The following document is an accepted version of the following published paper:

Alventosa, A., & Olcina, G. (2021). Rich and Poor: On the Emergence of a Sanctioning Institution. *Journal of Institutional and Theoretical Economics*, *177*(1), 56-80.

URL: https://www.mohrsiebeck.com/en/article/rich-and-poor-on-the-emergence-of-a-sanctioning-institution-101628jite-2020-0050

DOI: https://doi.org/10.1628/jite-2020-0050

Rich and poor. On the emergence of a sanctioning institution

by

Adriana Alventosa, Gonzalo Olcina*

September 17, 2020

This paper analyzes the implementation of centralized sanctioning on free-riding behavior in societies where wealth heterogeneity leads to different individual contributing and free-riding incentives on public goods. The decision on the implementation is made by a government representing a social class. We show that if the sanctioning institution can achieve contribution by the entire population, the government representing the poor will implement it more often. This will always be efficient. However, if the sanctioning institution only achieves contribution of a share of the population, the government of a free-riding rich class can implement it more often if expected fines are sufficiently low. Nevertheless, implementing the institution may be inefficient.

Keywords: public goods game, sanctioning institutions, wealth heterogeneity, pool punishment, moral hazard.

JEL classification code: C72, D02, H41

1 *Introduction*

Our behavior in society is strongly influenced by enforcement institutions that determine what should and what should not be done. Law enforcement is provided through the courts, the police or national tax agencies, for instance. They impose punishment of improper behavior through sanctions. Many of us believe that we would behave in the same way if such institutions did not exist because

*Adriana Alventosa: UMICCS, ERI-CES, University of Valencia. Gonzalo Olcina (corresponding author): 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), the Workshop in Industrial Organization (Valencia, 2017) and the Southern Europe Experimentalists Team SEET 2018 (Lecce, 2018). The authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness, the European Social Fund and the projects EC02014-58297-R, ECO2017-87245-R and BES-2015-073089.

it is the “right thing" to do, but our selfishness would surely move us in another direction.

This paper studies the implementation and performance of a sanctioning institution in a public goods provision context. Contributions take place in a society where there are different social classes with different levels of wealth. Decisions on the provision of incentives for the enforcement institution are made by a government representing the interests of a particular social class. We focus on and compare the decisions made by different governments.

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 free-rider issue (Fehr and Gachter, 2000). Peer punishment consists of the opportunity for each individual to costly penalize, at the end of the game, those participants who have been free riders. However, peer punishment causes high collateral damage as the costs surpass the possible gains in cooperation, unless individuals interact over prolonged periods of time (Gächter et al. 2008). Additionally, formal individual punishment is hard to implement, especially in large groups (Greif, 1993). Moreover, we know from evolutionary game theory (Boyd and Richerson, 1988), that it is difficult to explain the evolution of cooperation when there is peer punishment. This kind of punishment is very costly and, rare unconditional punishers bear substantial costs and hence are eliminated in the evolutionary dynamics.

An alternative to peer punishment is pool punishment, that is, centralized sanctioning. Throughout history, individuals have tended to group themselves and develop centralized sanctioning institutions with the power of punishing defectors. These institutions have not been automatic punishers, but have had their own incentives in the maintenance of social and political order. Many historical examples abide. Just as an example, Genoa in the period 1194-1339, where 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).

Additionally, the statemaking theory has exposed the emergence of centralized enforcers as providers of public goods. In primary societies where roving banditry reigned, leading to inefficient overpunishment through excessive plundering, stationary banditry eventually dominated. Stationary bandits served as centralized sanctioning institutions, which monopolized violence in a given region and provided public goods in a more efficient way than roving bandits. This result has been proven to apply to Vikings’ activities and settlements (Kurrild-

Klitgaard and Svendsen, 2007).

Nowadays, in our complex societies, the provision of public goods - and the Welfare State - are mostly financed by taxes, and transgressions, specially in large groups, are punished by specialized law enforcers.[1] Illustrations of centralized enforcing institutions are specialized law bodies, such as the police or the courts (Fehr and Williams, 2017) or, at a larger international scale, the Kyoto-protocol or the United Nations Security Council (Sutter et al. 2010). These sanctioning institutions are steered by individuals with their own goals and interests. Specialized enforcers need to be given incentives to carry out costly punishment. Their effort exerted in monitoring and sanctioning non-cooperative behavior is subject to moral hazard problems[2] and, therefore, their performance will crucially depend on the incentives provided by the society represented by the government. But in a society with a heterogeneous distribution of wealth, the government might represent the interests of a particular social class (Przeworski, 2015). If everybody in the society or group has the same level of wealth, incentives are the same for everybody, but in the presence of wealth heterogeneity, the incentives provided by different social classes to the institution might be quite different.

Our main interest in this paper is to understand how enforcement institutions that govern us work if the decision of implementing such institutions had to be made today in societies where governments' decisions are strongly influenced by social classes' interests. With this purpose, we theoretically model a society conformed by selfish individuals[3] with different levels of wealth, i.e. different individual endowments from which to contribute to a public good. A government representing one of these classes has the opportunity of implementing a sanctioning institution before contributions are made, which will take action at the end of the game.

We first determine the relation between wealth and incentives to contribute. Our basic assumption is that the expected cost from being discovered and punished when there is no contribution is non- increasing on wealth. In this setting, on an individual basis, the higher an individual's wealth is, the lower will his/her incentives to contribute under the threat of a sanctioning institution be. We characterize a critical wealth threshold from which individuals stop finding it profitable to contribute. Such threshold depends on the valuation of the public good and on the quality of the sanctioning institution, that is, on the monitoring and punishment technology. This result is endorsed by experimental evidence

[1] Such centralized authorities are desirable when contracts and property rights are not enforceable due to, for instance, high transaction costs (Aldashev and Zanarone, 2017) and are better positioned to overcome coordination failures than peer punishment (Baldassarri and Grossman, 2011).

[2] Effort exerted could also be subject to adverse selection problems. However, in this work we focus on the moral hazard issue.

[3] 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.

(Buckley and Croson, 2006; Heap et al. 2016) and field evidence, which suggests a strong positive correlation between wealth and fraud. Tax evasion, for instance, occurs more frequently in wealthier social classes than in the lower social strata (Alstadsæter et al. 2019).

In the next place, we study under which government will the sanctioning institution be implemented more often and its efficiency. In this case, the identity of the social class represented by the government becomes decisive, as the one with the lowest cost will implement the institution in a greater range of cases. If the sanctioning institution is able to enhance the contribution of all social classes in the population, implementing it will always be efficient. The government implementing it in a greater range of cases will be the poor-class government and it will, therefore, always make the efficient decision. In general, if there is full contribution, a government representing a social class with a level of wealth smaller or equal than the average wealth of the population will always make the efficient decision of implementing the sanctioning institution. Nevertheless, a government representing a rich class above the average level of wealth will not implement it in situations where it is socially efficient to do so.

Alternatively, it could be the case that the sanctioning institution's threat only achieved the contribution of some social classes. In this case, a government who will implement it in a greater range of cases will only be a poor-class government if the expected fines faced by free riders were high enough. If, instead, they are sufficiently low, it could be a government representing the free-riding rich class the one implementing it more often. This decision, however, could lead to an inefficient implementation. This paradoxical result occurs because free riders benefit from the public good to the same extent as contributors. If the costs faced by free riders under a sanctioning institution are exceedingly low, they will prefer to have an institution that sanctions them and benefit from others' contributions rather than not having it and not enjoying the public good at all.

1.1 Related literature and contribution

This paper is mainly related to previous literature dealing with pool punishment and sanctioning institutions.

The concept of pool punishment was presented by Yamagishi (1986) as a mechanism that captured how investments in monitoring and sanctioning institutions to uphold the common interest are made. Sigmund et al. (2010) modeled the comparison between peer and pool punishment in a public goods game with and without counter-punishment (the punishment of those who cooperate, but do not punish), and with voluntary and mandatory participation (Sigmund et al., 2011). Milinski et al. (2012) reproduce this model experimentally, using their same assumptions. Their main result, also supported by Zhang et al. (2014), is that when pool punishment is combined with counter-punishment, contributions increase. Furthermore, pool punishment clearly prevails over peer punishment in terms of preference and efficiency (Milinski et al., 2012; Andreoni and Gee, 2012).

Regarding sanctioning institutions literature, Kosfeld et al. (2009) portray the endogenous 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. Other works have focused on exploring institutional incentive provision. Wu et al. (2014) and Dong et al. (2019) do so by comparing institutions that punish, institutions that reward and institutions that both punish and reward. Moreover, Sasaki et al. (2012) and Wu et al. (2014) analyze the evolutionary stability of cooperation under different incentives.

Moreover, previous studies have also proved the superiority of centralized institutions with respect to decentralized ones in terms of efficiency (Tommasi and Weinschelbaum, 2014; Fehr and Williams, 2017; Romaniuc et al., 2016). Acemoglu and Wolitzky (2020) approach the problem by endogeneizing specialized enforcement remarking the importance of incentives in fraud chasing. They compare decentralized community enforcement with specialized enforcement in a repeated game scenario.

All but this last paper, present automatic institutions, ignoring the possibility of strategic institutions with asymmetric information. Our work contributes to these two strands of the economic literature by incorporating these characteristics into our model. We consider a centralized sanctioning institution, an external enforcer, implementing pool punishment. Such enforcer can exert different levels of non-verifiable effort in detecting free riders. The government must provide the correct incentives such that the external enforcer implements the desirable level of effort.

Furthermore, the previously cited work and, in general, the standard public goods game literature, have mainly considered homogeneity in the contributors' wealth. This approach is clearly non-realistic and, more importantly, homogeneous endowments can provide misleading results as they hide the different incentives individuals with different levels of wealth have when it comes to contributing.

Experimental evidence (See Cherry et al., 2005; Burlando and Guala, 2005; Heap et al., 2016) has already shown that wealth heterogeneity affects the provision of a public good. In line with our results, Buckley and Croson (2006) demonstrate that less wealthy individuals contribute with a higher percentage of their endowment than wealthier ones. Only recently, Hauser et al. (2019) have proposed a framework for public goods game with wealth and productivity heterogeneity pinpointing that extreme inequality or misaligned inequality breaks down cooperation.

These examples constitute empirical evidence of the impact of wealth heterogeneity in public good provision. In our work, we provide a theoretical framework explaining this phenomenon by introducing different social classes. Moreover, this paper is, to the best of our knowledge, the first one showing the influence of the identity of the social class represented by the government in the implementation of a centralized sanctioning institution.

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. Finally, the last section will sum up the main results obtained.

2 The model

Consider the following n-player public goods game. 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 $\omega_i^j \in [\omega^P, \omega^R]$ (where $i = 1, ..., n$ identifies the individual and $j = 1, ..., z$ the social class) from which he/she can contribute $g_i^j \leq \omega_i^j$ to a public good. We denote with superscript $j = P$ the poorest social class in the population and $j = R$ the richest one. Each social class is characterized by being composed of 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. Hence, a wealth distribution is characterized by a pair of vectors $\vec{\omega} = (\omega^P, ..., \omega^R)$ and $\vec{q} = (q^P, ..., q^R)$ where $\sum_{q^j} q^j = n$.

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

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

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.[4] 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/her 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 Nash Equilibrium outcome is inefficient. Punishment has formerly been introduced as a mechanism with the purpose of attaining this socially desirable cooperation. However, given the characteristics of this game (one-shot and with selfish preferences), if at the end players were given the opportunity to peer punish each other at a cost, nobody would do so. Hence, another type of sanctioning must be implemented in order to make punishment an effective mechanism for attaining cooperation

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

among selfish individuals.

To this purpose, we introduce the possibility of hiring an external enforcer, which will be in charge of monitoring and implementing the punishment under a pre-designed contract. Hereinafter, we will refer to this agent as a sheriff or as a sanctioning institution indifferently.

If the sheriff is hired, he/she will have the opportunity of exerting two possible levels of costly effort to detect free riders in the final stage: a low level of effort (e_L) or a high level of effort (e_H). For the sake of simplicity, exerting a low effort will have no cost, $c_L = 0$, and $c_H = c$ where $c > 0$. If the sheriff exerts low effort, free riding 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 principal-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 and that the sheriff's salary costs are equally divided among the players, being the sheriff a public good itself (Zhang et al., 2014).[5] 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 $k \in \{0, p, np\}$ represents the outcomes. As it is standard in the principal-agent literature and for simplicity, we assume that the sheriff's reservation utility is zero, $\overline{u} = 0$, and that the sheriff has limited liability.[6]

The precise sequence of actions for the whole game 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 him/her. We assume that the government represents a particular social class and denote ω^* as the wealth of such class.

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 ($g_i^j = \omega_i^j$).[7]

---

[5] Unequal sharing of the institution cost, the sheriff's salary, would be another possibility, but that would not change the results substantially.

[6] The sheriff's final utility cannot be negative.

[7] Another option would be to work with a proportional tax rate t. To contribute would be to pay the tax $g^j = t\omega_i^j$ and to free ride would be to defraud $g^j = 0$. A government representing a particular social class would decide both on the implementation of the sanctioning institution and on the tax rate. Results on the implementation of the sanctioning institution would not

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.

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. The sheriff chooses the effort he/she will make in order to produce objective evidence on free-riding behavior, with cost c_e. If free riding is detected it will automatically be punished with a fixed exogenous fine, $f > 0$. This fine is a dissipative cost, that is, it is an amount of the citizens' income that gets destroyed.[8] We assume that the sheriff will never punish someone who has been a contributor. Notice that we are leaving out the possibility of extortion from the sheriff to contributors by assuming a minimum institutional quality.

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

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

where γ_e is the expected cost of hiring the sheriff:

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

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

This is a one-shot game and another possible approach would be to analyze it in a repeated scenario. However, given that the social class represented by the government is exogenously given, a one-shot approach has the virtue of being simpler than a dynamic consideration, which would not provide any additional insights.

The model can also be extended in several directions. Some of them have already been mentioned such as working with a proportional tax rate t or allowing for unequal sharing of the institution's cost. However, in all the mentioned potential extensions, results concerning the implementation of the sanctioning institution, the level of public good provision achieved and the level of Social Welfare attained under governments representing different social classes do not change.

Summarizing, there are two essential assumptions made in this model. The

substantially change. For simplicity, in this paper, we focus on this latter issue and we leave the determination of the tax rate for future research.

[8]It is natural to assume that this fine has to be bounded, as punishment can rarely tend to infinity.

rest of the assumptions are just useful simplifications and none of them affect the final results. The first indispensable assumption is that the government represents the interests of a particular social class. This follows the trend of thought that citizens are not politically equal in society and that our political influence depends, among other factors, on our wealth level (Przeworski, 2015; Kelly et al., 2010; Bassett et al., 1999; among others).

The second indispensable assumption in our model is that the expected cost from being discovered and punished by the sanctioning institution when there is no contribution is non-increasing on wealth. For simplicity, we work with the same constant expected cost of not contributing, $p_e f$, for all social classes. Another possibility could be to assume that this expected cost is a decreasing function of wealth. That is, wealthier have lower probabilities of being caught but pay higher fines, whereas poorer have higher chances of being detected but pay lower fines if caught. Thus, the probability is an increasing function of wealth and the fine is a decreasing function of the level of wealth. Assuming that the expected cost is non-increasing with wealth implies that the negative probability effect outweighs the positive fine effect.[9] This assumption has a strong empirical support. For instance, Alstadsæter et al. (2019) show that only wealthier individuals can afford mechanisms to conceal wealth.

3 Public good provision under a sanctioning institution

In this section, we analyze the level of public good provision when the sanctioning institution has been formed. Recall that a wealth distribution is characterized by a pair of vectors $\vec{\omega} = (\omega^P, ..., \omega^R)$ and $\vec{q} = (q^P, ..., q^R)$ where $\sum_{q^j} q^j = n$.

To hire the sheriff with a contract such that he/she exerts low effort is not an appealing case, so let us assume that if the sheriff has been offered a low-effort inducing contract, all players' 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 than the net costs. The unique Nash Equilibrium in the continuation subgame after a low-effort contract will be that everybody free rides.

Let us now analyze 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 us 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/she will do so with $g^j = 0$. This

[9] We would obtain the same results as the ones here presented if we used, for instance, the following probability function and fine function: $p(\omega) = p^* + (1 - \omega/\omega^R)$ where $p^* < \omega^P/\omega^R$ and $f(\omega) = \tau\omega$ where $0 < \tau < 1$.

corner solution is derived from the standard assumption in public goods games literature of the individuals' utility function being linear.

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/she will have to pay the same expected salary of the sheriff plus the expected fine, no matter with how much he/she slopes off. In that case, it is rationally optimal for him/her to free ride with as much as possible. The individual decision, therefore, sums up in whether to free ride with $g^j = 0$ or to fully contribute with $g^j = \omega^j$.

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/she 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 to the sheriff's intervention. Now the net cost of free riding are the expected costs of being penalized (represented 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 us 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/she 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^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 $BR^j(\vec{g}) = \omega^j \ \forall \vec{g}$*
- *If $\hat{\omega} < \omega^j \leq \tilde{\omega}$:*

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

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

where $-i$ *refers to all players other than* i.

According to the features of the best response behavior of an individual with wealth ω^j obtained in this proposition, if an individual from social class j is sufficiently poor ($\omega^j \leq \hat{\omega} \leq \tilde{\omega}$) he/she will always contribute to the public good, regardless of what others do. This type of individuals are unconditional contributors. At the other end of the spectrum, if an individual from social class j is sufficiently rich ($\hat{\omega} \leq \tilde{\omega} < \omega^j$) he/she will always free ride. These individuals are unconditional free riders. Finally, if an individual had an intermediate wealth ($\hat{\omega} < \omega^j \leq \tilde{\omega}$) his/her best response is to contribute only if everybody else does so and to free ride if there is at least one free rider. These type of individuals are conditional contributors.

Now we are ready to compute the Nash Equilibria of the contribution subgame when the sheriff exerts high effort and propose a prediction in case of multiple equilibria. The following proposition presents the equilibria selection.

Proposition 2: *Given a wealth distribution* $\vec{\omega} = (\omega^P, ..., \omega^R)$ *and* $\vec{q} = (q^P, ..., q^R)$ *and assuming the sanctioning institution has been implemented and exerts high effort, the selected equilibrium at the contribution subgame is:*

- *If* $\omega^R \leq \hat{\omega}$, *then* $g^j = \omega^j\ \forall j$
- *If* $\omega^P \leq \hat{\omega} \leq \omega^R$, *then* $g^j = \omega^j$ *if* $\omega \leq \hat{\omega}$ *and* $g^j = 0$ *if* $\omega^j > \hat{\omega}$
- *If* $\hat{\omega} < \omega^P$, *then* $g^j = 0\ \forall j$

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

Notice that according to this proposition, in an society with wealth heterogeneity, incentives to free ride grow with wealth.

Although we leave the details of the proof for the appendix, let us provide some intuition of this result.

Notice that, in many situations, the equilibrium is going to be unique. For instance, if the distribution of wealth is such that everybody is a free rider ($\hat{\omega} < \omega^P$) or everybody is a contributor ($\hat{\omega} \leq \omega^R$). The interesting cases, however, arise for wealth distributions such that we have a different mix of individuals according to their best responding behavior.

There will also be a unique Nash Equilibrium whenever there is a proportion of free riders in the population, that is, there is at least one social class whose dominant action is to free ride. Anticipating this behavior, social classes which are conditional contributors will also free ride.

Nevertheless, in the cases where there are no unconditional free riders in the population ($\omega^R \leq \tilde{\omega}$), there will exist multiple equilibria in the contribution subgame. For example, it could be the case that everybody were conditional contributors ($\hat{\omega} \leq \omega^P \leq \omega^R \leq \tilde{\omega}$) or that a proportion were unconditional contributors while the rest were conditional contributors ($\omega^P \leq \hat{\omega} \leq \omega^R \leq \tilde{\omega}$). 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 cases of multiplicity, an equilibrium selection has been made, choosing the worst possible scenario, that is, the inefficient equilibrium with a lower level of contribution. This equilibrium selection criterion is justified because it provides more robust equilibria. In the previous situations, starting in the cooperative equilibrium of universal contribution, it is enough that one individual deviates to free riding and that the rest are given the opportunity to apply their best response, to switch to the non-cooperative equilibrium. However, from the non-cooperative equilibrium, an individual deviation towards contribution and then 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.

This selection makes $\hat{\omega}$ the unique critical value for the characterization of equilibria of the contribution subgame when the sheriff exerts high effort.[10]

Recall that the introduction of a sanctioning institution aims to solve the full free-riding outcome in the provision of a public good in the absence of the institution. Thus, the performance of this sanctioning institution can be measured by the level of public good provision. The next corollary is derived from Proposition 2.

Corollary 1*: Given a fixed population size and wealth distribution and a high-effort sanctioning institution, the level of public good provision will be:*

$$G(\hat{\omega}) = \sum_{\omega^j \leq \hat{\omega}} q^j \omega^j$$

where: $\hat{\omega} = \frac{p_H f}{1-\lambda/n}$ *and* q^j *is the number of individuals with wealth* ω^j*.*

The amount of public good provision will be non-decreasing on the social return of the public good λ*, on the probability of free-riding detection under high effort* p_H *and on the fine* f*. It will be non-increasing on the population size,* n*.*

[10]As we have previously discussed in Section 2, if we had considered that the probability of being detected was decreasing with wealth and the fine was increasing with the level of wealth, we would still have a critical wealth threshold $\hat{\omega}$ with the same properties from which contributors would stop contributing. This would hold as long as the expected cost fell with wealth, i.e. the negative probability effect outweighed the positive fine effect.

Notice that the level of public good provision achieved when the institution is formed depends exclusively on 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, $\omega^R \leq \hat{\omega}$, then all social classes contribute to the public good, i.e. there is no free riding in equilibrium. We define this equilibrium as a full-contribution 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. For lower levels of quality of the sanctioning institution and of the social return of the public good, $\omega^P \leq \hat{\omega} < \omega^R$, only some social classes contribute while the others free ride. We define this situation as a partial-contribution equilibrium.

4 When does a centralized sanctioning institution emerge?

After characterizing the provision of public good under a high-effort sanctioning institution, let us present our main result which concerns the condition that must be met for a sanctioning institution to emerge in this environment under different governments and the efficiency of the achieved outcome. Before doing so, we first need to characterize the contracts offered to the sheriff.

4.1 Contracts

Before players contribute, a contract $\{s\}$ characterizing the three possible salaries s_0, s_p, s_{np} contingent to the three verifiable outcomes, must be designed for the external enforcer. There is a government who proposes this contract representing the interests of a particular social class, whose wealth will be denoted with ω^*.

Let us first characterize the minimum-cost contracts offered to the sheriff with the purpose of encouraging high or low effort. These contracts do not depend on the type of government. We assume, as it is standard in principal-agent theory and for simplicity, that the government has all the bargaining power.[11] The formal proof is relegated to the appendix.

Lemma 2*: The contracts the government offers to implement the different levels of effort are:*

- *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 that we have assumed that contribution can only occur if the sheriff exerts high effort. Therefore, the incentive constraint must be satisfied. Given a

[11] This implies that the government makes an offer and the sheriff either accepts or rejects it, with no opportunity of negotiation.

contract scheme $\{s\}$, the sheriff will exert a high level of effort if his/her utility of exerting high effort is higher than his/her utility of exerting low effort, that is, if and only if: $p_H s_p + p_L s_{np} - c \geq p_H s_{np} + p_L s_p$. Equivalently, $(p_H - p_L)(s_p - s_{np}) \geq c$.

Under automatic punishment, which is equivalent to verifiable effort, it would be enough to pay in case there is punishment, $s_p = c$ to implement high effort and the sheriff would receive his/her reservation utility, which is equal to zero. However, with non-verifiable effort, his/her expected utility with an acceptable high-effort contract, would be $p_H \frac{c}{p_H - p_L} - c = \frac{p_L}{p_H - p_L} c > c$ instead.

Notice that whilst the low-effort contract is an acceptable contract with no economic rents (that is, no additional utility beyond the reservation utility), in case the government wants to encourage the exertion of a high level of effort, he/she will have to pay economic rents $(\frac{p_L}{p_H - p_L} c)$ due to the existence of moral hazard and limited liability. These economic rents collected by the institution depend on the relative cost of high effort, c, and on the likelihood ratio $(\frac{p_L}{p_H - p_L})$ which measures how important the existing moral hazard problem in the punishment phase is, i.e. how informative the observable result is of the effort chosen. If the difference between the probability of detecting free riding behavior under high and low effort is large, then the result is fairly informative about the effort exerted and less economic rents would have to be paid. Conversely, for very similar probabilities, the verifiable results would yield little information about the institution's effort and, consequently, more economic rents would have to be paid.

The government will prefer to offer the high-effort contract when the expected utility of an individual of the social class it represents 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 utility of the social class represented by the government would be:

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

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 not offering a contract at all, given that $p_L f \geq 0$. Therefore, the government must, in fact, decide whether to offer an acceptable contract, which additionally encourages high effort or to not hire a sheriff at all.

4.2 Implementation and efficiency of the sanctioning institution under different governments

The ultimate decision the government faces is whether or not to implement the sanctioning institution. In this context, it will compare the utility of an individual of the social class it represents with and without sheriff. The next result characterizes under which conditions will the sanctioning institution be formed.

Proposition 3: *Assume that $\omega^P \leq \hat{\omega}$ and the social class represented by the government has wealth ω^*. The sanctioning institution will be implemented if and only if:*

$$(6) \qquad \frac{\lambda}{n} G(\hat{\omega}) \geq \frac{p_H s_k}{n} + \begin{cases} \omega^* & \text{if } \omega^* \leq \hat{\omega} \\ p_H f & \text{if } \omega^* > \hat{\omega} \end{cases}$$

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

The case where $\omega^P > \hat{\omega}$ lacks of any interest because even in the presence of a sheriff exerting high effort, nobody is going to contribute. Therefore we focus on the 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 under the threat of a high-effort sanctioning institution.

According to Proposition 3, the implementation of a high-effort sanctioning institution depends on the interaction of three factors. Firstly, on a set of institutional and technological parameters, which include the social return generated by the public good, λ, the expected capacity of punishment, $p_H f$, and the severity of the moral hazard problem, captured by the economic rents $c\frac{p_L}{p_H - p_L}$, to be paid to the enforcer. Secondly, on the wealth distribution in the group, defined by the levels of wealth of each social class, $\vec{\omega} = (\omega^P, ..., \omega^R)$, and the number of citizens belonging to each group, $\vec{q} = (q^P, ..., q^R)$. And finally, on the cost in terms of wealth faced by an individual of the social class represented by the government when the sanctioning institution is implemented. Specifically, a contributor renounces to his wealth, ω^j, while a free rider pays the fine with probability p_H.

Below, we define the relation "implement the institution in a greater range of cases", useful in the characterization of our results.

Definition 1: *We say that a government representing social class j, in short, a government j, implements the institution in a greater range of cases than a government k , if, given a set of parameters and a wealth distribution, the government j implements the institution in every situation that the government k does and additionally, the government j implements the institution in situations that the government k does not.*

A natural question is which government will implement the punishing institution in a greater range of cases and whether doing so is efficient or not. We are going to answer these questions by separately studying the full-contribution and partial-contribution equilibria.

4.2.1 Full-contribution equilibria

We will start by analyzing the particular case where all social classes contribute under the threat of a sanctioning institution. This will occur under remarkably high levels of quality of the sanctioning institution and of the social return of the public good, resulting in $\omega^R \leq \hat{\omega}$ and a full-contribution equilibrium. The next corollary is derived from Proposition 3.

Corollary 2: *In a full-contribution equilibrium ($\omega^R \leq \hat{\omega}$), where all social classes contribute, governments' incentives to implement the sanctioning institution are decreasing with the level of wealth of the class they represent.*

This result is driven by the fact that the critical threshold $\hat{\omega}$ defining the level of public good provision and the expected wage of the sheriff are fully determined by the wealth distribution and the set of institutional and technological parameters. Therefore, the government under which the sanctioning institution emerges in a greater range of cases is the one representing the social class with the lowest cost for implementing the institution. If we consider, for instance, three social classes: a government representing a poor-class individual will implement the institution in a greater range of cases than a government representing a middle-class individual, which in turn, will do so in a greater range of cases than a government representing a rich-class individual.

We now address what is the level of Social Welfare (SW hereinafter) achieved by the implementation of a sanctioning institution and, especially, 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 sum of the endowments: $\sum_{\omega^j} q^j \omega^j$, denoted by W from now onward.

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 $\lambda > 1$. In a full-contribution equilibrium where all social classes contribute, then $SW = \lambda W - s_0$. Recall that for our model $s_0 = \overline{u} = 0$, because we have assumed that the sheriff's reservation utility is 0. For this case or even for other cases with $\overline{u} > 0$ sufficiently close to 0, SW will be very close to the one obtained in the efficient outcome, i.e. $SW \approx \lambda W$. Thus, the implementation of a sanctioning institution that enhances the exertion of high effort will always achieve the fully-efficient outcome.

Recall that for the case of full contribution, a government representing a social class with wealth ω^* will implement the institution if and only if $\frac{\lambda}{n} W \geq \omega^*$, or equivalently $\lambda W \geq n\omega^*$. We define as $\langle\omega\rangle = \frac{W}{n}$ the average wealth in the population. Thus, the condition for a government to implement the institution can also be expressed as $\lambda\langle\omega\rangle \geq \omega^*$.

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

Proposition 4: *In a full-contribution equilibrium, $\omega^R \leq \hat{\omega}$, a government will always make the socially efficient decision of implementing the sanctioning institution only if $\omega^* \leq \langle\omega\rangle$. Otherwise, if $\omega^* > \langle\omega\rangle$, it will not be implementing it in situations where it is socially efficient to do so.*

The proof of this result is straightforward. Recall that the condition for a government representing a social class with wealth ω^* to implement the sanctioning institution can be written in terms of the average population wealth, i.e. $\lambda\langle\omega\rangle \geq \omega^*$. If $\langle\omega\rangle \geq \omega^*$ holds, then $\lambda\langle\omega\rangle \geq \omega^*$ is also going to hold, given that $\lambda > 1$. If, however, $\langle\omega\rangle < \omega^*$, it is unclear whether the implementation condition is going to hold, and it will essentially depend on the value of λ.

Corollary 3: *In a full-contribution equilibrium, if $\omega^* > \langle\omega\rangle$, the government will efficiently implement the sanctioning institution only if the social return of the public good is sufficiently high, namely, $\lambda \geq \frac{\omega^*}{\langle\omega\rangle} > 1$.*

Intuitively, even though it is efficient to implement the institution if, with it, full contribution can be achieved, a government representing a social class with a level of wealth higher than the average wealth of the society may not be interested in doing so if wealth inequality is too pronounced (i.e. average wealth is very low) and/or the social return of the public good is too low. Each citizen receives an n^{th} part of the return of the public good composed of contributions of all the social classes. If we recall the example of three social classes, poor, middle and rich, for certain the poor class will be better off given that they are contributing with the lowest amount. Analogously, the rich will surely be worse but it is unclear whether it will pay off for the middle class.[12] Ultimately, this will depend on the average wealth of the society and on the social return of the public good λ.

Let us now propose a simple numerical example with three social classes to illustrate the different scenarios that may arise. Suppose a population of 100 citizens where the number of individuals belonging to each social class and their level of wealth is summarized in the following table.

Level of wealth	Social class size
$\omega^P = 200$	$q^P = 60$
$\omega^M = 400$	$q^M = 35$
$\omega^R = 1000$	$q^R = 5$
	$n = 100$

Given these values, total wealth in the population is equal to $W = 31000$

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

and the average wealth is $\langle\omega\rangle = 310$. Let us also suppose that $p_H f = 1000$ and $\lambda = 3$ such that the critical contributing wealth threshold is $\hat{\omega} = 1030.93$. Consequently, all social classes contribute to the public good as $\omega^R < \hat{\omega}$ and the total public good provision is equal to $G = W = 31000$. Recall that the condition for a government representing the interests of social class j to implement the sanctioning institution is found in Equation (**??**).

Thus, substituting the values for each social class we would have the following conditions for each type of government. A poor-class government will always implement the sanctioning institution, because $\frac{\lambda}{100}31000 \geq 200$ holds for any $\lambda > 1$. A middle-class government will implement the sanctioning institution only if $\frac{\lambda}{100}31000 \geq 400$, that is, if $\lambda \geq 1.29$. Finally, a rich-class government will do so only if $\frac{\lambda}{100}31000 \geq 1000$, that is, if $\lambda \geq 3.23$.

Therefore, whether the government implements the sanctioning institution or not, becomes more difficult as wealth increases, as expressed in Corollary 2. In this case, given that $\lambda = 3$, a poor and a middle-class government will make the efficient decision of implementing the institution. However, the rich-class will not do so because λ is not sufficiently high, as expressed in Corollary 3.

4.2.2 Partial-contribution equilibria

For lower levels of quality of the sanctioning institution and of the social return of the public good, we obtain a partial contribution equilibrium under the institution, where only some social classes contribute while the others free ride. Notice that full contribution was only a particular case of what a sanctioning institution can achieve, whereas attaining the contribution of some social classes might be a more realistic prospect in many societies. Notice that $p_H f < \hat{\omega}$. The next corollary is again derived from Proposition 3.

Corollary 4: *In a partial-contribution equilibrium, $\omega^P \leq \hat{\omega} < \omega^R$, for any government j such that $\omega^j \leq \hat{\omega}$ (a government that represents a contributing class) and any k such that $\omega^k > \hat{\omega}$ (a government that represents a free-riding class), if $\omega^j \leq p_H f$, then government j will implement the sanctioning institution in a greater range of cases than government k. However, if $\omega^j > p_H f$, then the opposite holds.*

Following the same intuition as with Corollary 2, these results are driven by the fact that the government under which the sanctioning institution emerges in a greater range of cases is the one representing the social class with the lowest cost for implementing the institution. Notice that if expected fines are sufficiently low, $p_H f \leq \omega^P$, free riders will have a lower cost than any contributor. Therefore, governments representing free riders will implement the sanctioning institution in a greater range of cases than any other. If, alternatively, expected fines were sufficiently large, $\omega^P < p_H f$, then a government representing the poorer social class will implement the institution in a greater range of cases than any other government. However, even in this case, a government representing a free rider can dominate a government representing a contributor with a sufficiently large wealth ω^j if $\omega^P < p_H f < \omega^j \leq \hat{\omega}$.

The result concerning low expected fines, where a government representing a free-riding class, for instance, a rich social class will implement the sanctioning institution in a greater range of cases, may seem counter intuitive. However, recall that everybody equally benefits from the public good independently of their contribution decision. A free rider receives an n^{th} of the public fund in the same way a contributor does. However, in the context here presented, the public good can only have a positive provision if the sanctioning institution is implemented. If the expected cost a free rider faces in the implementation of this enforcing mechanism is sufficiently low, the public good benefits will overcome the costs and, paradoxically, a government representing this free riding social class will implement the sanctioning institution in a greater range of cases than a government representing a contributing social class.

We now address what is the level of Social Welfare (SW) achieved with the implementation of a sanctioning institution by different governments in a partial-contribution equilibrium. In this case, the level of SW would be:

$$SW = \lambda W - [(\lambda - 1) \sum_{\omega^j > \hat{\omega}} q^j \omega^j + m(p_H f) + \frac{p_H}{p_H - p_L} c] \tag{7}$$

where $\hat{\omega} = \frac{p_H f}{1 - \frac{\lambda}{n}}$ *and* $m = \sum_{\omega^j > \hat{\omega}} q^j$ *is the number of free riders in the population.*

Therefore, at a first glance, we can see that there is a welfare loss captured by the second term of Equation (**??**) which entails, on the one hand, the net loss of not contributing by all the free-riding social classes plus the expected fines for all the free riders and, on the other hand, the institutional costs (sheriff's salary). This SW loss is increasing with the number of free riders in the population. 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:

$$\frac{\lambda}{n} G(\hat{\omega}) \geq \frac{p_H}{n(p_H - p_L)} c + \frac{G(\hat{\omega}) + m(p_H f)}{n} \tag{8}$$

where $G(\hat{\omega}) = \sum_{\omega^j \leq \hat{\omega}} q^j \omega^j$.

Alternatively, a government representing a social class with wealth ω^* will implement the sanctioning institution if and only if Equation (**??**) holds.

Notice these two conditions are equivalent on the left hand side and differ on the right hand side. While governments consider the cost of the individual they are representing, ω^* if it is a contributor or $p_H f$ if it is a free rider, from the point of view of a Social Planner, it is the average of everybody's cost what is being taken into account which we denote as $AC(\hat{\omega}) = \frac{G(\hat{\omega})+m(p_H f))}{n}$. In other words, the social criterion is different from the one followed by the government. Only if, coincidentally, these costs were equal, the decision of this government would always be efficient.

From the comparison of these two conditions, we can assert the following results, where we refer to over implementation when the institution is implemented when it is efficient but also in cases where it is inefficient to do so and, under implementation when it is not implemented in cases where it would be efficient.

Proposition 5*: In a partial-contribution equilibrium, $\omega^P \leq \hat{\omega} < \omega^R$, if the government represents a social class with an individual cost for implementing the institution smaller than the average population cost, $AC(\hat{\omega})$, then it will over implement the sanctioning institution, while if the individual cost is higher than $AC(\hat{\omega})$, then it will under implement the institution. Only the government of the social class with a wealth equal to $AC(\hat{\omega})$ will always make the social welfare maximizing decision.*

Therefore, partial-contribution equilibria almost always lead to inefficient over and under implementation of the sanctioning institution. As with Proposition 4, the driving force of this result is that a government representing a social class with the lowest cost will implement the institution in situations where it is inefficient to do so while a government representing a social class with the highest cost will decide not to implement the institution in situations where it is efficient to do so.

Let us now draw some conclusions on how the expected fine affects the incentives to over or under implement the institution under different types of governments. On the one hand, notice that $AC(\hat{\omega}) < \hat{\omega}$ always holds. On the other hand, the average population cost of the institution $AC(\hat{\omega})$ is actually a weighted average of the expected fines and the average level of wealth of the contributors. That is, $AC(\hat{\omega}) = \frac{n-m}{n}(\frac{G(\hat{\omega})}{n-m})+\frac{m}{n}(p_H f)$.

When $p_H f$ is higher than the average level of wealth of contributors, $p_H f > \frac{G(\hat{\omega})}{n-m}$, then $\frac{G(\hat{\omega})}{n-m} < AC(\hat{\omega}) < p_H f$ always holds and the government representing a free-riding class will always be under implementing the sanctioning institution. If, instead, $p_H f$ is found below the average level of wealth of contributors, $p_H f < \frac{G(\hat{\omega})}{n-m}$, then it will always be over implementing it, because $AC(\hat{\omega}) > p_H f$.

The following proposition derives from Proposition 5 and the previous analysis.

Proposition 6*: In a partial-contribution equilibrium, $\omega^P \leq \hat{\omega} < \omega^R$, a gov-*

ernment representing a free-riding social class ($\omega^* > \hat{\omega}$), will over implement the institution if $p_H f < \frac{G(\hat{\omega})}{n-m}$. i.e. for low expected fines; whereas it will under implement the institution if $p_H f > \frac{G(\hat{\omega})}{n-m}$, i.e. for high expected fines.

Conclusions in this line cannot be drawn for governments representing contributors. Over or under implementation exclusively depend on the relation between their level of wealth and the average population cost of the institution.

Taking up the numerical example proposed before, suppose now that the expected fine is relatively high, $p_H f = 800$, such that $\hat{\omega} = 824.74$ and, consequently, the poor and middle classes have incentives to contribute while the rich class free rides on the public good ($\omega^P < \omega^M < \hat{\omega} < \omega^R$). The public good provision would now be $G = 26000$. Substituting the values for each social class into the condition expressed in Equation (**??**) we obtain the following specific condition for implementation. A poor-class government will implement if $580 \geq \frac{p_H}{n(p_H - p_L)}$ and a middle-class government will implement if $380 \geq \frac{p_H}{n(p_H - p_L)}$. However, a free-riding rich class government will do so if $(-20) \geq \frac{p_H}{n(p_H - p_L)}$, which is obviously impossible. Therefore, a free-riding rich-class government will never implement the institution because the expected salary of the sheriff will always be positive.

If we explore SW in this case, we can compute the condition for a Social Planner to implement the sanctioning institution, which would be: $480 \geq \frac{p_H}{n(p_H - p_L)}$. By comparing the condition for the Social Planner with the previous conditions, we can see that a poor-class government will be implementing the institution in cases where it would not be efficient to do so (over implementation), given that $\omega^P < AC(\hat{\omega})$, as expressed in Proposition 5. However, the condition for the Social Planner is less strict than the conditions for the contributing middle class and the free-riding rich class. This implies that these governments could be not implementing a sanctioning institution in cases where a Social Planner would do (under implementation). For the case of free riders, this happens because, in this case, the expected fine that these pay with the institution is sufficiently high, $AC(\hat{\omega}) < p_H f$.

If, alternatively, we decrease the expected fine to $p_H f = 195$, then $\hat{\omega} = 201.03$ and only the poor class contributes. The middle class now also finds it profitable to free ride. Total contributions to the public good would fall to $G = 12000$ and the total average cost of the institution would be $AC(\hat{\omega}) = 198$. Now a government representing the poor class would under implement the sanctioning institution, while governments representing the interests of a free-riding class, would be over implementing the institution, given that this variation makes $p_H f < AC(\hat{\omega}) < \omega^P$. The maybe counter intuitive result of having a free-riding class over implementing a sanctioning institution is that it brings high profits (public good) at a low cost (low expected fine of being caught free riding).

5 Conclusions

This paper's aim has been to understand how do sanctioning institutions work in societies conformed by citizens who are diverse in terms of wealth, i.e. they can be classified into social classes according to their level of wealth. We propose the provision of a public good in a society of selfish individuals. The government, who represents the interests of a particular social class, can implement a sanctioning institution subject to moral hazard. Specifically, we analyze when will this kind of sanctioning institution be implemented, its performance and its efficiency.

Without any enforcing mechanism or with an institution with inappropriate incentives, selfish individuals will fully free ride leading to a non-provision of the the public good. If this is not the case and a correctly incentivized sanctioning institution is implemented, a positive provision of the public good can be achieved. The individual incentives to contribute to such good are decreasing with wealth. In other words, the rich have more incentives to free ride than the poor.

Regarding the implementation of a sanctioning institution, we show that the government representing the social class with the lowest cost will implement it in a wider range of cases. The specific answer to which government will it be depends on the degree of contribution that the sanctioning institution can potentially achieve, which depends on the monitoring and punishment technology and on the valuation of the public good. If the sanctioning institution can attain full contribution, then it will be the poor class. The decision of implementing the institution in this case will be efficient. If, however, the institution can only achieve contribution by some social classes, the government representing the poor class's interests will only be the one implementing the institution more frequently if the expected fines of free riding are sufficiently high. Alternatively, in societies with low-quality sanctioning institutions, a government representing a free-riding social class will implement the institution more often. We also show how in most cases the decision is inefficient, leading to over or under implementation of the sanctioning institution, both from governments representing contributing and free-riding social classes.

Our theory highlights the importance of wealth heterogeneity and of the incentives of the social class represented by the government in the collective action problem. We show how his/her incentives to free ride or to contribute not only affect his/her contributing behaviour, but also crucially affect the emergence of a sanctioning institution and the provision of the public good. Such incentives would be disguised in the usual homogeneous wealth society model.

Appendix

Proof of Proposition 2:

Let us analyze the different cases:

Case 1. $\omega^R \leq \hat{\omega} \leq \tilde{\omega}$: 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. $\omega^P \leq \hat{\omega} \leq \tilde{\omega} \leq \omega^R$: Recalling Proposition 1, there would be a set of players, those with $\omega^j \leq \hat{\omega}$, who 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 ($\tilde{\omega} < \omega^j$) who 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 $g_i^j = 0 \ \forall i, j$. *Q.E.D.*

Proof of Lemma 2:

The characterization of s_0 (the salary paid in case the sheriff's intervention is finally not necessary because everybody contributes) is straightforward. Recall $\overline{u} = 0$, so $s_0 = 0$.

Let us characterize the contract where the government, maximizing the utility of the social class with wealth ω^*, 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:

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

The government maximizes the utility of the social class it represents subject to the sheriff's participation and incentive constraint. Furthermore, the sheriff has limited liability, so salaries must all be non-negative. Although this problem can be solved building a Lagrangian and computing its first derivatives, given its simplicity, the solution can be rigorously obtained with a proof by contradiction.

Recall that the government has all the bargaining power when negotiating with the sheriff. The contract that maximizes the utility of the government (objective function), that is, the solution to the problem, must be such that both the limited liability constraint for s_{np} and the incentive constraint are binding. Otherwise, the government could always decrease s_p and s_{np} to increase its utility, while the participation constraint of the sheriff still holds. Therefore, from these two constraints satisfied with equality, we derive the contract enhancing high effort that maximizes the utility of the government: $\{s_0 = 0, s_p = \frac{c}{p_H - p_L}, s_{np} = 0\}$. Under this contract, as explained in the main text, the sheriff obtains positive economic rents: $\frac{p_L}{p_H - p_L}c$.

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

References

Acemoglu, Daron, and Alexander Wolitzky (2020), "Sustaining Cooperation: Community Enforcement versus Specialized Enforcement," *Journal of the European Economic Association,* 18(2), 1078-1122.

Aldashev, Gani, and Giorgio Zanarone (2017), "Endogenous enforcement institutions," *Journal of Development Economics,* 128, 49-64.

Alstadsæter, Annette, Niels Johannesen, and Gabriel Zucman (2019), "Tax evasion and inequality," *American Economic Review,* 109(6), 2073-2103.

Andreoni, James, and Laura K. Gee (2012), "Gun for hire: Delegated enforcement and peer punishment in public goods provision," *Journal of Public Economics,* 96(11-12), 1036-1046.

Baldassarri, Delia, and Guy Grossman (2011), "Centralized sanctioning and legitimate authority promote cooperation in humans," *Proceedings of the National Academy of Sciences,* 108(27), 11023-11027.

Bassett, William F., John P. Burkett, and Louis Putterman (1999), "Income distribution, government transfers, and the problem of unequal influence," *European Journal of Political Economy*, 15(2), 207-228.

Buckley, Edward, and Rachel Croson (2006), 'Income and wealth heterogeneity in the voluntary provision of linear public goods," *Journal of Public Economics*, 90(4-5),

935-955.
Boyd, Robert, and Peter J. Richerson (1988), “The evolution of reciprocity in sizable groups," *Journal of theoretical Biology,* 132(3), 337-356.
Burlando, Roberto M., and Francesco Guala (2005), “Heterogeneous agents in public goods experiments," *Experimental Economics,* 8(1), 35-54.
Cherry, Todd L., Stephan Kroll, and Jason F. Shogren (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.
Dong, Yali, Tatsuya Sasaki, and Boyu Zhang (2019), “The competitive advantage of institutional reward," *Proceedings of the Royal Society B,* 286(1899), 20190001.
Fehr, Ernst, and Simon Gachter (2000), “Cooperation and punishment in public goods experiments," *American Economic Review,* 90(4), 980-994.
Fehr, Ernst, and Tony Williams (2017), “Creating an efficient culture of cooperation", *No. 267. Working paper.*
Gächter, Simon, Elke Renner, and Martin Sefton (2008), “The long-run benefits of punishment," *Science,* 322(5907), 1510-1510.
Greif, Avner (2003), *Institutions: Theory and history,* Cambridge University Press, New York.
Greif, Avner (1993), “Contract enforceability and economic institutions in early trade: The Maghribi traders’ coalition," *The American economic review,* 525-548.
Hauser, Oliver P., Christian Hilbe, Krishnendu Chatterjee and Martin A. Nowak (2019), “Social dilemmas among unequals," *Nature,* 572(7770), 524-527.
Heap, Shaun P. Hargreaves, Abhijit Ramalingam, and Brock V. Stoddard (2016), “Endowment inequality in public goods games: A re-examination," *Economics letters,* 146, 4-7.
Kelly, N. J., & Enns, P. K. (2010). Kelly, Nathan J., and Peter K. Enns (2010), "Inequality and the dynamics of public opinion: The self-reinforcing link between economic inequality and mass preferences," *American Journal of Political Science,* 54(4), 855-870.
Kosfeld, Michael, Akira Okada, and Arno Riedl (2009), ‘Institution formation in public goods games." *American Economic Review,* 99(4), 1335-55.
Kurrild-Klitgaard, Peter, and Gert Tinggaard Svendsen (2003), ‘Rational bandits: Plunder, public goods, and the Vikings," *Public Choice,* 117(3-4), 255-272.
Traulsen, Arne, Torsten Röhl, and Manfred Milinski (2012), ‘An economic experiment reveals that humans prefer pool punishment to maintain the commons," *Proceedings of the Royal Society B: Biological Sciences,* 279(1743), 3716-3721.
Okada, Akira (1993), “The possibility of cooperation in an n-person prisoners’ dilemma with institutional arrangements," *Public Choice,* 77(3), 629-656.
Przeworski, Adam, (2015). “Economic inequality, political inequality, and redistribution," *Mimeo.*
Romaniuc, Rustam, Katherine Farrow, Lisette Ibanez and Alain Marciano (2016), “The perils of government enforcement," *Public Choice,* 166(1-2), 161-182.
Sasaki, T., Brännström, Å., Dieckmann, U., & Sigmund, K. (2012). The take-it-or-leave-it option allows small penalties to overcome social dilemmas. Proceedings of the National Academy of Sciences, 109(4), 1165-1169.
Sasaki, Tatsuya, Åke Brännström, Ulf Dieckmann and Karl Sigmund (2012), “The take-it-or-leave-it option allows small penalties to overcome social dilemmas," *Pro-*

ceedings of the National Academy of Sciences, 109(4), 1165-1169.

Sigmund, Karl, Christoph Hauert, Arne Traulsen and Hannelore De Silva (2011), “Social control and the social contract: the emergence of sanctioning systems for collective action," *Dynamic Games and Applications,* 1(1), 149-171.

Sutter, Matthias, Stefan Haigner, and Martin G. Kocher (2010), “Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations," *The Review of Economic Studies,* 77(4), 1540-1566.

Tommasi, Mariano, and Federico Weinschelbaum (2007), “Centralization vs. decentralization: A principal-agent analysis," *Journal of public economic theory,* 9(2), 369-389.

Wu, Jia-Jia, Cong Li, Bo-Yu Zhang, Ross Cressman and Yi Tao (2014), “The role of institutional incentives and the exemplar in promoting cooperation," *Scientific reports,* 4, 6421.

Yamagishi, Toshio (1986), “The provision of a sanctioning system as a public good," *Journal of Personality and social Psychology,* 51(1), 110.

Zhang, Boyu, Cong Li, Hannelore De Silva, Peter Bednarik and Karl Sigmund (2014), “The evolution of sanctioning institutions: an experimental approach to the social contract," *Experimental Economics,* 17(2), 285-303.

Adriana Alventosa
Av. Naranjos, s/n, 46022, Valencia
Spain
Adriana.Alventosa@uv.es

Gonzalo Olcina
Av. Naranjos, s/n, 46022, Valencia
Spain
Gonzalo.Olcina@uv.es