergence of a sanctioning institution. Recall the condition for the sanctioning institution to emerge assuming $\omega^P \leq \hat{\omega}$:

$$\frac{\lambda}{n}[\sum_{\omega^j \leq \hat{\omega}} q^j\omega^j] \geq \frac{p_H s_k}{n} + \begin{cases} \omega^* & \text{if } \omega^* \leq \hat{\omega} \\ p_H f & \text{if } \omega^* > \hat{\omega} \end{cases}$$

where $\hat{\omega} = \frac{p_H f}{1-\frac{\lambda}{n}}$, $s_k = 0$ *if* $\omega^R < \hat{\omega}$ *and* $s_k = \frac{c}{(p_H - p_L)}$ *if* $\hat{\omega} < \omega^R$.

Changes in the parameters could trigger out a change in this condition potentially through three different channels: either a variation in the individual return of the public good, the sheriff's per capita salary or the opportunity cost of the PDA. Additionally, notice that a change in the parameters that determine the contribution threshold, $\hat{\omega}$, could affect the contribution decision and, therefore, the size of the public good. For instance, if initially both poor and middle-class individuals contribute because $\omega^M \leq \hat{\omega} \leq \omega^R$ an increase in $\hat{\omega}$ will have no effect [7]. However, if initially $\omega^P \leq \hat{\omega} \leq \omega^M$ such that only poor people contribute, a change in $\hat{\omega}$ could make it become a society such that $\omega^P \leq \omega^M \leq \hat{\omega}$ and middle-class individuals also have incentives to contribute. Thus, only if initially $\hat{\omega} \leq \omega^M$ an increase in the critical threshold could switch the

[7] Recall we are always assuming $\omega^P \leq \hat{\omega} \leq \omega^R$.

relationship between these two variables, increasing the level of cooperation. We call this effect the *switch effect* and if it occurs it will increase the public good provision $\sum_{\omega^j \leq \hat{\omega}} q^j \omega^j$.

We now classify our variables into two different groups: institutional and technological variables on one side and wealth distribution variables on the other.

5.1 Institutional and technological variables

A sanctioning institution could be implemented more easily/difficultly with variations in the society's institutional and technological variables. For instance, changes in the value citizens assign to the public good captured by λ, improvements in the legal capacity of punishment represented by the fine f, increases in the effectiveness of high effort exertion in detecting fraud p_H, changes in the moral hazard likelihood ratio $\frac{p_L}{p_H - p_L}$ or increases in the cost of exerting high effort c are subject of study in this subsection.

Let's start by considering an increase in the social return of the public good, λ. There is a direct effect on the individual return of the public good for a given level of public provision, making the condition hold in a greater range of cases. Additionally, there will be an increase in $\hat{\omega} = \frac{p_H f}{1 - \lambda/n}$ which might cause a *switch effect*, that is a social class changes its behaviour from free riding to contribution. If this happens and there is an increase in the level of public good provision, the sanctioning institution will emerge in even a greater range of cases. But then, if this social class corresponds to the PDA, its opporunity cost will increase making more difficult the implementation of the sanctioning institution. This can only happen when the initial situation is that of a government representing a free-riding middle-class.

Result 1*: An increase in the social return of the public good would make the sanctioning institution emerge more easily under any government representing the poor, the rich or a contributing middle class. But the result is unclear for a government representing a free-riding middle-class.*

With an increase in the fine paid in case of fraud detection, f, two different effects could happen. On the one hand the *switch effect* could also occur if initially only the poorer segment of the population contributed and the fine were sufficiently big in order to make more people contribute. If so, the sanctioning institution would emerge in a greater range of cases. Furthermore, it would affect the opportunity cost of a free rider. If the government represented a free-riding social class, the increase in the fine would make the sanctioning institution emerge more difficultly.

Result 2*: An increase in the fine paid by free-riders would make the sanctioning institution emerge more easily under a government representing a contributing social class, only if the fine were sufficiently large. Under a government representing a free-riding social class, the effect on the emergence of a sanctioning institution would be ambiguous given the positive effect on the potentially higher public good provision and the increase in the opportunity cost of free riding.*

Let's now consider the comparative statics of parameters that change the sheriff's salary. Recall that the existence of moral hazard implies paying the sheriff a per capita economic rent of $\frac{p_L c}{n(p_H - p_L)}$ when high effort is incentivized. The size of this rent depends on two factors: the likelihood ratio, $\frac{p_L}{p_H - p_L}$ and the cost of high effort exertion, c. The likelihood ratio represents how

informative the result is of the action (effort) chosen. If the difference between the probability of detecting defection under high and low effort is large, then the result is fairly informative about the effort exerted. Conversely, for very similar probabilities, the verifiable results would yield little information about the sheriff's effort [8].

Assume now that the likelihood ratio decreases due to a fall in p_L, then informativeness increases and the economic rents that have to be paid to the sheriff to give him incentives to exert effort would fall. It is straightforward that the same happens if the sheriff's cost of exerting high effort, c, decreases. Said in a different way, a decrease in the sheriff's salary due to either an increase in the informativeness of the result because of a decrease in the probability of punishing free-riding behavior under low effort or a fall in the cost of exerting high effort would make the sanctioning institution emerge in a greater range of cases, due to a diminishment in the cost of having such institution.

***Result 3**: A decrease in the probability of punishing free-riding behavior under low effort and/or a decrease in the sheriff's cost of exerting high effort would make the sanctioning institution emerge more easily under any type of government.*

Finally, a variation in the probability of detecting fraud under high effort, p_H, may trigger up to three effects. On the one hand, it would also affect the sheriff's salary, enhancing the emergence of the sanctioning institution as it would make its implementation less costly for the citizens[9]. On the other hand, it would have a similar effect to the variation in the fine: it would increase the opportunity cost of a free rider and could possibly trigger a *switch effect* through the increase in $\hat{\omega}$. The combination of these three effects would determine whether a particular government would implement the institution more easily or not.

If, for instance, the government represented a contributing social class, the sheriff's salary would always fall and the *switch effect*, if it happened, would also have a positive effect on the emergence of the sanctioning institution. For a government representing a free-riding social class, to these effects we add the rise in the opportunity cost of having a sanctioning institution, which would oppose to the previous effects, making the net effect uncertain.

***Result 4**: An increase in the probability of detecting free-riding attitudes under high effort would make the sanctioning institution emerge more easily under a government representing a contributing social class. Under a government representing a free riding social class, the effect on the emergence of the sanctioning institution would be uncertain.*

5.2 Wealth distribution

Recall the wealth distribution is composed by the number of people belonging to each social class: q^P, q^M, q^R and their corresponding levels of wealth: $\omega^P, \omega^M, \omega^R$. In this section we aim to analyze the effect on the emergence of the sanctioning institution derived from changes of these variables.

[8] In the extreme case, if $p_H = 0.5 + \epsilon$ and $p_L = 0.5 - \epsilon$, the result provides no information about the effort exerted.

[9] An increase in p_H will decrease the expected wage in $\frac{np_L}{[n(p_H - p_L)]^2}$ units.

5.2.1 Natural growth of the population

First, we consider the impact of a proportional change in the population as a consequence of natural growth, i.e. $\Delta n = \Delta q_P = \Delta q_M = \Delta q_R$. In this case, there is a clear effect on the per capita salary paid to the sheriff, loosening the emergence condition. In other words, the effect would favor the institution emergence. However, notice that the increase in n might also make the amount of public good provision fall through a diminishment in $\hat{\omega} = \frac{p_H f}{1-\lambda/n}$, which could cause a *switch effect* if wealth was initially distributed such that $\omega^M \leq \hat{\omega} \leq \omega^R$. If this effect occurs and less people contribute, the return of the public good would also fall. Thus, with a *switch effect*, the net effect of a proportional increase in the population would be uncertain and will depend on the size these two opposing effects have. However, if the fall in the critical threshold $\hat{\omega}$ didn't affect the number of contributors (i.e. if no *switch effect* happened) the effect on the return of the public good would be neutral given that the number of people belonging to each social class grows proportionally. Consequently, the sanctioning institution would emerge more easily, due to the fall of the sheriff's per capita salary.

***Result 5**: Natural growth of the population would make the sanctioning institution emerge more easily under any type of government if it didn't hinder contribution. Otherwise, the net effect on the emergence would be uncertain because of this negative effect on contribution and the positive effect of a reduction in the sheriff's salary.*

5.2.2 Variation in the composition of social classes

Another possible scenario could be a transfer of individuals between social classes without the population size changing. For example, suppose certain middle-class individuals become poor, that is $\Delta q_P = \nabla q_M$. In other words, suppose there is a reduction in the size of the middle class. The effect this would have on the emergence of the sanctioning institution depends on who is contributing to the public good. If only the poor segment of the population were contributing, an increase in the number of people belonging to this social class would have a positive effect on the return of the public good and therefore on the easiness with which the institution is implemented. Nevertheless, if the public pot is composed by contributions of both poor and middle class, the effect on the return of the public good would be negative and the emergence condition would tighten. In this scenario, a transfer of individuals from the middle class to the poor class would decrease the size of the public pot at rate $\frac{\lambda}{n}\omega^M$ but would also increase it at rate $\frac{\lambda}{n}\omega^P$. Given that, by definition, $\omega^M > \omega^P$, the net effect on the return of the public good will be negative. Intuitively, if the middle class shrinks in size by one unit and the poor class expands in size by one player, the loss of the middle-class contributor won't be compensated by the additional contribution of the poor individual.

***Result 6**: If only poor-class individuals were contributing to the public good, a transfer of individuals from middle to poor class would make the sanctioning institution emerge more easily under any type of government. If, however, both poor and middle class were contributing, the institution would emerge with greater difficulty under any type of government.*

Consider now an external shock in the population affecting uniquely part of it, for instance, an immigration wave increasing the number of low-wealth individuals, $\Delta n = \Delta q^P$. In first place, the increase in the population would directly make citizens pay a lower salary per capita. However, as it happens with the natural growth phenomenon, to this effect we have to add the effect

on the return of the public good, which depends on who is contributing to it. If, for instance, only poor people contribute, the increase in poor people would enlarge the total amount of public good provision, $\sum q^P \omega^P$, but would leave the individual return $\frac{\lambda}{n} \sum q^P \omega^P$ unchanged, given that population grows in the same proportion as the number of poor people do.

However, if initially both poor and middle class individuals were contributing, besides the positive effect on the salary, there would be a negative effect on the return of the public good. The reason behind this is that even though the proportion $\frac{\lambda}{n} q^P \omega^P$ would remain unchanged, $\frac{\lambda}{n} q^M \omega^M$ would fall with the increase in the total population. Hence, in this case, the final effect would be uncertain and depend on these two opposing effects. Finally, notice that the increase in the total population could affect the number of people willing to contribute if initially $\omega^P \leq \hat{\omega} \leq \omega^M$ potentially causing a *switch effect*. This would emphasize the negative effect on the return of the public good previously exposed.

Result 7*: If only poor-class individuals were contributing to the public good, an immigration wave of poor-class individuals would make the sanctioning institution emerge more easily under any government. If, however, both poor and middle class were contributing, the final effect on the emergence of the sanctioning institution would be uncertain given the positive effect of the reduction in the sheriff's salary along with the negative effect of the reduction in the return of the public good.*

5.2.3 Enrichment and impoverishment of social classes

Finally, let's analyze the impact of a variation in the wealth level of some social class. In particular, we consider appealing to study an impoverishment of the middle class, $\nabla \omega^M$, given that many societies have suffered this misfortune after the scraps of the Great Recession. As formerly exposed, results will depend on who has incentives to contribute, so let's analyze each possible scenario: only poor-class individuals have incentives to contribute or both poor and middle-class individuals have. Notice that even though variations in ω^M do not affect the critical $\hat{\omega}$, the impoverishment of this social class could change the position of such threshold with respect to ω^P and ω^M.

If, for instance, only poor class individuals contributed and the fall in the middle class wealth produced no *switch effect*, there would be no effect on the emergence condition. However, the variation described could make a society where only poor class individuals contribute ($\omega^P \leq \hat{\omega} \leq \omega^M$) become one where both poor and middle class do so ($\omega^M \leq \hat{\omega} \leq \omega^R$) such that the *switch effect* turned on. This would increase the public good provision favoring the emergence of the institution. Moreover, there could also be an effect on the opportunity cost of the decisive agent if it belonged to the middle class, which would change from being $p_H f$ (before the *switch effect*) to ω^M (after it). If $\omega^M < p_H f$, the opportunity cost would fall, so the condition for emergence would loosen even further. However, if $p_H f < \omega^M$, the opportunity cost would rise and the net effect on the condition would depend on the balance between the increase in the return of the public good and the increase in the opportunity cost.

Result 8*: If initially only poor individuals contributed to the public good, an impoverishment of the middle class would only have an effect on the emergence of the sanctioning institution if such impoverishment enhanced contribution. Whether this effect is positive or negative would depend on the relationship between the middle-class wealth and the expected fine.*

Another possible scenario could be that individuals were initially distributed such that $\omega^M \leq \hat{\omega} \leq \omega^R$ and both poor and middle class contributed. The impoverishment of the middle class would make the return of the public good fall in $\lambda\frac{q^M}{n}$ units, but at the same time the opportunity cost of a middle-class PDA would also fall at rate 1. Therefore, if the condition $\frac{q^M}{n} \leq \frac{1}{\lambda}$ held, the effect of the impoverishment of this social group would ease the emergence of the sanctioning institution. Notice that this condition establishes that the proportion of individuals belonging to the middle class must be lower than than the social marginal rate of substitution between private and public goods. However, if the PDA were different to the middle class, the effect on the opportunity cost would not occur and the sanctioning institution would emerge for a lower range of cases given that only the fall in the return of the public good would happen.

***Result 9**: If both poor and middle class individuals were contributing to the public good, an impoverishment of the middle class would make the sanctioning institution emerge more difficultly under a poor and rich-class government. Under a middle-class government, however, the sanctioning institution would emerge more easily if $\frac{q^M}{n} \leq \frac{1}{\lambda}$, i.e. if the proportion of individuals belonging to the middle class were lower than than social marginal rate of substitution between private and public goods.*

For a matter of completeness, let's comment what would happen with the impoverishment of the other social classes. If the rich class suffered a diminishment in their level of wealth, $\nabla\omega^R$, this would have no effect on the return of the public good (given that we are assuming they always have incentives to free ride) and would have no effect on the opportunity cost either, as his free-riding condition makes his opportunity cost always be $p_H f$. Thus, a variation in the wealth of the wealthiest social class would have no effect on the emergence of the sanctioning institution.

An impoverishment of the poor class, however, would have a similar impact to the one formerly described for the middle class. The return of the public good would fall with a $\nabla\omega^P$ making the emergence condition hold in a lower range of cases. However, the opportunity cost of a poor-class PDA would fall making the condition hold in a greater range of cases. Nonetheless, if the PDA were not poor, there would be no variation on the opportunity cost. Hence, an impoverishment of the poor class would make the sanctioning institution emerge more difficultly under a middle and rich-class government, while under a poor-class government the sanctioning institution would emerge more easily if $\frac{q^P}{n} \leq \frac{1}{\lambda}$.

6 Heterogeneous valuation of the public good

Even though the model has been proposed with heterogeneity in terms of wealth, symmetric results arise when individuals show different marginal returns of the public good. Recall λ represents the personal valuation each individual gives to his corresponding share of the public good. Given that the number of players is fixed, let's assume individuals have a valuation of the public good which can be classified into one of three groups. Namely, $\lambda^j \in \{\lambda^L, \lambda^M, \lambda^H\}$, such that individuals now have either a low, a middle or a high valuation of the public good. This way we define the valuation distribution as follows $\{(\lambda^L, \lambda^M, \lambda^H), (q^L, q^M, q^H)\}$

Summing up and following the same reasoning as before, we obtain symmetrical results to the ones expressed in previous sections: individuals with higher valuations of the good will now

have higher incentives to contribute, while those with lower valuations will have lower gains of doing so. Thus, we obtain a critical threshold $\hat{\lambda} = n - \frac{np_H f}{\omega}$ in terms of valuation (instead of wealth) from which individuals switch from free riding to contributing.

Finally, the condition for the emergence of the sanctioning institution is analogous, considering the fact that the PDA is now going to be determined in terms of valuation on the public good. Our main result for this extension is presented in the following proposition. Formal proof has been relegated to the appendix.

***Proposition 6**: Assume that $\lambda^L \leq \hat{\lambda} \leq \lambda^H$ and the political decisive agent has valuation λ^*. The sanctioning institution will emerge if and only if:*

$$\frac{\lambda^j}{n}[\sum_{\lambda^j \geq \hat{\lambda}} q^j \omega] \geq \frac{p_H s_k}{n} + \begin{cases} \omega & \text{if } \lambda^* \geq \hat{\lambda} \\ p_H f & \text{if } \lambda^* < \hat{\lambda} \end{cases}$$

where $\hat{\lambda} = n - \frac{np_H f}{\omega}$, $s_k = 0$ if $\hat{\lambda} \leq \lambda^j$ and $s_k = \frac{c}{(p_H - p_L)}$ if $\lambda^j < \hat{\lambda}$.

Individuals with a relatively low wealth usually assign a higher valuation to public goods and viceversa. This is commonly attributed to the fact that wealthier individuals can afford private substitutes to a larger extent, for instance medical insurance plans. Hence, if we jointly considered wealth and public good valuation heterogeneity, results would be reinforced. An individual with a low wealth that values public goods highly, will have high incentives to contribute to the provision to a public good, and viceversa.

7 Conclusions

This paper theoretically analyzes how do centralized sanctioning institutions emerge in selfish societies that must decide on the provision of a public good only once. The implementation depends on an independent player, in this case a government who chases being reelected in the future. Examples such as international agreements or bargaining among companies are of this nature. Moreover, we study the level of contribution such institution can potentially achieve if implemented.

The emergence of a sanctioning institution depends on a set of institutional and technological parameters (valuation society assigns to the provision of a pubic good, cost and efficiency in fraud detection and the fine paid in case of free-riding attitudes), the wealth distribution in the society (levels of wealth of each social class and proportion of each type among the population) and the identity of the political decisive agent, that is, the agent that ensures the government being reelected. Given these exogenous variables, the sanctioning institution will emerge more easily if the persistence of the government depends on the social class with the lowest opportunity cost in the provision of a public good, independently of it being a contributing or a free-riding social class.

Without any enforcing mechanism, selfish individuals will fully free ride on the one-shot public good. Nonetheless, the sanctioning institution can achieve a positive provision of such good if the society counts with a relatively high quality in their sanctioning institutions and high social return of the public good.

A Appendix

Proof of Proposition 2:

Case by case:

Case 1. $\omega^R \leq \hat{\omega} \leq \tilde{\omega}$

For this first case, all individuals satisfy $\omega^j \leq \hat{\omega}$. Thus, according to proposition 1, contributing is a dominant action for every player, so the unique Nash equilibrium is $g_i^j = \omega_i^j \ \forall_{i,j}$.

Case 2. $\omega^P \leq \hat{\omega} \leq \omega^R \leq \tilde{\omega}$

The wealth distribution in the society is such that for some classes $\omega^j \leq \hat{\omega}$ while for others $\hat{\omega} < \omega^j \leq \tilde{\omega}$. This gives rise to two Nash equilibria in the contribution subgame. On the one hand everybody could contribute with $g_i^j = \omega_i^j$. However, if somebody free rides, individuals with wealth $\hat{\omega} < \omega^j \leq \tilde{\omega}$ will fully free ride. Thus, the two Nash equilibria are as follows: either $g_i^j = \omega_i^j \ \forall_{i,j}$ or $g_i^j = \omega_i^j$ for those with $\omega^j \leq \hat{\omega}$ and $g_i^j = 0$ for individuals with $\hat{\omega} < \omega^j$.

Case 3. $\omega^P \leq \hat{\omega} \leq \tilde{\omega} \leq \omega^R$

This is the situation where the wealth distribution in the society is sparser in the wealth spectrum. Recalling proposition 1, there would be a set of players, those with $\omega^j \leq \hat{\omega}$, which will contribute with $g_i^j = \omega_i^j$. Given that individuals in the segment $\hat{\omega} < \omega^j \leq \tilde{\omega}$ are conditional contributors, and there is a proportion of players ($\tilde{\omega} < \omega^j$) which would free ride no matter what, these conditional contributors would also free ride. Thus, there would be a unique Nash equilibrium where $g_i^j = \omega_i^j$ for those with $\omega_i^j \leq \hat{\omega}$ and $g_i^j = 0$ for individuals with $\hat{\omega} < \omega^j$.

Case 4. $\hat{\omega} \leq \omega^P \leq \omega^R \leq \tilde{\omega}$

Following best responses in proposition 1, this setup gives rise to two Nash Equilibria in the contribution subgame: either $g_i^j = \omega_i^j \ \forall_{i,j}$ or $g_i^j = 0 \ \forall_{i,j}$.

Case 5. $\hat{\omega} \leq \omega^P \leq \tilde{\omega} \leq \omega^R$

This wealth distribution leads to a unique Nash equilibrium where $g_i^j = 0 \ \forall_{i,j}$. This result is obtained by successive elimination of dominated actions. Given that individuals with $\omega^j \leq \tilde{\omega}$ will free ride if at least one other player free rides and individuals with $\tilde{\omega} < \omega^j$ will always free ride, everybody would do so.

Case 6. $\hat{\omega} \leq \tilde{\omega} \leq \omega^P \leq \omega^R$

For every single individual, ω^j is too high for them to have incentives to contribute. The unique Nash equilibrium is to free ride with $g_i^j = 0 \ \forall_{i,j}$.

■

Proof of Lemma 3:

The characterization of s_0 (the salary paid in case the sheriff's intervention is finally not necessary because everybody contributes) is straightforward. If this occurs, the government needn't offer anything above the sheriff's reservation utility, $\overline{u}$. Recall $\overline{u} = 0$, so $s_0 = 0$.

The optimal salary $s_0 = 0$ is independent on the effort the government desires the sheriff to exert. This does not hold for s_p and s_{np}, for which we must consider the maximization

problem subject to participation and incentive constraints, as well as limited liability constraints.

Let's first characterize the contract where the government, maximizing the utility of the political decisive agent, enhances the exertion of e_H. In this case, the maximization problem would be:

$$\begin{aligned} \underset{s_p, s_{np}}{\text{maximize}} \quad & \omega^* - g^* + \frac{\lambda}{n}[\sum_{i=1}^{n} g_i] - \gamma_H \\ \text{subject to} \quad & p_H s_p + (1-p_H)s_{np} - c \geq 0 \\ & (p_H - p_L)(s_p - s_{np}) \geq c \\ & s_p, s_{np} \geq 0 \end{aligned}$$

where:

$$\gamma_H = \begin{cases} p_H \frac{s_p}{n} + (1-p_H)\frac{s_{np}}{n} & \text{if } \sum_{i=1}^{n} g_i < \sum_{i=1}^{n} \omega_i \text{ and } g^* = \omega^* \\ p_H \frac{s_p}{n} + (1-p_H)\frac{s_{np}}{n} + p_H f & \text{if } \sum_{i=1}^{n} g_i < \sum_{i=1}^{n} \omega_i \text{ and } g^* < \omega^* \\ \frac{s_0}{n} & \text{if } \sum_{i=1}^{n} g_i = \sum_{i=1}^{n} \omega_i \end{cases}$$

The government maximizes the utility of the political decisive agent subject to the sheriff's participation and incentive constraint. The former one ensures the sheriff will accept the offered contract instead of staying out of the game, while the latter one ensures that he's better off by exerting the desired level of effort. Furthermore, the sheriff has limited liability, so salaries must all be positive. The function γ_H represents the individual cost of having the sheriff. This function can take three forms depending on individual and total contributions.

Assuming the government has all the bargaining power when negotiating with the sheriff, the contract enhancing high effort would be: $\{s_0 = 0, s_p = \frac{c}{p_H - p_L}, s_{np} = 0\}$. Under this contract, economic rents are: $\frac{p_L}{p_H - p_L} c$.

If instead of enhancing high effort, he prefers to encourage low effort, the maximization problem would be:

$$\begin{aligned} \underset{s_p, s_{np}}{\text{maximize}} \quad & \omega^* - g^* + \frac{\lambda}{n}[\sum_{i=1}^{n} g_i] - \gamma_L \\ \text{subject to} \quad & p_L s_p + (1-p_L)s_{np} \geq 0 \\ & (p_L - p_H)(s_p - s_{np}) \leq c \\ & s_p, s_{np} \geq 0 \end{aligned}$$

where:

$$\gamma_L = \begin{cases} p_L \frac{s_p}{n} + (1-p_L)\frac{s_{np}}{n} & \text{if } \sum_{i=1}^{n} g_i < \sum_{i=1}^{n} \omega_i \text{ and } g^* = \omega^* \\ p_L \frac{s_p}{n} + (1-p_L)\frac{s_{np}}{n} + p_L f & \text{if } \sum_{i=1}^{n} g_i < \sum_{i=1}^{n} \omega_i \text{ and } g^* < \omega^* \\ \frac{s_0}{n} & \text{if } \sum_{i=1}^{n} g_i = \sum_{i=1}^{n} \omega_i \end{cases}$$

In case the government wants to enhance low effort, it would be enough for him to offer an acceptable contract yielding no economic rents: $\{s_0 = 0, s_p = 0, s_{np} = 0\}$.

■

Proof of Proposition 6:

Amount of the public good provision

Analogously to the setup with heterogeneous levels of wealth, we assume that there are three valuation groups in this game. Namely, a low-valuation group (j=L), a middle valuation group (j=M) and a high valuation group (j=H). We define the valuation distribution using a pair of vectors $\{(\lambda^L, \lambda^M, \lambda^H), (q^L, q^M, q^R)\}$ where $q^L + q^M + q^H = n$.

Firstly, let's assume that under the low effort contract, everybody would free ride and focus on those cases where the sheriff has been offered a high-effort contract. Notice lemma 1 still holds: if an individual is going to free ride, he will maximize his utility by free riding with $g_i = 0$.

In a situation where everybody else contributes, the size of the fund before everybody else contributes is $\sum_{i=1}^{n} g_{-i} = (n-1)\omega$. Player i will contribute if the expected costs of free riding were greater than the net gains:

$$\frac{p_H s_p + (1-p_H)s_{np} - s_0}{n} + p_H f \geq \omega(1 - \frac{\lambda_i^j}{n})$$

Let's denote by $\tilde{\lambda}$ the critical value of λ_i^j such that this holds with equality. In this case, if individual i from valuation group j has a personal valuation such that $\lambda_i^j \geq \tilde{\lambda}$, he will contribute as long as everybody else contributes as well.

In a situation where there is at least one free rider, individual i from valuation group j will contribute as long as:

$$p_H f \geq \omega(1 - \frac{\lambda_i^j}{n})$$

Similarly, let's name $\hat{\lambda}$ the critical valuation such that this holds will equality. If individual i from valuation group j has a valuation such that $\lambda_i^j \geq \hat{\lambda}$, he will always contribute. Notice that now $\tilde{\lambda} \leq \hat{\lambda}$.

Notice as well that the intuition is symmetrical to the wealth heterogeneity one. If individuals value the public good sufficiently high, they will contribute because gains of free riding fall as the valuation of the public good increases.

Regarding the whole population, case by case:

Case 1. $\lambda^H \leq \tilde{\lambda} \leq \hat{\lambda}$
Everybody values the good sufficiently little, thus, everybody will free ride no matter what, such that the Nash equilibrium is $g_i^j = 0 \ \forall_{i,j}$.

Case 2. $\lambda^L \leq \tilde{\lambda} \leq \lambda^H \leq \hat{\lambda}$
If some individuals free ride no matter what (those with $\lambda^j \leq \tilde{\lambda}$) while others are conditional contributors (those with $\lambda^j > \tilde{\lambda}$), by successive elimination of dominated actions, everybody will free ride with $g_i^j = 0 \ \forall_{i,j}$.

Case 3. $\lambda^L \leq \tilde{\lambda} \leq \hat{\lambda} \leq \lambda^H$.
In this case, those with the lower valuation ($\lambda^j \leq \tilde{\lambda}$) plus the conditional contributors

($\tilde{\lambda} \leq \lambda^j \leq \hat{\lambda}$) will free ride, while those with the higher valuation ($\hat{\lambda} \leq \lambda^j$) will contribute. Thus, the unique Nash equilibrium is $g_i^j = 0$ for all players with $\lambda^j \leq \hat{\lambda}$ and $g_i^j = \omega_i^j$ for those who $\hat{\lambda} < \lambda^j$.

Case 4. $\tilde{\lambda} \leq \lambda^L \leq \lambda^H \leq \hat{\lambda}$.

This scenario gives rise to either everybody contributing or everybody free riding. All players are conditional contributors in this case and they will contribute as long as everybody does so. As soon as one of them deviates to free riding, everybody will free ride. Let's apply the equilibria selection criterion explained in the heterogeneous wealth model, considering the worst possible case where at least one individual deviates to free riding. Following the presented intuition, this would lead to $g_i^j = 0 \ \forall_{i,j}$.

Case 5. $\tilde{\lambda} \leq \lambda^L \leq \hat{\lambda} \leq \lambda^H$

If some individuals are contributors, while the rest only contribute conditionally, two Nash equilibria may occur: either everybody contributes, or only unconditional contributors contribute and conditional ones free ride. Following the equilibria selection criteria explained previously, let's select that one where at least one individual free rides so that the outcome is $g_i^j = \omega_i^j$ for players with $\hat{\lambda} < \lambda^j$ and $g_i^j = 0$ for the rest of the players with $\lambda^j \leq \hat{\lambda}$.

Case 6. $\tilde{\lambda} \leq \hat{\lambda} \leq \lambda^L \leq \lambda^H$

Finally, if everybody is found in the top segment of the valuations spectrum, everybody will contribute with $g_i^j = \omega_i^j \ \forall_{i,j}$.

Notice that, as before, $\hat{\lambda}$ becomes the unique critical value for the characterization of the Nash equilibria.

Summing up, given an initial valuation distribution $\{(\lambda^L, \lambda^M, \lambda^H), (q^L, q^M, q^H)\}$ and assuming the sanctioning institution has been implemented and exerts high effort, the selected equilibria at the contribution stage are:

- If $\lambda^H \leq \hat{\lambda}$, then $g^j = 0 \ \forall_j$
- If $\lambda^L \leq \hat{\lambda} \leq \lambda^H$, then $g^j = \omega^j$ for individuals with $\lambda^j \leq \hat{\lambda}$ and $g^j = 0$ for individuals with $\hat{\lambda} < \lambda^j$.
- If $\hat{\lambda} \leq \lambda^L$, then $g^j = \omega^j \ \forall_j$.

Consequently, the amount of public good provision will be:

$$\sum_{\lambda^j \geq \hat{\lambda}} q^j \omega$$

where $\hat{\lambda} = n - \frac{np_H f}{\omega}$.

Emergence of the sanctioning institution

In order for the sheriff to have incentives to really exert high effort, the incentives constraint must hold:

$$(p_H - p_L)(s_p - s_{np}) \geq c$$

Moreover, provided that the maximization problem is the same as before with the particularity of λ^j instead of ω^j, the minimum-cost contracts are the same as the ones described in the heterogeneous wealth model:

- High-effort contract:$\{s^H\} = \{s_0 = 0, s_p = \frac{c}{p_H - p_L}, s_{np} = 0\}$
- Low-effort contract: $\{s^L\} = \{s_0 = 0, s_p = 0, s_{np} = 0\}$

Recall the government considers the political decisive agent in order to decide which contract to offer to the sheriff. The government can also offer an unacceptable contract with any $s_k < 0$ if he anticipates the political decisive agent is better without the sheriff's intervention. As before, this is as least as good as offering the low effort contract so the government will choose between implementing the sanctioning institution or not given that:

$$\omega \geq \omega - np_L f$$

As before, let's assume we're in a situation such that $\hat{\lambda} \leq \lambda^R$ and the political decisive agent has a valuation λ^*. Our main result for this extension states that the sanctioning institution will emerge if and only if:

$$\frac{\lambda^j}{n}[\sum_{\lambda^j \geq \hat{\lambda}} q^j \omega] \geq \frac{p_H s_k}{n} + \begin{cases} \omega & \text{if } \lambda^* \geq \hat{\lambda} \\ p_H f & \text{if } \lambda^* < \hat{\lambda} \end{cases}$$

where $\hat{\lambda} = n - \frac{np_H f}{\omega}$, $s_k = 0$ if $\hat{\lambda} \leq \lambda^j$ and $s_k = \frac{c}{(p_H - p_L)}$ if $\lambda^j < \hat{\lambda}$.

■

References

[1] Acemoglu, D., & Wolitzky, A. (2015). "Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement." *National Bureau of Economic Research,* (No. w21457).

[2] Black, D. (1948). "On the rationale of group decision-making". *The Journal of Political Economy*, 23-34.

[3] Buckley, E., & Croson, R. (2006). "Income and Wealth Heterogeneity in the Voluntary Provision of Linear Public Goods." *Journal of Public Economics*, 90(4-5), 935-955.

[4] Burlando, R. M., & Guala, F. (2005). "Heterogeneous Agents in Public Goods Experiments." *Experimental Economics*, 8(1), 35-54.

[5] Cherry, T. L., Kroll, S., & Shogren, J. F. (2005). "The Impact of Endowment Heterogeneity and Origin on Public Good Contributions: Evidence From the Lab." *Journal of Economic Behavior & Organization*, 57(3), 357-365.

[6] Fehr, E., & Williams, T. (2013). "Endogenous Emergence of Institutions to Sustain." *Working Paper.*

[7] Fellner, G., Iida, Y., Kröger, S., & Seki, E. (2011). "Heterogeneous Productivity in Voluntary Public Good Provision - An Experimental Analysis." *Working Paper.*

[8] Fisher, J., Isaac, R. M., Schatzberg, J. W., & Walker, J. M. (1995). "Heterogenous Demand for Public Goods: Behavior in the Voluntary Contributions Mechanism." *Public Choice*, 85(3-4), 249-266.

[9] Greif, A. (2006). "Institutions: Theory and History." *Cambridge University Press*, 217-268.

[10] Greif, A. (1993). "Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition". *The American Economic Review*, 83(3), 525-548.

[11] Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). "The competitive advantage of sanctioning institutions". *Science,* 312(5770), 108-111.

[12] Hotelling, H. (1929). " Stability in Competition". *Economic Journal* 29: 41-57.

[13] Kosfeld, M., Okada, A., & Riedl, A. (2009). "Institution Formation in Public Goods Games." *American Economic Review*, 1335-1355.

[14] Milinski, M., Traulsen, A., & Röhl, T. (2012). "An Economic Experiment Reveals that Humans Prefer Pool Punishment to Maintain the Commons Groups." *Proceedings of the Royal Society of London B: Biological Sciences*, 279, 3716-3721.

[15] Okada, A. (1993). "The Possibility of Cooperation in an N-Person Prisoners' Dilemma with Institutional Arrangements." *Public Choice*, 77(3), 629-656.

[16] Przeworski, A. (2015). "Economic Inequality, Political Inequality, and Redistribution." Draft. *Department of Politics. New York University.*

[17] Rapoport, A. (1988). "Provision of step-level public goods: Effects of inequality in resources." *Journal of Personality and Social Psychology*, 54(3), 432.

[18] Reuben, E., & Riedl, A. (2013). "Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations." *Games and Economic Behavior*, 77(1), 122-137.

[19] Sigmund, K., Hauert, C., & Traulsen, A. (2011). "Social Control and the Social Contract: The Emergence of Sanctioning Systems for Collective Action." *Dynamic Games and Applications*, 1(1), 149-171.

[20] Tommasi, M., & Weinschelbaum, F. (2014). "Centralization vs. Decentralization: A Principal-Agent Analysis." *Journal of Public Economic Theory*, 9(2), 369-389.