Punishment and efficiency: theoretical and experimental approaches

Adriana Alventosa Baños

Supervisors: Penélope Hernández Rojas, Gonzalo Olcina Vauteren

ERI Comportament Econòmic i Social (ERI-CES)

Departament d'Anàlisi Econòmica

Facultat d'Economia

Universitat de València

july 2018

Documento maquetado con TEXIS v.1.0+.

Punishment and efficiency: theoretical and experimental approaches

Adriana Alventosa Baños

Supervisors: Penélope Hernández Rojas, Gonzalo Olcina Vauteren

ERI Comportament Econòmic i Social (ERI-CES)

Departament d'Anàlisi Econòmica

Facultat d'Economia

Universitat de València

july 2018

A mi padre,

mi estrella en el cielo.

Agradecimientos

No encuentro, y seguramente sea porque no hay, palabras suficientes para agradecer todo lo que han hecho por mí mis directores, mis mentores, mis padres académicos. Y cuando no hay palabras, solamente puedo dar las gracias. Gracias por cuidarme, por vuestro tiempo, por vuestra dedicación e implicación no solo durante estos tres años de tesis. Gracias Penélope por los últimos cinco años, por creer en mí desde el primer día que me viste cruzar la puerta y empezaste a ponerme retos encima de la mesa. Gracias por creer que ser perfeccionista no es un defecto y por ayudarme a no perder nunca ni las ganas ni la sonrisa. Gracias Gonzalo por todo lo que me has enseñado en siete años, por dejar que me equivocara y que acertara, por estar dispuesto a dirigir una tesis después de tanto tiempo y por estar siempre que el reloj ha contado las últimas horas. Siempre a contrarreloj pero siempre lo hemos conseguido.

Quisiera dedicar también unas líneas de agradecimiento a la Estructura de Recerca Interdisciplinar Comportament Econòmic i Social (ERI-CES), al Departament d'Anàlisi Econòmica (tanto a su PAS como a su PDI), a la Facultat d'Economia y a los profesores del Master en Economía. Considero que merecen también una mención especial la directora de ERI-CES y del

Programa de Doctorado Amparo Urbano y sobre todo, mi tercer director en la sombra, Vicente Calabuig. No hay corazón más grande ni chistes más malos que los tuyos.

Gracias a mi madre simplemente por ser madre y sufrir conmigo sin entender a veces muy bien qué estaba pasando. A Alberto por ser mi otra mitad y la persona no académica que mejor entiende este mundo. Gracias a los dos por estar en los momentos más importantes de esta y de otras etapas y celebrarlo conmigo. Gracias al resto de mi familia y amigos, especialmente a mis hermanos, a mis tías, a mi prima Bea y a mi amiga Cris.

Quizás lo que menos esperaba de esta etapa era hacer tantos amigos. Ha sido un placer conocer a todas las personas que han formado y forman parte del equipo LINEEX en los últimos cinco años, especialmente a Alberto, Rebeca y Helena por vuestro cariño. A todos los compañeros de fatigas de la Facultat d'Economia: Adri, Carlos, Danny, Eli, Ernesto, Jesús, Jorge, Laura, Luis, Mariola, Marta Solaz, Marta Suárez-Varela, Paula y Rubén; así como de otras facultades de la Universitat. Gracias Rocío porque, sin tú saberlo, me ayudaste mucho más de lo que podías imaginar.

Y no puedo cerrar esta etapa sin recordar a mi estrella en el cielo. Por irte demasiado pronto pero seguir a mi lado cada día. Sé que todo va a ir bien mientras tú me guíes, papá.

Contents

List of Figures

List of Tables

Introducción

Los dilemas sociales se caracterizan por la desalineación de incentivos privados y sociales bajo preferencias egoístas. Mientras el óptimo social se alcanza mediante la implementación de un conjunto particular de acciones, los incentivos individuales mueven a los agentes a comportarse de una manera diferente, lo cual lleva a resultados ineficientes. Un ejemplo clásico es la provisión de un bien público, donde el óptimo se alcanza si todos contribuyen, sin embargo hay una desviación unilateral provechosa a hacer de polizón, es decir, a no contribuir al bien público y beneficiarse de los demás. De este modo, dar a los agentes egoístas la posibilidad de elegir libremente sus asignaciones lleva al subabastecimiento del bien público y a que surja el "problema del polizón". Por este motivo, los dilemas sociales han sido cruciales en el estudio del comportamiento humano desde el origen de la economía del comportamiento.

Con el fin de mitigar el problema del polizón se han propuesto numerosos mecanismos, de los cuales el castigo ha sido el que ha adquirido mayor relevancia. Dar a los individuos la oportunidad de sancionarse los unos a los otros consigue, en términos generales, que disminuya el comportamiento indeseado y acorta la distancia al resultado eficiente. No obstante, no cualquier tipo de sistema sancionador es suficiente para cumplir dicho propósito. En esta tesis presento tres trabajos de investigación, dos teóricos y uno experimental, que estudian tres estructuras de castigo que promueven la eficiencia en dilemas sociales.

La literatura tanto teórica como experimental ofrece distintos esquemas de castigo que han sido empleados tanto en marcos teóricos como experimentales. El pionero ha sido el castigo entre pares, donde al final del juego los individuos observan lo que otros han decidido y tienen la oportunidad de sancionar a nivel individual. Dado que castigar es globalmente considerado una acción costosa, los individuos egoístas no deberían implementar esta acción en juegos finitos como consecuencia a mejor respuesta. Sin embargo, la sociedad no solamente está formada for agentes egoístas ya que, de hecho, la mayoría tenemos algún tipo de preocupación social como aversión a la desigualdad, reciprocidad o altruismo. Esto hace que, aún siendo costoso, los individuos castiguen "racionalmente" a compañeros polizones como mejor respuesta considerando que la función de utilidad recoge esta aversión. No obstante, un inconveniente frecuente de este tipo de castigo es el exceso de castigo ineficiente. Además, muchos trabajos han hecho hincapié en la falta de realismo existente en la posibilidad de implementar castigo a nivel individual.

Como respuesta a las desventajas del castigo entre pares, la literatura ha propuesto dos esquemas de castigo alternativos que utilizo en este trabajo: (i) castigo delegado y (ii) castigo coordinado.

El castigo delegado es una estructura centralizada de sanción donde un agente en particular, preferiblemente externo al desarrollo del juego, es dotado de poder sancionador para castigar a polizones. Asimismo, este agente o institución no es un ser automático pero una entidad estratégica que también debe tener los incentivos correctos para actuar apropiadamente. Este tipo de castigo solventa el resultado de exceso de castigo del castigo entre pares y el no realismo de las autoridades sancionadoras descentralizadas. Un ejemplo de castigo delegado son los recaudadores de impuestos.

El castigo coordinado, por otro lado, es un sistema de sanción descentralizado que presenta dos características llamativas no existentes en esquemas previos. En primer lugar, señala que los costes de castigo deben ser divididos entre los individuos dispuestos a implementar el castigo. Es decir, deberían mostrar rendimiento a escala crecientes. En segundo lugar, para que las acciones de sanción sean efectivas, deben requerir cierto número de castigadores. De no ser así, no se destruirá el pago del polizón. Un ejemplo de castigo coordinado es una huelga.

En el Capítulo 2 de esta tesis presento una revisión bibliográfica describiendo el problema del polizón así como el impacto de relajar los distintos supuestos del juego de bienes públicos estándar: heterogeneidad de riqueza, heterogeneidad de productividad, información y preferencias sociales. Además, propongo cuatro mecanismos para afrontar el problema, haciendo particular énfasis en el castigo. Finalmente, examino el estado del arte de varios esquemas de sanción: castigo entre pares, contracastigo, castigo coordinado y castigo delegado.

El Capítulo 3 es un trabajo teórico analizando el surgimiento y desempeño de una institución sancionadora en un contexto de juego de bienes públicos. Se presenta una sociedad con distintos niveles de riqueza cuyos ciudadanos, por sí mismos, no pueden conseguir una provisión positiva del bien público. Dicha sociedad, a través del gobierno representando el interés de una clase social en concreto, debe decidir si implementar una institución de castigo centralizada, en línea con la literatura del castigo delegado. Este trabajo analiza bajo qué condiciones se implementará una institución sancionadora de alto rendimiento, qué nivel de provisión de bien público conseguirá, y finalmente, su eficiencia.

Este modelo considera preferencias egoístas con el objetivo de evaluar el peor escenario posible. Si los ciudadanos de esta sociedad tuvieran algún tipo de preocupación social, los resultados aquí presentados se verían impulsados. Por este motivo, la metodología utilizada en este capítulo es la Teoría de Juegos, con el propósito de modelizar la interacción estratégica de los agentes económicos y caracterizar los Equilibrios de Nash del juego.

El Capítulo 4 es un trabajo experimental que explora el impacto de dos esquemas de pago diferentes en un entorno centralizado de castigo. En particular, presento un juego de bienes públicos con 336 sujetos agrupados en grupos de cuatro personas, donde tres de ellos son contribuyentes y uno de ellos es un sancionador. Los primeros solamente contribuyen al bien público mientras que el último únicamente lleva a cabo acciones de castigo siguiendo un esquema de castigo delegado. La principal cuestión a abordar es cómo debería funcionar la implementación de estas instituciones centralizadas, para lo cual comparamos dos esquemas: (i) un esquema fijo donde se proporciona al sancionador cierta dotación para decidir sobre sus decisiones de castigo y (ii) un esquema variable, donde, en cambio, el sancionador recibe una dotación proporcional al nivel de cooperación conseguido. Este trabajo resalta los beneficios de sistemas centralizados de castigo con esquemas de pago fijos en términos de contribuciones y eficiencia.

En este capítulo, hago uso de la metodología de la Teoría de Juegos para caracterizar los Equilibrios de Nash del juego bajo el supuesto de preferencias estándar. Además, dado que los decisores del mundo real no son siempre egoístas, utilizo la Economía Experimental para contrastar estos resultados. Con este propósito, se ha seguido cuidadosamente el protocolo de Economía Experimental en el diseño e implementación de las sesiones con el fin de garantizar que los datos derivados fueran apropiados para el análisis. El diseño, a su vez, ha obtenido la aprobación del Comité de Ética de la Universitat de València.

En último lugar, el Capítulo 5 de esta tesis es un trabajo teórico comparando dos esquemas de castigo descentralizados en un juego de confianza en equipo con asimetrías de información: (i) un sistema no coordinado donde el castigo individual destruye el pago del agente castigado y (ii) un esquema coordinado donde es necesario que el número de individuos dispuestos a castigar exceda un límite determinado para que el castigo sea efectivo. Los resultados desvelan que un sistema de castigo coordinado lleva a equilibrios eficientes en un mayor rango de casos que el castigo no coordinado siempre que la proporción de reciprocadores en la población de inversores sea suficientemente alta.

Más allá del uso de la Teoría de Juegos en la caracterización de los Equilibrios Nash del juego, este trabajo introduce preferencias sociales. En concreto, considera que los inversores del juego de confianza en equipo pueden ser o bien reciprocadores o egoístas, y que el asignador de recursos puede ser o bien imparcial o maximizador de beneficios. Caracterizo los Equilibrios de Nash Bayesianos del juego haciendo uso de la Teoría de Juegos en general, y de la Economía del Comportamiento en particular.

Chapter 1

Introduction

Social dilemmas are characterized by the misalignment of private and social incentives under selfish preferences. While the social optimum is reached with the implementation of a particular set of actions, private incentives move agents to behave in a different way, leading to inefficient outcomes. A classic example is the provision of a public good, where the optimum is reached if everybody contributes to it, but there is a profitable unilateral deviation to free ride, that is, to not contribute to the public good and benefit from its outcome. Thus, giving selfish agents the possibility to freely choose on their allocations leads to the underprovision of the public good and the arising of the "free rider issue". For this reason, social dilemmas have been central in the study of human behaviour since the origin of behavioural economics.

In order to conceal the free rider issue, many have been the mechanisms proposed, from which sanctioning has been the one that has acquired greater relevancy. Providing individuals the opportunity to sanction each other achieves, in general terms, a diminishment in deceitful behaviour and bridges the gap to the efficient outcome. However, in these terms, not any kind of implementation of a sanctioning system is enough in accomplishing such purpose. In this dissertation I present three research works, two theoretical and one experimental, which feature three different punishment structures which enhance cooperation in social dilemmas.

There are several sanctioning schemes that have been employed in theoretical and experimental settings. The trendsetter has been peer punishment, where at the end of the game, individuals observe what others have decided and are given the chance to sanction them at an individual level. Given that sanctioning is globally considered a costly action, selfish individuals would not implement punishment in finite games. However, society is not only conformed by selfish agents as, in fact, most of us have some kind of social concern such as inequity aversion, reciprocity or altruism. This makes that, even if costly, individuals rationally punish free-rider peers as a best response. Nevertheless, inefficient overpunishment is a frequent drawback of this kind of punishment. Moreover, many have emphasized the lack of realism in the possibility of implementing individual punishment.

As a response to the downside of peer punishment, sanctioning literature has proposed two alternative punishment schemes that I employ in this work: (i) pool punishment and (ii) coordinated punishment.

Pool punishment is a centralized sanctioning structure where a particular agent, preferably external to the development of the game, is endowed with sanctioning power to punish free riders. Furthermore, this agent or institution is not an automatic being but a strategic entity who must also be provided with the correct incentives to perform appropriately. This type of punishment overcomes the overpunishing outcome of peer punishment and the non-realism of decentralized sanctioning authorities. An example of pool punishment are tax collectors.

Coordinated punishment, on the other side, is a decentralized sanctioning system which proposes two appealing features not present in previous schemes. In first place, it highlights that sanctioning costs should be divided among the individuals willing to carry out the punishment. That is, they should present increasing returns to scale. In second place, for sanctioning to be effective it should require certain number of punishers. Otherwise, no payoff of the deceiver is destroyed. An example of coordinated punishment is a strike.

In Chapter 2 of this dissertation I present a literature review describing the free rider issue as well as the impact of relaxing the different assumptions of the standard public goods game model: wealth heterogeneity, productivity heterogeneity, information and social preferences. Moreover, I propose four mechanisms to address the issue, making particular emphasis on sanctioning. Finally, I examine the state of the art of various punishment schemes: peer punishment, counter punishment, coordinated punishment and pool punishment.

Chapter 3 is a theoretical work analysing the emergence and performance of a sanctioning institution in a public goods provision context. We present a society with different wealth levels who, by themselves, cannot achieve a positive provision of the public good. Such society, through a government representing the interest of a particular social class, must decide whether or not to implement a centralized sanctioning institution in line with the pool punishment literature. This work analyses under which conditions will a high-performance sanctioning institution be implemented, what is the level of public good provision achieved and, in last place, its efficiency.

This model considers selfish preferences with the aim of evaluating the worst possible scenario. If citizens of this society had some type of social concern, the results here presented would be boosted. For this reason, the methodology employed in this chapter is Game Theory, with the purpose of modeling the strategic interaction of economic agents and characterizing the Nash Equilibria of the game.

Chapter 4 is an experimental work exploring the impact of two different payoff schemes in a centralized sanctioning environment. In particular, I present a public goods game experiment with 336 subjects grouped into groups of four, where three of them are contributors and one of them is a sanctioner. The former only contribute to the public good while the latter uniquely carries out punishment actions, following a pool punishment scheme. The main question to approach is how should the implementation of punishment from these centralized institutions work, for which we compare two payoff schemes: (i) a fixed scheme where the sanctioner is provided certain level of endowment to decide on the punishment actions and (ii) a variable scheme, where instead he receives an endowment proportional to the level of cooperation attained. This work emphasizes the benefits in terms of contributions and efficiency of centralized punishment systems with fixed payoff schemes.

In this work, I make use of the Game Theory methodology to characterize the Nash Equilibria of the game, under the assumption of standard preferences. Furthermore, given that real-life decision makers are not always selfish, I use Experimental Economics to contrast these results. With this end, the Experimental Economics protocol has been carefully followed in the design and implementation of the sessions to guarantee the derived data was appropriate for analysis and the design has obtained the approval of the University of Valencia Ethical Committee.

In last place, Chapter 5 of this dissertation is a theoretical work comparing two different decentralized punishment schemes in a team trust game with information asymmetries: (i) an uncoordinated punishment system where individual punishment destroys the punished agent's payoff and (ii) a coordinated punishment scheme where it is necessary that the number of individuals willing to carry out punishment exceeds a particular threshold for the punishment to be effective. Results reveal that a coordinated punishment system leads to efficient equilibria in a wider range of cases than uncoordinated punishment when the proportion of reciprocators in the population of investors is sufficiently high.

Beyond the use of Game Theory in the characterization of the Nash Equilibrium of the game, this work introduces social preferences. In particular, it considers that the investors of the team trust game can either be selfish or reciprocal and that the allocator can either be profit maximiser or fair minded. By making use of Game Theory in general and, Behavioural Economics in particular, I characterize the game's Perfect Bayesian Equilibria of the game.

Chapter 2

Coordination concerns: concealing the free rider issue

2.1 Introduction

Coordination is a key element in most of our day-to-day interactions with other individuals. Think about the chief executive officer (CEO) who has to bring together different working units, the managers at each of those units organizing their teams, or the workers in each of those teams trying to work together with a common goal. But coordination is not only crucial during working hours, think about reaching an agreement at your neighborhood community about setting up a new elevator, renewing the contract to the maintenance staff or modernizing the almost torn-down façade. Remember having tried to organize a meeting with your classmates to reminisce and catch up or just think about how instruments synchronize in a sonata.

Regardless of it being a working or a social environment, coordination is pursued as a guarantee of efficiency: maximizing utility using the minimum resources for it. Recall the firm example, for instance. Any profit-oriented firm will try to get the most out of its profits with the least assets and productive factors possible, where time is one of the most valuable assets in a competitive context in which it is standard to see how rivals sprint to be original and inventive. In this setting, coordinated teams will work faster and will avoid duplication and shortages, common in teams with a lack of organization. Think about going to a restaurant and receiving your drinks twice, or not receiving them at all.

From a social perspective, if neighbors propose the modernization of that façade, it may be moved by their aesthete self but there is probably also a component of wanting to appreciate their property. Evidently, the upgrading should imply the minimum cost that indeed causes the expected revalue. In the field of game theory, the public goods game (PGG) has been the baseline to reproduce any of the situations previously described. This simple game, to be explained in the following section, clearly captures the importance of coordinated actions in terms of efficiency along with the associated issues that coordination rises: if coordination is so beneficial, why is it sometimes difficult to achieve? Intuitively, coordination is costly. Coordination requires effort, time and resources. And more importantly, given its social benefits, you cannot avoid that somebody that does not put in those ingredients takes advantage from the outcomes. Recall the elevator example formerly presented: you cannot prevent a neighbor who has not paid for the installation from using the elevator. The fact that coordination is costly and that its outcomes are non-excludable tempts selfish individuals to free ride from coordinating. Either way, they are going to take the elevator.

In the following section, we formally describe the PGG and the theoretical predictions for it, illustrating the free rider issue. Furthermore, we present a comparative statics analysis using recent scientific findings regarding the different elements that define a PGG. After that, we describe the four main mechanisms the literature has proposed to face the free riding matter. From these, we select sanctioning as the one with most potential and devote the last section to the detailed description of the different types of punishment schemes that can be implemented.

2.2 Coordination issue

The PGG, in its simplest form, is a 2x2 game where two players must simultaneously decide whether to contribute or not to contribute to a public good. The best outcome, where both players receive the highest payoffs is reached if both of them contribute. However, if a player believes the other one is going to contribute, then he receives a higher payoff by not contributing, given it is a costly action and the public good is going to be funded thanks to the other player's contribution. Finally, if none of them contribute, the public good's costs are not covered. In a normal form, a PGG would look like Table 1, where the row-player is Player 1, the column-player is Player 2 and C (contribute) and NC (not contribute) are the possible actions for each player. Each payoff cell contains the payoff for Player 1, the payoff for Player 2 for every possible combination of actions. Notice that when both players coordinate in contributing, they both receive a payoff of 2. However, if one of them contributes and the other one does not, the player who has contributed bears all the costs and is left with a payoff of 0, whereas the player who has free ridden by not contributing benefits from the public good without engaging in the costly action, i.e. he receives a payoff of 3. Finally, if none of them contribute, they both receive 1, which is a worse outcome than both of them contributing and earning 2. Both players have perfect information about the payoffs for each possible scenario.

	C	NC
C	$2,2$	$0,3$
NC	3,0	$1,1$

Table 2.1: Classic Public Goods Game

In game theory, the standard solution concept of a simultaneous game with perfect information is the Nash Equilibrium (NE), named after John Forbes Nash Jr. It states that a pair of actions is a NE if no player has profitable unilateral deviation from it. Assuming both players are rational and have selfish preferences, that is, they maximize their material payoff, the NE of this game is that both players choose not to contribute (NC, NC) receiving a payoff of 1 each. One could think that the solution of this game is that both players contribute to the public good, as they are both better off than not contributing (2>1). However, notice that if any player believes the other player is going to contribute, they have incentives to free ride by not contributing and make a payoff of 3 instead of 2. Therefore, (C,C) cannot be a NE. However, if both players are intelligent, they can both apply this reasoning, such that they both end up in (NC, NC) with a payoff of 1 each. This pair of actions is a NE because no player has incentives to deviate to contribution and make a payoff of 0.

We can generalize this simple game to a broader situation that can be applied to, for instance, a firm facing this social problem. Let's consider a group of n workers who receive an endowment in effort. From this endowment, they must decide how much effort to keep for their own interest and how much to destine to the team. The sum of all of the efforts the workers invest in the firm project is then multiplied by a multiplier and equally divided among all the workers, regardless of their contribution. The material payoffs of any player would be given by equation 2.1, where ω is the endowment in effort every subject receives, g_i is the individual contribution of subject i and α is the marginal per capita return of the project.

$$\pi_i = \omega - g_i + \alpha \sum_{i=1}^{n} g_i \tag{2.1}$$

The NE in this game is that every individual chooses $g_i = 0$, even though the social optimum is reached if $g_i = \omega$. For this to be true, the marginal per capital return (MPCR) must be $\alpha > 1$.

Consider, for instance, a task where a working team of 3 salesmen must work jointly to reach a particular goal in sales. Every salesman will get an increase in their salary, proportional to the aggregate sales achieved by the group, which makes the goal beneficial for all of them. However, achieving that goal implies a substantial level of effort. The social optimum would be reached if all of them cooperated to achieve that task, attaining the maximum salary increase possible. However, they are benefitting from the aggregate sales, that is, from the same social project. Thus, any of them could decide to free ride on effort and take advantage of the sales the other two members achieve. However, if the 3 of them think in the same way, nobody would devote any effort and there would be no salary increase.

Despite these predictions, experimental evidence has repeatedly and extensively proven deviations from the NE. In particular, positive levels of cooperation to public goods are usually achieved. Subjects contribute with some amount between 40% and 60% of their endowment. In the following, and before moving on to how can the coordination issue be concealed, let's see what has recent experimental evidence proven about variations in the presented setup. There are four elements that we consider homogenous for all of the subjects in this standard game: endowment, MPCR, information and preferences. In other words, we consider all subjects have the same effort capabilities, that the return of the public good is common from everybody, that all of them are provided the same information about the project and that they all want to maximize their material payoff. In this section, we aim to see how do the violation of these homogeneity assumptions change the outcome of the game in an experimental setting.

2.2.1 Wealth heterogeneity

Wealth homogeneity is one of the first assumptions that raises suspicions. Considering equally rich societies is an unrealistic and rather utopian assumption. If we accept that we have different levels of wealth but that we are able to group ourselves into homogeneous groups, there are significant differences in what low-endowment and high-endowment groups contribute to public goods. In particular, low-income groups tend to over-contribute while high-income groups under-contribute. In other words, the proportion of their endowment that poorer groups destine to public goods is higher than the proportion that richer ones destine (Chan, Mestelman, Moir and Muller, 1996, 1999; Buckley and Croson, 2006; Reap, Ramalingam and Stoddard, 2016). Additionally, as it has been proven in Cherry, Kroll and Shogren (2006), this result is robust to the origin of the endowment. In other words, it doesn?t matter whether subjects have had to work for that endowment and it is in fact an income, or whether it has just been given to them effortlessly as wealth. In either case, those whose endowment is lower contribute to a larger extent than those whose endowment is higher.

Nevertheless, one could argue that we usually face coordination issues in heterogeneous groups. In other words, we sometimes are not able to classify ourselves into low and high-wealth groups and we in fact belong to unequal communities. In this case, heterogeneous groups contribute less than homogeneous groups (Cherry, Kroll and Shogren, 2006). This accounts for an inequality issue, where having variety could be detrimental for group performance.

These results highlight the importance of working in homogeneous groups. Going back to the salesman example, perhaps not all of them have the same time availability or the same effort capability. If we take these features as given, the unit manager should try and form homogeneous groups according to the workers? characteristics in order to maximize the total sales.

2.2.2 Productivity heterogeneity

The second dubious assumption is the fact that the public good's productivity, captured by the MPCR, is common for everybody. This implies that everybody values the public good in the same way and, therefore, obtains the same return from it. If we consider a neighborhood community discussing about the elevator, it is comprehensible that the return that somebody living on the first floor receives from having an elevator is not the same as the return of somebody living on the last floor. If, instead, we consider a working team incentivized with a salary increase, some of them could argue that devoting that extra effort is too time consuming and that they prefer to spend that time with their families rather than earning more money. Furthermore, personal circumstances affect our daily attitude, concentration and productivity at work, and they do not affect all of us equally.

In this line, literature has demonstrated that when endowed with different productivity levels, low-MPCR subjects contribute less than high-MPCR subjects. This holds even in heterogeneous groups with different productivity levels (Fellner, Iida, Kröger and Seki, 2014). However, heterogeneous groups contribute less than homogeneous ones, analogously to the case of wealth heterogeneity. Moreover, as Kölle (2015) stresses, these lower contributions are not a consequence of the heterogeneity itself, but of the nature of such asymmetry.

This implies that teams should share interests, motivations and goals.

Likewise, different team-performance related bonuses should be avoided among identical workers. This way, coordination will be higher and so will efficiency.

Finally, let us make a brief comment about productivity related to group size. The MPCR, which we saw in the model as ?, is in fact the result of a multiplier representing each individual's valuation of the public good divided by the number of individuals among which the public good is going to be shared. Possibly, what we expect is that increasing group size reduces cooperation given that it requires a higher degree of coordination. As shown by Isaac and Walker (1988), if the larger group size entails a decline of the MPCR, the effect will indeed be negative on cooperation. Nonetheless, for the same MPCR, large groups contribute more, on average, to public goods than smaller ones (Barcelo and Capraro, 2015; Isaac, Walker and Williams, 1994), despite the potential coordination issues. This positive effect is called group-size effect and rules out the common belief of small groups being superior.

An example of this could be a logistics manager coordinating different working divisions of a supply chain. Most of the times, firms commit to hand in the final product before a particular date. In this case, where coordination is fundamental to meet such deadline, the logistics manager should not be afraid of working with large groups in each of the chain links, as long as their productivity is similar. They will coordinate more to complete their part of the process such that the customer has his product at the right time.

2.2.3 Information

The classic PGG and its solution concept assumes information is perfect and symmetric. In other words, everybody has costless access to the details about the game's evolution and outcomes and this information is the same for everybody. In our daily interactions, however, we sometimes fall short of information. An aspect in which literature has focused in during the last decades, is the feedback subjects receive after playing the PGG and how they receive such information. In this line, knowing what peers have contributed increases contributions in future rounds, effect that is detrimental if instead of knowing the level of contributions, they are informed about the earnings (Sell and Wilson, 1991; Bigoni and Suetens, 2012; Nikiforakis 2010). Other factors that have turned up to increase contributions are providing feedback about virtuous behavior in the group, that is, the higher group contributions (Faillo, Grieco and Zarri, 2013) or identity revealing (Rege and Telle, 2004).

Behavioral economics has also spotted that the way in which information is disclosed also causes a significant impact in decision-making. This result is called framing effect and is widely used in behavioral economics, especially for marketing and public policy purposes. Regarding this, (Cookson, 2000) carries out a meta-analysis of framing effects on PGG. This study claims that if the PGG is played in phases of several rounds each after which they receive a results summary, a re-start effect is triggered increasing the contributions at the beginning of the next phase. Moreover, subjects contribute more when the public good payoffs are presented in terms of gifts instead of private and public investments. Finally, comprehension tasks significantly enhance cooperation.

Information is therefore a potential tool for coordination issues. In working or social situations where coordination is required, transparency is always going to improve cooperation. These findings point out how information develops trust and how trust increases cooperation towards a common objective. In this line, many firms choose to make their workers more conscious of the whole process they are involved in. Likewise, many researchers ease their survey participants a copy of the final research outcome they have participated in.

2.2.4 Social preferences

The last underlying assumption of the standard PGG are the preferences individuals have. Following the model, we are purely selfish individuals, only concerned about the material payoffs of our actions. This also entails that we are always absolutely capable of measuring and balancing monetary costs and benefits associated to each possible action. Some game theorists have criticized this consideration and have introduced concepts inherent to behavioral economics. In particular, ideas that have to do with social preferences.

The most well-known social preference model is the inequality aversion model (Fehr and Schmidt, 1999). This model states that only a proportion of the population has selfish preferences, while the rest of the individuals dislike inequitable outcomes. If their material payoff is lower than their peers', they suffer a disutility proportional to the distance between the payoffs (disadvantageous inequality). Additionally, they also experience disutility if their payoff is higher that their peers' (advantageous inequality). Nevertheless, the first disutility is stronger than the second one: you don't like being worse off, you don't like being better off but if you had to choose, you would prefer to be better off. Individuals with inequality aversion will contribute more than what the NE for selfish individuals predicts if this can reduce the inequality between them. This model was proposed as an explanation to the cooperative behavior constantly observed in laboratory experiments and has proven to be fairly explanatory for most of them.

A social preference alternative to explain human behavior is reciprocity. Reciprocal agents are friendly to friendly peers and hostile to hostile peers (Fehr and Gächter, 2000). Notice reciprocity is, therefore, positive and negative. It is a tit-for-tat, an eye for an eye. Many supermarkets expect reciprocity by offering free samples or discounts of their products. Moreover, in our social relationships, we are usually willing to return favors to those who have been kind with us at some point in time.

Finally, the most endpoint case of social preferences is altruism. Altruism or selflessness is the complete opposite of selfishness: instead of maximizing your own material payoff, you maximize the welfare of others. This kind of preference is harder to see at a working level but is commonly used to describe paternal love.

Social preferences are important because they exist in social relationships and actually explain why we behave as we do. In predesigned settings, like working environments, social concerns are fundamental in team-working. If the manager can design working teams, he should have deep knowledge of each person's ethics when working together. An individual with high disadvantageous inequality concerns could feel highly frustrated if working with purely selfish colleagues.

2.3 Mechanisms to address the coordination issue

At this point, the reader should understand the relevance and problematics of coordination as well as the effects that variations in the basic assumptions of the model have. These variations, however, are usually endogenous features of the game rather than aspects we can influence on. In this section we present exogenous mechanisms that change the game's rules pursuing an increase in coordination and, consequently, in efficiency.

2.3.1 Reputation

Up until now, predictions have been made for games where interactions are unique, rather than prolonged over time. If we think about a working team, a social event with our family and friends or any kind of community we belong to, it is reasonable to assume that we will meet that people in the future and we may face similar situations where coordination becomes crucial again.

In this respect, game theory makes a clear distinction between games that are played only once (one-shot games), and games that are played for several periods of time (repeated games). Repeated games, at the same time, can also be divided into finitely repeated games and infinitely repeated games. If a relationship is maintained over a predetermined period of time (finitely repeated games), the theoretical prediction for the PGG is the same as the one of the one-shot game. Notice that if individuals cooperate in this context it is because they expect to maintain this friendly relationship in the future by building a reputation. However, if there is a last period, a last day, a last meeting, a last task, there are no incentives to be friendly for tomorrow. Would you strive in your last day of work? This phenomenon is named the end-of-the-world effect, where cooperation drastically falls in the last period. Now, if there is going to be full free riding in the last day, your incentives to cooperate in the second-to-last day disappear, and so do your incentives in the third-to-last day, ... and so do your incentives in the first period.

The workaround for this is for there to be no last day or alternatively (given vital restrictions), that individuals don't know when the game is going to end. Using game theory terminology, the game has an infinite horizon or there is a positive probability of the game ending. In this case, positive levels of cooperation can effectively be achieved.

In experiments carried out in the laboratory, the existence of repeated interactions is common and is combined with the social concerns inherent to each subject. This combination leads to positive levels of cooperation that decay as time goes by, leading to an inverse U shape (see Figure 1). If the number of periods they are going to interact is known, an end-of-the-world effect is always noticeable. This implies that reputation is necessary but it is not sufficient as a mechanism to sustain cooperation over time.

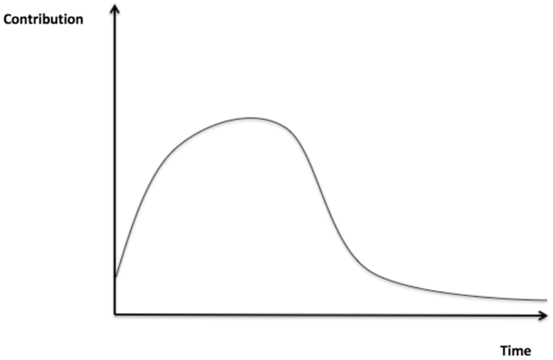

Figure 2.1: Standard result of public goods game experiments

2.3.2 Step-level PGG

Another basic rule that can be changed to enhance cooperation is setting a particular threshold level necessary for the public good to be shared (Rapoport, 1988). In other words, unless a minimum level of aggregate contribution is achieved, there are no public-good advantages. For instance, recall the salesmen example seeking for a salary increase of Section 2. The unit manager could ask for a minimum sales target necessary for any salary increase to happen. If that threshold is sufficiently high, and the maximum effort of every member is necessary to achieve it, free riding is no longer attractive for them.

Obviously, a guaranteed success would be to set the threshold at maximum contribution such that there are no advantages on free riding. If we talk about an investment, this is feasible, as money is quantifiable. However, cases requiring human effort are more difficult to measure and compute.

Targets are natural in firm environments, where employees must meet a monthly, weekly or even daily goal. However, occasionally, these targets are settled too low such that part of the staff can achieve them by themselves without the need of high levels of coordination. Thus, it may incentivize the over-effort of some employees given that it is a take it or leave it approach. The definition of the correct targets is therefore essential in concealing the free rider problem.

2.3.3 Communication

The standard PGG is based on the premise of individuals deciding independently and simultaneously on their cooperation to the public good. However, it is ordinary to see, especially in working environments, how coworkers communicate between themselves. This communication opportunity has been repeatedly ascertained to increase the level of cooperation in PGG. Multiple communication mechanisms have been tested in the laboratory, such as non-binding face-to-face communication, audiovisual conferences, audio communication or e-mail communication, among others (Isaac and Walker, 1988; Brosig, Weimann and Ockenfels, 2003; Frohlich and Oppenheimer, 1998); being face-to-face communication the most efficient mechanism. Nevertheless, this is not due to the loss of anonymity: verbal communication through an anonymous chat room has been demonstrated to be almost as efficient (Bochet, Page and Putterman, 2006).

This result pinpoints the importance of enhancing a friendly environment where communication flows as part of a firm's corporate culture. In this line, many companies implement regular informal meetings or outdoor activities as part of their staff's routine for employee engagement. Furthermore, in order for it to be as binding as possible, the barriers between the interlocutors should be minimized.

2.3.4 Sanctioning

The most common mechanism to conceal the free rider issue and which counts with a vast theoretical and experimental literature is sanctioning. Sanctioning can be understood in many different ways: formal economic punishment in the form of a penalty, social punishment related to hostility or even breaking bonds in the working or personal domain. It is used with the purpose of smoothening the contribution nosedive in repeated PGG and, for some cases, revert it.

According to purely selfish preferences, if punishment is costly, nobody should engage in such action. However, as social agents that we are, we do implement punishment, even in one-shot situations. The next question we propose is related to how is punishment, as a matter of fact, implemented. Should coworkers have the power to sanction each other? Should there be a responsible in charge of doing so? Can coworkers then return the hostile behavior somehow? Should there be certain level of agreement in a punishment decision?

The following section in this chapter tackles this issue by presenting different types of sanctioning schemes.

2.4 Sanctioning

2.4.1 Peer punishment

Peer punishment is the most standard way of implementing a mechanism to address the coordination dilemma. 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. This would be comparable to endowing coworkers the possibility of punishing each other at the end of the day. Notice that this type of punishment is a public good itself, as everybody is better off if free riders are sanctioned, but they prefer the rest to undertake the cost of doing so.

Consider, as an example of a social setup, a group of friends meeting for dinner, where each one of them is expected to bring a dish and a beverage so that there is a variety of food and drink for dinner. If somebody free rides from their part of the contribution, but indeed benefits from what others have prepared, he could possibly not be invited again by his friends for future dinner parties as a form of social punishment. In a working scenario, coworkers could ostracize employees that free ride on effort exertion after an important task carried out by the team. At an economic level, that free-riding employee could be penalized by the firm in terms of salary or even fired.

Experimental works have extensively proven that a combination of peer punishment, social preferences and long-term interactions lead to higher contributions. The key result in this field is that peer punishment can indeed raise contributions to levels above those attainable in the absence of such punishments (Fehr and Gächter, 2000). Furthermore, these improvements in terms of efficiency are also valued by individuals, who, if allowed to choose between a sanctioning environment or a sanction-free environment, establish themselves in the former one after a learning process (Gürerk, Irlenbusch and Rockenbach, 2006). Regarding the long-run effects, contributions reach significantly higher levels the greater the number of periods subjects interact (Gächter, Renner and Sefton, 2008).

Many authors suggest that the ability of costly punishments to sustain high contributions to the public good depends crucially on the effectiveness of that punishment, i.e., the factor by which each punishment point reduces the recipient's payoff (Nikiforakis and Normann, 2008). According to the seminal work in this area (Fehr and Gächter, 2000), the cost of peer punishment should follow an exponential trend. For low levels of punishment, it should be a 1-1 relationship, but as the impact of punishment increases, this relation becomes a 3-1, that is, the cost the punisher bears is thrice the impact the punished undertakes. Other studies, however, argue that for punishment to make a difference, it must inflict a penalty that is substantially higher than the cost of meting out that punishment. In particular, they assert that the only punishment treatment that succeeds in sustaining cooperation over time is the low cost-high impact treatment (Egas and

Riedl, 2008). In particular, following Casari (2005), the cost-effectiveness ratio should be no less than 1-3 (Casari, 2005). That is in fact the inverse of the seminal punishment model in (Fehr and Gächter, 2000): for every unit of utility that is deducted from the punisher's payoff, the punished individual should have his utility reduced in 3 units.

What we pick up from this is that peer punishment has the power of smoothening the coordination issue as long as the cost-effectiveness ratio is suitable and relations are maintained over time. Regarding the dinner party, the free rider will have higher incentives to bring a dish if they have scheduled more dinner parties for the next months and he identifies the risk of not being invited anymore.

2.4.2 Counter punishment

Counter punishment, also known as perverse punishment, is a second-round punishment phase, where sanctioned free riders can penalize their punishers back. If one allows the possibility of counter punishment by punished free riders, cooperators will be less willing to punish in first instance (Nikiforakis, 2008). While with peer punishment, contributors use punishment as a signal of not accepting low contributions in the future, counter punishment is used to strategically signal that future sanctions will not be tolerated. This way, peer punishment is reduced and contributions show a decaying pattern. However, counter punishment also has its bright side if used in a good way. On the one hand, it can be used to sanction those who fail to sanction free riders, in other words, those who have free ridden on punishment. On the other hand, it can also be used to penalize those who have exerted coercive punishment by sanctioning high contributions (Denant-Boemont, Masclet and Noussair, 2007). However, fairness concerns are necessary for this type of punishment to be used in this way.

At a company setting, counter punishment could occur if a group of coworkers believed that somebody is being too hostile with the novice who did not rise to the challenge at a first attempt. This way, they could also decide to exclude the punitive coworker when organizing the next outdoor activity.

2.4.3 Coordinated punishment

Individual effective punishment is sometimes not very truthful. In real life, it is usual that a certain number of individuals are needed to effectively sanction opportunistic behavior. Everyday examples of this condition are worker strikes, a state coup or any kind of boycott. In this sense, coordinated punishment is implemented in the following way: at the end of the game, players individually decide whether to punish or not to punish opportunists forwarding that if they succeed in reaching a threshold in the number of punishers, the damage inflicted can be very large and the individual cost of coordinated punishment can be relatively low.

Following this approach, coordinated punishment can be effective if the threshold that must be reached is sufficiently high. According to (Casari and Luini, 2009), coordinated punishment performs remarkably better than peer punishment when the requirement to punish a person is the emergence of a coalition of at least 40% of the group members. Authors associate the effectiveness of coordinated punishment with its ability to censor coercive punishment of the higher contributors, which, as a matter of fact, was relatively frequent in their experiment.

Coordinated punishment has also been revealed to be effective in other kind of social dilemmas like team trust games, also called team investment games. The common method of this kind of games is as follows. Subjects are assorted into groups of 3, from which 2 are named assigned the role of investors and 1 is assigned the role of allocator. In the first stage of the game, the 2 investors must decide whether to invest or not in a common project, which is only successful if both of them invest. Such project generates a surplus, from which the allocator decides how much to return to each investor and how much to keep for himself, in the second stage of the game. A punishment stage could be added to this standard team trust game in next place. If this punishment scheme follows the basics of coordinated punishment such that both investors must coordinate to sanction the allocator for there to be effective punishment, cooperation can be maintained (Calabuig and Olcina, 2015).

Around us, there are many situations where coordinated punishment occurs, situations where certain level of agreement must be attained for punishment to be effective. Think about any type of social community, where one of the members has misbehaved and the community is considering expelling the mischievous member. In this case, communities frequently undertake some kind of voting procedure for such decision.

2.4.4 Pool punishment

There are numerous situations where punishment is not individually decided once the outcomes are observed, but must be agreed before the game even starts. This way, individuals commit to punishment actions and there is no place for any kind of renegotiation of the conditions or of backing down. This reflects how investments in monitoring and sanctioning institutions to uphold the common interest are made.

In this line, different studies have explored how individuals indeed implement institutions of this type, if offered such possibility. The credible threat of this institution sanctioning opportunistic behavior at the end of the day enhances individual cooperation, which in turn, has positive effects on group cooperation (Kosfeld, Okada and Riedl, 2009; Ozono, Jin, Watabe and Shimizu, 2016). This effect is even more pronounced with the option of counter punishment (Traulsen, Röhl and Milinski, 2012).

In the last years, the comparison between peer and pool punishment has caught attention. With the purpose of overcoming difficulties and inefficiencies related to individual punishment (like the coercive punishment of high contributors we saw before), groups have continuously developed forms of self-regulation, where sanctioning is delegated to a central authority (Baldassarri and Grossman, 2011; Fehr and Williams, 2017). Examples are specialized law forces such as the police, courts, state and non-state institutions. Hence, it could be said that it is the punishment option preferred by individuals (Traulsen, Röhl and Milinski, 2012; Sigmund, De Silva, Traulsen and Hauert, 2010).

At a firm level, the organizational hierarchy limits the sanctioning power. Besides all the examples we have provided about social punishment between coworkers, actual penalizing decisions come from higher bodies. A coworker can never fire you, a CEO can. In this sense, the commitment of the application of sanctioning is usually regulated by a series of protocols and internal regulation specifying the consequences of unruly behavior in detriment of the firm. At a societal level, the same applies; you cannot economically sanction your neighbor for tax payment default or illegal parking. The most you can do is to report it to the relevant authorities for them to make use of their power.

The objective of all of these pre-designed rules that surround us is exactly to increase cooperation and avoid free riding. If the threat that we are going to be certainly caught and punished were large enough, prisons would be empty.

2.5 Conclusions

When we interact with other people we constantly face coordination dilemmas: at our neighborhoods, with our families, with our friends or with the people we work with. We should all put the best of ourselves so that everything works properly but there is always somebody who decides to free ride on others' money or effort. Think about a supply chain selling a product you want to buy through different channels. You can go to a local retailer to see the product, obtain information about it or even test it. However, when you get home, you are going to buy it online. The online supplier is indirectly benefiting from the service of the local retailer. We are all free riders at some point.

However, this opportunistic response is not associated to any kind of particular mischief, it is just a selfish reaction to a cooperative situation with a non-excludable outcome. Who is willing to organize the next family trip? Fortunately, the world population is not composed uniquely by selfish individuals who look the other way, most of us have social concerns for inequality, reciprocity or even altruism. This conditions how we behave in all of the described situations and brings to the surface at least someone willing to cooperate by taking the lead in planning the next trip.

Nonetheless, this is not enough. If we want to conceal the free rider problem and enhance further cooperation, we should try and form groups of people that share the same capabilities, interests, motivations and ethics. Relationships should be maintained for as long as possible so that there is a better tomorrow for which everybody wants to fight today. Considering the multichannel supply chain example presented before, if you trust your local retailer, you could prefer to buy the product to him than to the unknown online supplier. Additionally, a sanctioning mechanism would also be helpful.

Punishment opportunities are present in most of the interactions we talk about. Punishing must not necessarily be an economic action, it can just be a social response of hostility, ostracism or bond breaking. Generally speaking, if any sort of sanctioning is at reach for the group members, cooperation significantly increases. In more detail, punishment can either be an individual decentralized decision or it can be a power endowed to a centralized authority, like a government. For interactions involving numerous agents, the establishment of hierarchies with different responsibility levels is a more feasible way of ensuring collective cooperation. But we don't need to go to massive populations to find such hierarchies: any task that requires team working at any small-scale enterprise will already need a "good" manager.

References

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

2. Barcelo, H., & Capraro, V. (2015). Group size effect on cooperation in one-shot social dilemmas. *Scientific Reports*, 5, 7937.

3. Bigoni, M., & Suetens, S. (2012). Feedback and dynamics in public good experiments. *Journal of Economic Behavior & Organization*, 82(1), 86-95.

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

5. Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science,* 328(5978), 617-620.

6. Brosig, J., Weimann, J., & Ockenfels, A. (2003). The effect of communication media on cooperation. *German Economic Review*, 4(2), 217-241.

7. 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.

8. Calabuig, V. & Olcina, G. (2015). Coordinated punishment and the evolution of cooperation. *Journal of Public Economic Theory*, 17(2),

147-173.

9. Casari, M., & Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization,* 71(2), 273-282.

10. Casari, M. (2005). On the design of peer punishment experiments. *Experimental Economics,* 8(2), 107-115.

11. Chan, K. S., Mestelman, S., Moir, R., & Muller, R. A. (1999). Heterogeneity and the voluntary provision of public goods. *Experimental Economics*, 2(1), 5-30.

12. Chan, K. S., Mestelman, S., Moir, R., & Muller, R. A. (1996). The voluntary provision of public goods under varying income distributions. *Canadian Journal of Economics,* 54-69.

13. 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.

14. Cookson, R. (2000). Framing effects in public goods experiments. *Experimental Economics,* 3(1), 55-79.

15. Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic theory,* 33(1), 145-167.

16. Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society of London B: Biological Sciences,* 275(1637), 871-878.

17. Faillo, M., Grieco, D., & Zarri, L. (2013). Legitimate punishment, feedback, and the enforcement of cooperation. *Games and economic behavior,* 77(1), 271-283

18. Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives,* 14(3), 159-181

19. Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics,* 114(3), 817-868.

20. Fehr, E., & Williams, T. (2017). Creating an efficient culture of cooperation. *Mimeo.*

21. Fellner, G., Iida, Y., Kröger, S., & Seki, E. (2011). Heterogeneous productivity in voluntary public good provision: an experimental analysis. *IZA Discussion Paper No. 5556. 2014.*

22. Frohlich, N., & Oppenheimer, J. (1998). Some consequences of e-mail vs. face-to-face communication in experiment. *Journal of Economic Behavior & Organization*, 35(3), 389-403.

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

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

25. Heap, S. P. H., Ramalingam, A., & Stoddard, B. V. (2016). Endowment inequality in public goods games: A re-examination. *Economics Letters*, 146, 4-7.

26. Isaac, R. M., Walker, J. M., & Williams, A. W. (1994). Group size and the voluntary provision of public goods: Experimental evidence utilizing large groups. *Journal of public Economics,* 54(1), 1-36.

27. Isaac, R. M., & Walker, J. M. (1988). Communication and free?riding behavior: The voluntary contribution mechanism. *Economic inquiry,* 26(4), 585-608.

28. Kölle, F. (2015). Heterogeneity and cooperation: The role of capability and valuation on public goods provision. *Journal of Economic Behavior & Organization,* 109, 120-134.

29. Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review,* 99(4), 1335-55.

30. Nikiforakis, N., & Normann, H. T. (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics,* 11(4), 358-369.

31. Nikiforakis, N. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior*, 68(2), 689-702.

32. Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics*, 92(1-2), 91-112.

33. Ozono, H., Jin, N., Watabe, M., & Shimizu, K. (2016). Solving the second-order free rider problem in a public goods game: An experiment using a leader support system. *Scientific reports,* 6, 38349.

34. Rapoport, A. (1988). Provision of step-level public goods: Effects of

inequality in resources. *Journal of Personality and Social Psychology*, 54(3), 432.

35. Rege, M., & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of public Economics*, 88(7), 1625-1644.

36. Sell, J., & Wilson, R. K. (1991). Levels of information and contributions to public goods. *Social Forces,* 70(1), 107-124.

37. Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, 466(7308), 861.

38. Traulsen, A., Röhl, T., & Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B,* rspb20120937.

Chapter 3

On the emergence of sanctioning institutions

3.1 Introduction

This paper studies the emergence and performance of a sanctioning institution in a public goods provision context. Action takes place in a group where there is wealth heterogeneity and decisions on the provision of incentives for the enforcement institution are made by a government representing the interests of a particular social class. We do not analyse the collective decision problem faced by the group on the creation or not of the institution but, instead, we focus on and compare the decisions made by governments representing different political decisive agents. We analyse under which conditions will a high-performance sanctioning institution be implemented in the different cases, what is the level of public good provision achieved and its efficiency from the social welfare point of view.

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 (Fehr and Gächter, 2000). 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, peer punishment causes high collateral damage when individuals interact over prolonged periods of time as the costs surpass the possible gains in cooperation (Gächter, Renner and Sefton, 2008). Additionally, formal individual punishment is hard to implement, specially in large groups. The reason for this is that bilateral punishment becomes more infrequent as groups increase in size and its potential future gains cannot be internalized (Greif, 1993).

In modern societies, transgressions 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, Haigner and Kocher, 2010). These sanctioning institutions are governed 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 behaviour is

[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).

subject to moral hazard problems[2] and, therefore, their performance will depend crucially on the incentives provided by the society represented by the government. But in a society with a heterogeneous distribution of wealth, the government represents the political decisive social class. If everybody in the group has the same level of wealth this makes no difference, but under the presence of wealth heterogeneity the incentives provided by different political decisive agents might be quite different. Our main interest in this paper is to study how does wealth heterogeneity and the identity of the political decisive agent affect the emergence and performance of sanctioning institutions with a moral hazard problem.

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. Without loss of generality, we will assume individuals either belong to a poor class, a middle class or a rich class. 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 our model, if the sanctioning institution is not implemented by the government or it is done with inappropriate incentives for high effort, there

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

would be no contribution at all. Only if a high-effort or high-performance institution emerges, the public good provision can be positive.

We first show that under a high-performance institution optimal individual contributions depend on the level of wealth. At an individual basis, the higher an individual's wealth is, the lower will his incentives to contribute under the threat of a sanctioning institution be.[4]

The incentives to contribute or to free-ride of the different social classes are determined by the social value of the public good and by the parameters that characterize the punishing technology such as the probability of fraud detection under high effort and the fine imposed on free riders. Therefore, total provision of the public good will depend on the interplay between the quality of institutional and technological variables and the features of the wealth distribution (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 show that the sanctioning institution will be implemented as long as the individual return of the public good exceeds the individual cost of the institution (salary paid to the sanctioner) 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

[4] This contrasts with the result obtained in the literature regarding step-level public goods, where wealthier individuals have stronger incentives to contribute (Rapoport, 1988)

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 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 implement it in more cases 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 the institution in this case, a middle or a rich-class government may not do so. This, perhaps puzzling, result is due to the wealth inequality among contributors.

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.

We also run some comparative statics, analysing the effects of variations of the different parameters on the implementation of a high-performance sanctioning institution. We show the results produced by changes in technological and institutional variables and also by changes in the wealth population distribution, both due to exogenous shocks.

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 and Riedl 2013).

3.1.1 Related literature

This paper is mainly related with previous literature dealing with pool punishment. 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, Hauert and Traulsen (2011) modelled 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, Okada and Riedl (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 analyse the emergence and performance of sanctioning institutions when individuals are free to migrate between different institutions. Acemoglu and Wolitzky (2015) approach the problem by endogenizing 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 assymetric information, characteristics considered in this study.

Furthermore, the standard public goods game literature has 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 has already shown that wealth heterogeneity affects the provision of a public good. Reuben and Riedl (2013), for instance, approach this problem from a social norm point of view. Their experimental findings show how individuals can overcome the collective action problem by agreeing on and enforcing a contribution norm, even in heterogeneous groups. Cherry, Kroll and Schrogen (2005), for instance, show that groups with heterogeneous levels of wealth contribute less than homogeneous ones. Burlando and Guala (2005) confirm this and additionally prove that when groups are formed by reciprocators, this effect is enhanced. In our work, we study the individual incentives under different wealth levels. In this line, Buckley and Croson (2006) demonstrate that less wealthy individuals contribute with a higher percentage of their endowment that wealthier ones. Finally, Heap, Ramalingam and Stoddard (2016) indicate that the adverse effect of inequality arises because individuals with higher levels of wealth reduce their contributions when they belong to unequal environments.

In this paper we also show that 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.

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 have not 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 Section 5, 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.

3.2 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 $\omega_i^j \in [\underline{\omega}, \overline{\omega}]$ (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:

$$\pi_i^j(g) = \omega_i^j - g_i^j + \frac{\lambda}{n}[\sum_{i=1}^{n} g_i] \tag{3.1}$$

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.[5] 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 their full wealth to the public good. In fact, the strategy profile $g_i = \omega_i$ $\forall_i$ is welfare maximizing.

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

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 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. Hereinafter, we will refer to this agent as a sheriff or as a sanctioning institution indifferently. 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 offence 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 re-elected in future elections) and so maximizes the equilibrium utility of the decisive political agent that can guarantee a majority of votes or its permanence in power. Let's denote ω^* as the wealth of this political decisive agent.

However, notice that we will not analyse how the political decisive agent 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 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. 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. 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, $f > 0$. This fine is a dissipative cost, that is, it is an amount of the citizens' income that gets destroyed.

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

$$\pi_i^j(g, \{s\}, p_e) = \omega_i^j - g_i^j + \frac{\lambda}{n}[\sum_{i=1}^{n} g_i] - \gamma_e \tag{3.2}$$

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 public goods game where every player knows the course of the game in previous stages. In the following, we will characterize the set of Subgame Perfect Equilibria of the game allowing for moral hazard ($0 < p_L < p_H < 1$).

3.3 Performance of the sanctioning institution: the level of public good provision

In this section we analyse 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, 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 analyse 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. 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:

$$\frac{p_H s_p + (1 - p_H) s_{np} - s_0}{n} + p_H f \geq \omega_i^j (1 - \frac{\lambda}{n}) \tag{3.3}$$

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 (represented by the term $p_H f$).

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

$$p_H f \geq \omega_i^j (1 - \frac{\lambda}{n}) \tag{3.4}$$

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 $\hat{\omega} < \omega^j \leq \tilde{\omega}$:
  - $BR^j(g) = \omega^j$ 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 from social class j 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 from social class j 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 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. These equilibria will depend on the existing wealth distribution in the group. Although we leave the details of the proof for the appendix (see appendix 3), 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.

There will also be a unique Nash Equilibrium whenever there is a proportion of free riders in the population. For instance, if the population is composed by unconditional free riders and conditional contributors ($\hat{\omega} \leq \omega^P \leq \tilde{\omega} \leq \omega^R$), then using successive elimination of dominated actions, conditional contributors 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 contributors 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 ($\omega^R \leq \tilde{\omega}$), 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. 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 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.

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 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 $\omega^P \leq \hat{\omega} \leq \omega^M$, 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 appendix 3.

Notice that according to this proposition, the higher social class j's wealth is, the less incentives it will have to contribute, provided that net gains of free riding increase with the level of wealth.

Recall 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, that is by the sum of the individual contributions.

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

Corollary 1*: Given a fixed population size and wealth distribution and a high-effort sanctioning institution, the amount of public good provision will be non-decreasing on the social return of the public good* λ*, on the probability of fraud detection under high effort* p_H *and on the fine* f*.*

Intuitively, if a society assigns a greater value to the provision of the public good 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* (Weingast, Shepsle and Johnsen 1981).

3.4 When does a centralized sanctioning institution emerge?

After characterizing the provision of public good under a high-performance 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.

3.4.1 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. 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 (see appendix 3).

Lemma 2: *Assuming the government has all the bargaining power, the* *contracts offered 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:* $\{s^L\} = \{s_0 = 0, s_p = 0, s_{np} = 0\}$.

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 $(\frac{p_L}{p_H - p_L}c)$ due to the existence of moral hazard and limited liability.[6] 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 political decisive player 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:

$$\pi^*(\{s^L\}, g^* = \omega^*) = \omega^* - p_L f \tag{3.5}$$

Recall that without the sheriff, free riding is the unique Nash Equilibrium, i.e. $\pi^* = \omega^*$. Consequently, offering a low-effort contract is always

[6] 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 $\frac{p_H}{p_H - p_L}c = c + \frac{p_L}{p_H - p_L}c$. Thus, $\frac{p_L}{p_H - p_L}c$ are the economic rents.

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-performance institution or not.

3.4.2 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 3: *Assume that $\omega^P \leq \hat{\omega}$ and the political decisive agent has wealth ω^*. The sanctioning institution will emerge if and only if:*

$$\frac{\lambda}{n}[\sum_{\omega^j \leq \hat{\omega}} q^j \omega^j] \geq \frac{p_H s_k}{n} + \begin{cases} \omega^* & \text{if } \omega^* \leq \hat{\omega} \\ p_H f & \text{if } \omega^* > \hat{\omega} \end{cases} \tag{3.6}$$

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 interesting cases where the punishment technology and the social return of the public good are sufficiently high to make at least one social class willing to contribute under the threat of a high-effort sanctioning institution.

According to Proposition 3, 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 $(\lambda,\ p_H, p_L, c, f)$, the existing wealth distribution in the group and the opportunity cost from providing the incentives for high effort to the institution faced by the political decisive agent. Specifically, a contributor renounces to his wealth while a free rider pays the fine with probability p_H. 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 $(p_H f)$ 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, $\frac{p_H}{(p_H - p_L)}c$.

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. 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. The next corollary is derived from Proposition 3.

Corollary 2*: Given a wealth distribution* $\{(\omega^P, \omega^M, \omega^R), (q^P, q^M, q^R)\}$ *and a set of institutional and technological parameters* $(\lambda,\ p_H, p_L, c, f)$ *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 political decisive agent 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-class individual will implement the institution in a larger 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.

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 decisive political agent 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 3*: For very low expected fines ($\omega^P > p_H f$), a government representing a free-rider political decisive agent will implement the sanctioning institution in a larger range of cases than a government represent-*

ing a contributor political decisive agent. For sufficiently high expected fine ($\omega^P \leq p_H f$) the government representing the poor class will implement the sanctioning institution in a greater range of cases than the government of the middle or the rich class.

3.4.3 Social welfare

In this section, we address 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.

$$SW \approx \lambda W \tag{3.7}$$

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 would implement it or not. Recall that for the case of full contribution, a government representing a political decisive agent with wealth ω^* will implement the institution if and only if $\frac{\lambda}{n}W \geq \omega^*$, or equivalently $\lambda W \geq n\omega^*$.

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

Proposition 4: *If $\omega^R \leq \hat{\omega}$, a poor-class government will always implement the institution, making the socially efficient decision. A middle-class or rich-class government may not implement the institution in situations where it is socially efficient to do so.*

Formal proof can be found in appendix 3. Intuitively, even though it is 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 return of the public good composed by contributions of the three social classes. 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.[7] Ultimately, this will depend on the wealth distribution and on the social return of the public good λ.

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

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:

$$SW = \lambda(q^P\omega^P + q^M\omega^M) + q^R(\omega^R - p_H f) - \frac{p_H}{p_H - p_L}c \tag{3.8}$$

or equivalently,

$$SW = \lambda W - [q^R(\lambda - 1)\omega^R + q^R p_H f + \frac{p_H}{p_H - p_L}c] \tag{3.9}$$

Therefore, at a first glance we can see that there is a welfare loss captured by the second term of equation 3.9 which entails, on the one hand, the net loss of not contributing plus the expected fines for all the rich class and, on the other hand, the institutional costs (sheriff's 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:

$$(\lambda - 1)(q^P\omega^P + q^M\omega^M) \geq q^R p_H f + \frac{p_H}{p_H - p_L}c \tag{3.10}$$

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

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:

$$SW = \lambda W - [q^M((\lambda-1)\omega^M + p_H f) + q^R((\lambda-1)\omega^R + p_H f) + \frac{p_H f}{p_H - p_L}] \quad (3.11)$$

Notice that the SW loss captured in equation 3.11 is greater than the one obtained in the previous case of partial contribution with the poor and middle class (equation 3.9), 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:

$$(\lambda - 1)q^P \omega^P \geq (q^M + q^R)p_H f + \frac{p_H}{p_H - p_L}c \quad (3.12)$$

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 ($\omega^M \leq \hat{\omega} < \omega^R$). To do so, let's take the condition for the implementation to be SW maximizing and rewrite it in the following way:

$$\frac{\lambda}{n}[q^P \omega^P + q^M \omega^M] - \frac{p_H}{n(p_H - p_L)}c \geq \frac{q^P \omega^P + q^M \omega^M + q^R p_H f}{n} \quad (3.13)$$

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

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

Notice these two conditions are equivalent on the LHS and differ on 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 criterion is different to the one followed by the government. Only if, coincidentally, the political decisive agent's opportunity cost were equal to the average of everybody's opportunity cost, the decision of this government would always be efficient. From the comparison of these two conditions, we can assert the following results:

***Proposition 5**: 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.*

For sufficiently high expected fines ($\omega^P \leq p_H f$) the government representing the poor class will implement the institution in situations where it is inefficient to do so, while the government representing the free-riding rich class will not implement the institution in situations where it is efficient to do so. However, for very low expected fines ($\omega^P > p_H f$), the result is just the opposite. A government representing a free-rider political decisive agent will implement the institution in situations where it is inefficient to do so.

3.5 The determinants of the emergence of a sanctioning institution

Let us next study the effects of changes in the different parameters on the implementation of a high-performance sanctioning institution. Changes in the parameters could trigger out a change in the condition given by equation 3.6 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 political decisive agent. Additionally, notice that a change in the parameters that determine the contribution threshold, $\hat{\omega}$, could affect the contribution decision of the different social classes and, therefore, the size of the public good. For instance, if initially 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. We call this effect the *switch effect* and if it occurs it will have an impact on the public good provision.

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

3.5.1 Institutional and technological variables

The conditions for the implementation of a high-performance sanctioning institution will be affected 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.

We previously obtained that, given a fixed population size and wealth distribution and a high effort sanctioning institution, the amount of public good provision will be non-decreasing on the social return of the public good λ, on the probability of fraud detection under high effort p_H and, on the fine f. Therefore, the same will occur with the individual return of the public good. This effect will be even stronger if the *switch effect* occurs, that is, a social class changes its behaviour from free riding to contribution.

On the other hand, recall that the existence of moral hazard implies paying the institution a per capita economic rent of $\frac{p_L}{n(p_H - p_L)}c$ when high effort is enhanced. 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 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 institution's effort.[8]

Assume now that the likelihood ratio decreases due to a fall in p_L or an increase in p_H. 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 an increase in the probability of punishing free-riding behaviour under high effort, a decrease in the probability of punishing free-riding behaviour 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 decrease in the cost of having such institution.

Therefore, these two previous variations, an increase in the individual return of the public good and a decrease in the cost of the institution will facilitate the implementation of a high performance institution for any type of government. However, our model highlights that this is true only if the political decisive agent is a contributor.

If the political decisive agent is a free rider, then an increase in the probability of fraud detection under high effort p_H or in the fine f, will increase the opportunity cost of the political decisive agent making more difficult the implementation of a high-performance institution. The final effect in this case will be unclear and depend on the particular initial configuration of the

[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.

parameters. We summarize all the previous analysis in the following two results.

Result 1: *An increase in the social return of the public good, a decrease in the probability of punishing free-riding behaviour under low effort and/or a decrease in the institutional cost of exerting high effort would make the sanctioning institution emerge in a greater range of cases under any type of government.*[9]

Result 2: *An increase in the fine paid by free riders and/or an increase in the probability of detecting free-riding attitudes under high effort would make the sanctioning institution emerge in a greater range of cases 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.*

Summarizing, intuition apparently indicates that an increase in the social value of the public good, an improvement in the monitoring and sanctioning technology or a diminution in the severity of the moral hazard problem existing with the institution, are all of them factors that will ease the implementation of a high-performance sanctioning institution. In fact, in a model with a homogeneous wealth population, the sanctioning institution will emerge more easily for any of the changes previously discussed. Our model highlights that this is only true if the political decisive agent is a contributor in the presence of the institution. The homogeneous model dis-

[9] For an increase in the social return of the public good, this result does not hold for the very particular case where $\omega^M > p_H f$.

regards that if the government represents a free-riding political decisive agent then the effect will be uncertain.

3.5.2 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 subsection we aim to analyse the effect on the emergence of the sanctioning institution derived from changes of these variables.

3.5.2.1 Variation in the composition of social classes.

A possible scenario could be a transfer of individuals between social classes without the population size changing. Consider as an example that certain middle-class individuals become poor, i.e. a reduction in the size of the middle class $\Delta q_P = \nabla q_M$. This variation only compromises the emergence of the sanctioning institution through the individual return of the public good. Consequently, the effect any variation would have ultimately depends on how does this movement affect the share of contributors and free riders to the public good. Any variation that makes free-riding individuals become contributors (*switch effect*) will increase the return of the public good, which in turn, will make the sanctioning institution emerge more easily. If, instead, the share of contributors does not change but contributors climb up the social ladder (for instance, poor-class contributors become middle-class contributors), the effect is also positive in the return of the public good. Alternatively, if the variation makes either a contributing social class become free rider or a contributing social class climb down the social ladder, the return of the public good will fall. Then, the impact on the emergence of a sanctioning institution will be negative.

Result 3: *Variations in the composition of social classes (q^P, q^M, q^R) keeping the population (n) constant, make the sanctioning institution emerge in a greater range of cases when either a free-riding social class becomes a contributor or a contributing social class climbs up the social ladder and remains as a contributor. Otherwise, the impact will be negative.*

Consider now an external shock in the population affecting uniquely part of it. A natural example is, for instance, an immigration wave of poor-class citizens, $\Delta q^P = \Delta n$. On the one hand, the increase in the population would directly make citizens pay a lower salary per capita, which makes the sanctioning institution emerge more easily under any type of government. However, to this effect we have to add the effect on the return of the public good. Notice that the proportion of the social class to which the immigrants belong would remain unchanged. For the example we proposed, the ratio q^P/n would remain unaltered. Nonetheless, the proportion of other contributing social classes would fall diminishing the individual return of the public good. This negative effect could be accentuated if the increase in the population caused a *switch effect* making contributors become free riders. This negative effect on the individual return of the public good will make the sanctioning institution emerge more difficultly. The combination of these two opposing effects will determine the net effect on the emergence condition.

An exception for this ambiguity would be that there was an immigration of poor-class individuals and that these were the only ones with incentives to contribute to the public good. In this case, the only effect that would be triggered would be the decrease of the per capita salary to pay to the sheriff. Thus, in this particular case, the immigration wave would make the sanctioning institution emerge in a greater range of cases under any type of government.

Result 4: *An immigration wave of any social class $\Delta q^j = \Delta n$ would have an ambiguous effect on the emergence of a sanctioning institution under any government, as it would diminish both the individual return of the public good and the salary per capita paid. Exceptionally, if $j = P$ and $\omega^P \leq \hat{\omega} < \omega^M$, that is, only poor individuals contribute, the impact would be positive and the sanctioning institution would emerge in a greater range of cases under any type of government.*

3.5.2.2 Enrichment and impoverishment of social classes

Finally, let's analyse the impact of a variation in the wealth level of some social class due to an external shock. An appealing example is an impoverishment of the middle class, $\nabla\omega^M$, given that many societies have suffered this misfortune after the scraps of the Great Recession. In this subsection we will consider an impoverishment of any social class j. As formerly exposed, results will depend on the incentives to contribute that this social class has.

Let's start by considering that the impoverished social class is a free-riding social class. If the change in his wealth were sufficiently large, a *switch effect* could occur, such that this group became a contributing social class. This, in turn, would trigger two effects. On one side, the return of the public good would increase, fact that would make any other government (k government $\forall_{k \neq j}$) implement the sanctioning institution more easily. For the government representing that social class (j government), however, this positive effect on the public good is combined with a change in the opportunity cost from $p_H f$ to ω^j. If $p_H f \geq \omega^j$, the net effect would be positive, and the j government would implement the institution in a larger range of cases. Otherwise, it would be uncertain.

***Result 5**: The impoverishment of any free-riding social class j that led them to become contributors, would make any k government ($\forall_{k \neq j}$) implement the sanctioning institution more easily. The j government, however, will only do so if the expected fine is sufficiently large ($p_H f \geq \omega^j$). Otherwise the net effect would be uncertain.*

If instead, the impoverished social class had incentives to contribute, the impact on the return of the public good would be negative. This would make it more difficult for any other government (k government $\forall_{k \neq j}$) to implement the institution. As before, for the j government, this effect is combined with the effect on the opportunity cost, which in this case, will always be lower. For the net effect to have a positive impact on the implementation of the sanctioning institution, the condition $\frac{q^j}{n} \leq \frac{1}{\lambda}$ should hold. Notice that this condition establishes that the proportion of individuals belonging to the impoverished class must be lower than than the social marginal rate of substitution between private and public goods.

Result 6: *The impoverishment of any contributing social class j, would make any k government ($\forall_{k \neq j}$) implement the sanctioning institution in a smaller range of cases. The j government, however, will only do so if the share of this class is lower than the social marginal rate of substitution between private and public goods ($\frac{q^j}{n} \leq \frac{1}{\lambda}$). Otherwise the net effect would be uncertain.*

Notice that an enrichment of any social class j follows the same intuition leading to opposite results.

3.6 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 political decisive agent 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 (see appendix 3).

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 vice versa. 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 vice versa.

3.7 Conclusions

This paper theoretically explores how do centralized sanctioning institutions subject to moral hazard problems emerge in selfish societies that must decide on the provision of a public good. The implementation depends on a player, the government, who represents the interests of a particular social class, the political decisive agent. Moreover, we study the level of contribution such institution can potentially achieve if implemented and the social welfare achieved with the institution.

Without any enforcing mechanism or with an institution with inappropriate incentives, selfish individuals will fully free ride on the one-shot public good. Nonetheless, with a high-performance sanctioning institution a positive provision of public good can be achieved. The incentives to contribute or to free ride of the different social classes are determined by the social value of the public good and by the parameters that characterize the punishing technology. Given a fixed group size and wealth distribution, societies with a relatively high quality sanctioning institution and high social return of the public good will have higher levels of contribution under a high-performance sanctioning institution.

Regarding the emergence and implementation of a high-performance sanctioning institution, we show that that the government representing the social class with the lowest opportunity cost (independently of it being a contributing or a free-riding social class) will make the efficient decision in a wider range of cases. This result depends on the interplay between the punishing technology, the severity of the moral hazard problem and the behaviour of the political decisive agent on the contribution game.

Our theory highlights the importance of the political decisive agent in the collective action problem. We show how his incentives to free ride or to contribute crucially affect the emergence of a high-performance sanctioning institution and the provision of the public good.

References

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

2. Aldashev, G., & Zanarone, G. (2017). “Endogenous enforcement institutions". *Journal of Development Economics.*

3. Baldassarri, D., & Grossman, G. (2011). “Centralized Sanctioning and Legitimate Authority Promote Cooperation in Humans." *Proceedings of the National Academy of Sciences*, 108(27), 11023-11027.

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

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

6. 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.

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

8. Fehr, E., & Williams, T. (2017). “Creating an efficient culture of cooperation." *Mimeo.*

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

10. 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.

11. Gächter, S., Renner, E., & Sefton, M. (2008). “The Long-Run Benefits of Punishment." *Science*, 322(5907), 1510-1510.

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

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

14. Heap, S. P. H., Ramalingam, A., & Stoddard, B. V. (2016). “Endowment inequality in public goods games: A re-examination." *Economics Letters,* 146, 4-7.

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

16. 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.

17. Okada, A. (1993). “The Possibility of Cooperation in an N-Person

Prisoners' Dilemma with Institutional Arrangements." *Public Choice*, 77(3), 629-656.

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

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

20. 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.

21. Sutter, M., Haigner, S., & Kocher, M. G. (2010). "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." *The Review of Economic Studies*, 77(4), 1540-1566.

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

23. Weingast, B. R., Shepsle, K. A., & Johnsen, C. (1981). "The political economy of benefits and costs: A neoclassical approach to distributive politics." *Journal of political Economy*, 89(4), 642-664.

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

Appendix 3. Proofs

Proof of Proposition 2:

Let's analyse the different cases:

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 2:

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 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:

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

The government maximizes the utility of the 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.

We assume that the government has all the bargaining power when negotiating with the sheriff. It is easy to deduce that the contract that maximizes the objective function 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. From this, we derive the contract enhancing high effort: $\{s_0 = 0, s_p = \frac{c}{p_H - p_L}, s_{np} = 0\}$. Under this contract, economic rents are: $\frac{p_L}{p_H - p_L}c$.

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 4:

Given that $\lambda < 1$, then $\lambda W > W$ always holds, where $W = q^P \omega^P + q^M \omega^M + q^R \omega^R$.

For the case of a poor-class government, $W > n\omega^P$, therefore, $\lambda W > n\omega^P$ always holds, which implies that this type of government will always implement the sanctioning institution, which is the socially efficient action. However, for a middle or a rich class government $n\omega^R > n\omega^M > W$, so the relationship between λW and $n\omega^M$ or between λW and $n\omega^R$ is unclear. This implies that this type of governments may not implement the institution in situations where it is socially efficient to do so.

■

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 thanthe 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 middle and high-valuation individuals and $g^j = 0$ for low-valuation players.
- 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

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: $\{s^H\} = \{s_0 = 0, s_p = \frac{c}{p_H - p_L}, s_{np} = 0\}$ for high effort and $\{s^L\} = \{s_0 = 0, s_p = 0, s_{np} = 0\}$ for low effort.

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 $\lambda^j < \hat{\lambda}$

■

Chapter 4

Sanctioning as a noisy signal

4.1 Introduction

Public goods provision has been broadly discussed as a social dilemma. While the social optimum is reached when everybody fully contributes to the public good, there are incentives to deviate to free riding, leading to an inefficient outcome. Multiple mechanisms have been proposed to conceal the free rider issue, being sanctioning the most regarded one. Amongst all of the various ways in which this punishment mechanism can be implemented, the most traditional approach has been decentralized punishment among peers, where every player contributes and has the option to punish. Nonetheless, in large societies, decentralized peer punishment is at times inefficient or not implementable and sanctioning is delegated to a central authority (Gächter, Renner and Sefton 2008; Sigmund, Traulsen, Hauert 2010; Fehr and Williams 2017).[1]

[1] Peer punishment usually causes high collateral damage in interactions over extended periods of time as the costs tend to exceed the gains of cooperation (Gächter, Renner and Sefton 2008). Additionally, for it to be effective, it requires punishers to be pro-social

The main question to approach in this paper is how should the implementation of punishment from these centralized institutions work, in the sense that some individuals uniquely contribute while others can implement sanctions. In this work, our aim is to explore the impact of two different payoff schemes in a centralized sanctioning environment: (i) a *fixed scheme*, where the sanctioner is provided certain level of endowment to decide on the punishment actions and (ii) a *variable scheme*, where instead he receives an endowment proportional to the level of cooperation attained. The positive effect of punishment is a renown result, however, this paper sheds light on the impact of the payoff scheme in the level of cooperation achieved. In particular, providing the sanctioner a fixed payoff increases significantly the provision of the public good from the contributors, even with less punishment activity of the sanctioner, and consequently social welfare improves. This occurs despite contingent-payoff sanctioners implement more punishment and contributors display a greater responsiveness to sanctions. The reason behind this is that when the sanctioner's endowment is fixed, contributors increase their willingness to cooperate.

In a punishment environment, the sanctioner's action will depend on two features: (i) the punishment scheme and (ii) the payoff scheme. Concerning the first feature, punishment is centralized to a unique figure, who must decide whether to sanction or not at a fixed cost per punished contributor. The use of costly punishment to enhance cooperation can be understood as a

in order to be willing to bear the high costs, fact that is at times unlikely (Sigmund, Traulsen, Hauert 2010). From an evolutionary perspective, as groups increase in size, individuals leave the inefficient peer-punishment environments and migrate to groups with more efficient law forces (Fehr and Williams 2017).

signal of discomfort with the level of cooperation within the group (Schoenmakers, Hilbe, Blasius, and Traulsen 2014).[2]

Regarding the second feature, the payoff scheme, we present a fixed and a variable payoff scheme. While the fixed endowment goes in line with standard centralized punishment literature, the variable endowment follows the concept of pool punishment. This notion describes those situations where sanctioning is centralized and outsourced to a monitoring figure or institution, whose payoff must depend on both the amount of cooperation achieved and the sanctioning carried out (Kosfeld, Okada and Riedl 2008; Sutter, Haigner and Kocher 2010).[3] The novelty of this work is that it outlines the impact of the payoff scheme on contributions. Notice that while a fixed-payoff-scheme sanctioner is an independent entity, a variable-payoff-scheme sanctioner is also benefitting from the public good without any contribution request. Thus, contributors could feel that sanctioners are free riders of their costly contributions. With this experiment we show that contributors belonging to groups where the sanctioner's payoff was fixed contributed more than those in groups where the sanctioner received a contingent payoff. This happened notwithstanding the fact that contributors were more responsive to being punished when sanctioners had contingent payoffs. This

[2] They study the signaling effect of centralized sanctioning claiming that effective sanctioning institutions make use of it, along with other mechanisms, in order to reduce the temptation to free ride.

[3] This reflects how investments in monitoring and sanctioning institutions are actually made. Examples are specialized law forces such as the police, courts, state and non-state institutions. Several studies have paid attention to the endogenous formation of pool punishment institutions. In this line, Kosfeld, Okada and Riedl (2008) show that the endogenous formation of these institutions, which enhance cooperation and have positive effects on group cooperation. Sutter, Haigner and Kocher (2010) include rewards as well as punishments and find a positive effect on cooperation of endogenous institutional choices in comparison to the same exogenously implemented institutions.

reflects that contributors ascertained the extrinsic motivation of sanctioners endowed with a variable payoff and their willingness to cooperate was lower.

The payoff scheme is also determinant for the sanctioners' behaviour. Notice that sanctioners with a variable endowment could have both the intrinsic motivation of fighting for a better world and the extrinsic motivation of pushing contributions upwards, as it also pours on their future payoff. Our results demonstrate that, for the sanctioners, the payoff scheme in fact matters: with a contingent payoff, sanctioners implement punishment more frequently than with a fixed payoff, i.e. send the signal more often. Furthermore, as their endowment increases in size, they sanction less. This result is also present with decentralized peer punishment.

Regarding different punishment schemes, literature has previously made comparisons between peer punishment and centralized punishment (Sigmund, De Silva, Traulsen and Hauert 2010; Milinski, Traulsen and Röhl 2012; Gross, Méder, Okamoto-Barth and Riedl 2016)[4] or between different types of centralized sanctioning institutions (Kamijo, Nihonsugi, Takeuchi and Funaki 2014).[5] Nonetheless, there is a gap concerning comparisons

[4] Sigmund, De Silva, Traulsen and Hauert (2010) and Milinski, Traulsen and Röhl (2012) show that centralized punishment is not only a more realistic approach, as a matter or fact, it is the punishment option preferred by individuals. Additionally, this effect is even more pronounced with second-order punishment. Gross, Méder, Okamoto-Barth and Riedl (2016) explore the effects of transferring the sanctioning power between contributors. According to their experimental results, introducing such voluntary transfer enables the maintenance of cooperation in a decentralized context, by empowering those who are willing to punish in the interest of the group.

[5] They carry out a comparison between an absolute punishment institution where everybody who contributes below a given threshold is punished, and a relative punishment institution, where only the lowest contributor receives such punishment. They show how the latter performs as least as good as the former one, proving that it is only necessary that one free rider is punished for punishment to act as a credible and effective threat.

among centralized institutions presenting different payoff schemes, which is precisely, the aspect of pool punishment with greater discrepancies. Many papers ignore the public good feature of pool punishment and provide the sanctioner a fixed endowment, independent of the level of cooperation attained (Baldassarri and Grossman 2011).[6] In contrast with this study, our target is to prove that the payoff function has an impact on the players' behaviour as well. To do so, we use an approximation of Yamagishi (1986) payoff by providing the sanctioner an endowment that is equivalent to certain percentage of the public good, as a form of pool punishment. By doing so, we aim to compare this context of pool punishment with a fixed-payoff form of centralized punishment. This provides new insights regarding how different centralized sanctioning institutions work and how should their workers be paid if they pursue improvements in terms of efficiency.

Specifically, we implement a between-subjects design where subjects are randomly assorted into groups of 4 and play a 10-period public goods game. In each group, one of the four subjects will be detached the contribution decision and will be given the power to sanction others in his group. Subjects either participate in a not pool punishment environment or in a pool punishment environment, where the difference lies on how the sanctioner's endowment is decided. For the first one, the endowment was fixed, whereas for the second one it was proportional to the level of cooperation achieved by the group in such period. We additionally implement two treatments where this type of player was in fact a mere observer, with either a fixed or

[6] They show that in a context of pool punishment with a fixed endowment, contributions depend on the perceived level of legitimacy, i.e. in their lab-in-the-field experiment, subjects were more responsive when the monitor was elected than when he was randomly chosen. However, the monitor's sanctioning decision did not depend on his legitimacy.

a variable payoff. This way, we can carry out a study that properly explores the impact of pool punishment in contributions and sanctions with respect to other centralized punishment designs.

Beyond the standard result of the positive impact of punishment on contributions, in this paper we verify that there is a scheme effect both for sanctioners and contributors. On one side, sanctioners implement punishment as an attempt of signalling at a larger extent when their payoff is contingent to the public good provision. They are not concerned on whether their payoff is of a fixed or variable nature, but on the size that such payoff has. On the other side, contributions are higher when their sanctioner receives a fixed payoff instead of a variable one. The free-riding origin of the sanctioner's earnings detriments their willingness to contribute to that cause. This has an impact on efficiency, being a fixed-payoff centralized environment with sanctioning opportunities the one that maximizes social welfare. The fact that the sanctioning signal is sent more frequently under pool punishment and contributors have lower willingness to cooperate, accounts for such signal to be noisy. This work is, to the best of our knowledge, the first paper gathering this effect and providing an explanation for it.

Considering a framework of centralized punishment and, in particular, of pool punishment, there are two research lines that have been followed. On the one hand, some authors have focused on the implementation of different systems. Ozono, Watabe and Shimizu (2016) study how the establishment of a leader-support system increases both contributions and efficiency achieved with punishment in a public goods game. The idea behind this system is that, once contributions are made, individuals are given the chance to support a leader who can use the public good's capital to freely punish. The implementation of this system is experimentally proven to be efficient, especially when leaders punish both non-contributors and non-supporters. In this line, Baldassarri and Grossman (2011) carried out a lab-in-the field experiment finding out that groups reach higher level of cooperation in the presence of a leader. Our paper can be classified in this line because, even if it does not present a leader-support system, it proposes an alternative structure that enhances cooperation.

The rest of the paper is organized as follows. Section 4.2 describes the experimental design. In next place, Section 4.3 describes the theoretical predictions and enumerates the hypothesis that will be contrasted. Section 5.3 presents the results and, finally, Section 5.4 contains concluding remarks.

4.2 Experimental design

The experiment was carried out at the Laboratory for Research in Behavioural Experimental Economics (LINEEX) from the University of Valencia during December 2017. A total of 336 participants took part in 6 sessions of 56 subjects each. Subjects only participated in one of the four treatments, thereby being 84 subjects per treatment. The experimental currency was expressed in points, where 15 points=1€. Each session lasted approximately 90 minutes and the average earnings were 15€. See translated instructions in appendix 4.A.

Subjects were randomly assorted into groups of 4, with which they played a 10-period PGG with partner matching. In each group, 3 subjects were randomly assigned role Type A (contributors) and 1 was assigned role Type B (monitor). Each period had two stages. In the first stage, the 3 contributors were given an endowment of 10 points, from which they had to decide how many points invest into a common project, and how many to keep as savings. The sum of the group investments was multiplied by 2/3. In the second stage, every group member observed the individual contributions arranged in a decreasing order and without individual identification. The monitor then received an endowment and had to decide who to sanction, if anybody, in the group. The sanction had a cost of 1 point for the monitor and a negative impact of 3 points for the sanctioned contributors.

Before the 10 experimental rounds, subjects participated in 10 unpaid trial rounds where all of them where Type A players (contributors) and no sanctioning stage was played, i.e. standard PGG. After these trials, subjects were rematched for the experiment.

In this experiment we implemented two treatments with different endowment levels for the monitor: *Sanctioner Not Pool (SNP)* and *Sanctioner Pool (SP)*. Sanctioners participating in the SNP were provided a fixed endowment of 15 points to carry out their punishing actions, whereas sanctioners participating in the SP were given a contingent endowment equivalent to 5 points plus 2/3 of the group contributions. How the sanctioner's endowment was going to be defined was communicated to the contributors before their decisions were made.

Additionally, we implemented two treatments without punishment: *Observer Not Pool (ONP)* and *Observer Pool (OP)*. In these cases, the monitor is an observer, a mere spectator who makes no decision. The observer's payoff is symmetric to the sanctioner's endowment: 15 points for observers participating in the ONP and 5 points plus 2/3 of the group contributions for observers participating in the OP. In each session, two treatments were simultaneously implemented: either ONP and OP or SNP and SP, such that within a session the difference lied on the monitor's endowment.[7]

4.3 Theoretical predictions and hypothesis

Assuming everybody has purely selfish preferences, the Nash equilibrium of the game would be that every contributor fully free-rode on the public good (in the four treatments) and that no sanctioner carried out any punishing, under both treatments SNP and SP. Nonetheless, experimental results have repeatedly revealed the existence of other forces, different to the theoretically predicted, driving and explaining individuals' behaviour. In order to explore such forces in the presented context, we contrast the following series of hypotheses, aligned with the rationality assumptions, such that rejection on any of them implies deviation from the Nash Equilibrium.

H1: In a context with punishment opportunities, average contributions

[7] Player B's endowment was not specified on the printed instructions, which were read out loud by the experimenter, but displayed on screen for every group member.

are the same than in a context without punishment opportunities.

Hypothesis 1 studies the impact of sanctioning opportunities in a public goods game, regardless of the payoff scheme of the monitor. According to the theoretical predictions formerly stated, contributions should be the same and should be equal to zero. This holds independently of there being a sanctioner, as they expect these not to punish. By contrasting this hypothesis we aim to find out a potential *sanction effect*.

H2: In a context where the monitor receives a fixed endowment, average contributions are the same than in a context where the monitor receives a variable endowment.

Hypothesis 2 aims to study the impact of the payoff scheme on the average contributions in both the non-sanctioning and the sanctioning scenario. According to rationality assumptions, these should both be equal to zero independently of the mointor's endowment. With this comparison we pretend to analyse a potential *scheme effect*.

H3: Contributors are equally responsive to sanctioners' behaviour when the sanctioners have a fixed endowment and when the endowment depends on the level of cooperation.

H4: Sanctioners are equally responsive to contributors' behaviour when they have a fixed endowment and when the endowment depends on the level of cooperation.

As decisions are not isolated and are rather a response to others' actions, with these two final hypotheses we intend to explore, with each payoff scheme, the reciprocal reaction of contributors to sanctioners and vice versa.[8] These comparisons, jointly with the other results, will shed light on the strategy each type of players followed.

4.4 Results

In the study of the subjects' behaviour, we are going to disentangle the different effects influencing contributors and sanctioners in their decision-making. Starting with the contributors, we will study two possible effects on their contributions: the effect of introducing sanctioning opportunities and the effect given by the payoff scheme of the monitor. Moving on to the players' strategies, we will shed light on the factors affecting each type of player in their actions. Additionally, we will study the responsiveness or reactiveness exhibited to the decisions of the other type of players.

4.4.1 Preliminary statistical overview

Once the 10 trial rounds were over, subjects were randomly re-matched and were assigned one of two roles: either Type A (contributor) or Type B (monitor). This role was fixed during the 10 experimental rounds. Figure 4.1 shows the evolution of contributions over the 10 rounds for ONP and OP on the left and SNP and SP on the right. The first aspect to be highlighted

[8] Recall observers in ONP and OP have no set of actions.

is the difference in the trends when there was an observer and when there was a sanctioner. In both cases, contributions follow a decreasing pattern, but this effect is noticeably steeper in the treatments with observers.

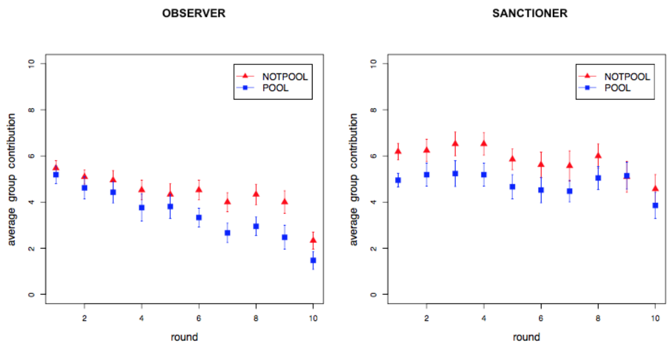

Figure 4.1: Evolution of contributions through the experimental rounds

Going into further detail, we can appreciate how contributions when the monitor receives a fixed endowment are higher than we he receives a contingent one in both contexts (ONP vs. OP and SNP vs. SP). This happens right from the first period. However, it is worthwhile to highlight how for the case with an observer, the gap between contributions is small in the first periods and widens as time goes by. For the case of the sanctioning environment, however, the gap remains fairly constant over time. Additionally, we can appreciate how, broadly speaking, contributions with a sanctioner (SNP and SP) are larger than those with an observer (ONP and OP). To check this, we compute the average contributions for each of the four treatments, presented in Table 4.1.

	Not Pool	Pool
Observer	4.33	3.46
Sanctioner	5.84	4.82

Table 4.1: Average contributions in points

Differences in average contributions among the four treatments are statistically significant.[9] In particular, when the monitor had the chance to punish, average contributions were higher than when he did not ($p^{***} = 6.6581$ x 10^{-18} for ONP vs. SNP and $p^{***} = 1.4764$ x10^{-14} for OP vs. SP). This accounts for a *sanction effect*. Furthermore, when the monitor was endowed with a fixed amount of points, average contributions turned out to be larger than when the monitor had a variable endowment ($p^{***} = 2.3535$ x 10^{-7} for ONP vs. OP and $p^{***} = 9.3964$ x 10^{-9} for SNP vs. SP). This accounts for a *scheme effect*. The quantification of these effects will be measured in the following subsection.

Regarding how did the sanctioners behave, the first characteristic to emphasize is the difference in the average number of sanctions between the two treatments, where the only difference was the payoff scheme. As it can be appreciated in Figure 4.2, sanctioners with a variable endowment carried out a significantly greater number of punishment actions than those with a fixed endowment ($p^{***} = 0.0017$).

[9] For these comparisons, we carried out a t-test where the null hypothesis considered equal contributions and the alternative hypothesis considered greater contributions

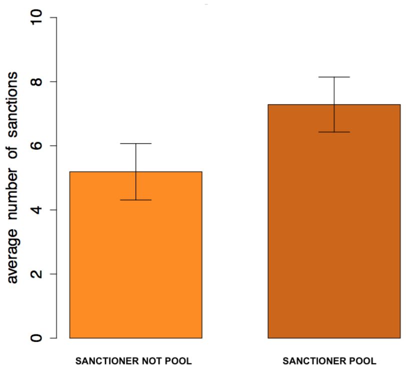

Figure 4.2: Average number of sanctions in *Sanctioner Not Pool* and *Sanctioner Pool*

4.4.2 Sanction and scheme effect

It has been proven that sanctioning significantly increases contributions whereas changing from a fixed to a variable payoff scheme significantly diminishes them. In this subsection we seek to quantify the impact of sanctioning and the scheme in the discrepancies among the contributions.

It is clear that the decision a contributor makes in a particular moment of time is influenced by several environmental characteristics, where some of them are particular for each treatment. Here, we aim to measure what is the effect of the treatment with its underlying traits on contributions. With this end, we run a mixed-effects regression.[10] where the contribution of a subject i at moment t is explained by the two dummy variables that define a treatment: *sanction* and *pool* (See Table 4.2)

[10] Linear mixed model fit by maximum likelihood. T-tests use Satterwhaite approximations to degrees of freedom. The dependent variable was contribution at round t. AIC= 12724.3.

	Estimate	**Std. Error**	**Pr($>\|t$)**
Random Effects			
Subject ID	-	0.8803	-
Group	-	0.6916	-
Fixed Effects			
Intercept	4.4333	0.3060	0.000 ***
Sanction	1.4232	0.1205	0.000 ***
Pool	-1.0889	0.3913	0.015*

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.2: Sanction and scheme effect

From this we can assert that a contributor belonging to a sanctioning environment contributes 1.42 points more than a contributor belonging to a non-sanctioning environment. Furthermore, a contributor whose monitor has received a variable endowment contributes 1.09 points less than a contributor whose monitor is endowed with a fixed amount. Thus, we can reject H1 and H2 and quantify the sanction and scheme in the following two results.

RESULT 1: *Average contributions when there is a sanctioner are 1.42 points larger than when there is only an observer.*

RESULT 2: *Average contributions when the monitor's payoff depends on the level of cooperation are 1.09 points lower than contributions when the monitor's payoff is fixed.*

of citizens' to sanctioners' behavior and we present this in the following result:

4.4.3 The contributing strategy

During this experiment, contributors' behaviour depended on a set of variables, which we dispose to classify into 4 categories: (i) contribution variables, (ii) fine variables, (iii) experimental setting variables and (iv) socio-demographic variables. To deeply analyse the contributors' strategy we are going to distinguish between contributors deciding in a sanctioning-environment and in a non-sanctioning environment, given that the latter ones have no type (ii) variables in their information set.

The selected models[11] that explain the contributing strategy are found in Table 4.3. Multiple other models have been tested with different combinations of variables, which can be found in appendix 4.B.

	Sanctioning environment			**Non-sanctioning environment**		
	Estimate	Std. Error	Pr($>\|t$)	Estimate	Std. Error	Pr($>\|t$)
Random Effects						
Subject ID	-	0.2551	-	-	0.577	-
Fixed Effects						
Intercept	1.8567	0.2743	0.000***	1.2966	0.5775	0.0255*
Contribution $t-1$	0.3125	0.0254	0.000***	0.331	0.0249	0.000***
Avg. Contr. $t-1$	0.4433	0.032	0.000***	0.3096	0.0326	0.000***
Group Fines $t-1 \geq 2$	0.9477	0.2174	0.000***	-	-	-
Round	-0.11461	0.0246	0.000***	-0.1515	0.0252	0.000***
Pool	-0.2299	0.1615	0.155	-0.4546	0.1623	0.0053**
Age	-	-	-	0.048	0.0245	0.0508 .

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.3: Contributing behaviour

In both environments, contributors were influenced by their own con-

[11] Linear mixed models fit by maximum likelihood. T-tests use Satterwhaite approximations to degrees of freedom. The dependent variable was contribution at round t. AIC= 5938.4 for the sanctioning environment model and AIC=5883.8 for the non-sanctioning environment model. Group has been deleted as a random effect given that the group effect is now collected in other variables.

tribution in the previous period and the average group contribution in the previous period (contribution variables). They both had a positive impact in their current contributing decision, which implies that high contributions today lead to higher contributions tomorrow. Furthermore, we compared these two variables separately with a third variable that collected the deviation of the contribution with respect to the group average, which was less explanatory in the contributing behaviour.

Moving on to the fine variables, recall that these can only be determinant in the sanctioning environment. In this regard, we consider two variables: a dummy representing whether an individual has been sanctioned in the previous period and a variable collecting the group number of fines in the previous period. To have an initial overview of the effect of group fines on contributions, we study how do fines affect contributions from one period to another by examining which is the number of sanctioned contributors necessary for group contributions to increase in the next period. We can conclude that sanctions are only effective in rising contributions when at least two group members are punished. Otherwise, contributions fall (see Figure 4.3). This invites us to include another explanatory variable in the analysis, *Group Fines* $t-1 \geq 2$. This variable displays a greater explanatory power than the two first variables which indicates that contributions get particularly boosted when at least 2 of the 3 contributors are sanctioned.

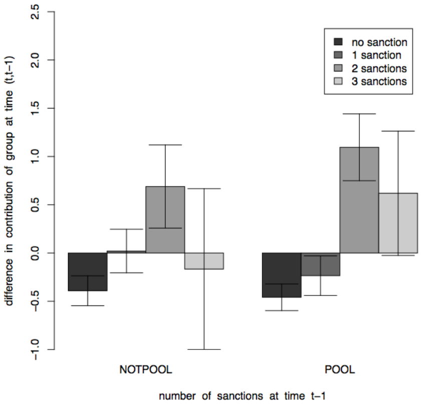

Figure 4.3: Number of sanctions and effect on variations in contributions in *Sanctioner Not Pool* and *Sanctioner Pool*

The experimental setting variables include *round* and *treatment*. In first place, in what refers to the round, this has a negative and significant impact in both the sanctioning and the non-sanctioning environment.[12] That is, contributions decrease over time. As we already anticipated, this fall is more accused in the case of the treatments with observer (non-sanctioning environment). In second place, the treatment variable (which collects the pool effect) has a negative impact in both environments. In other words, contributions are lower in a pool context environment where monitors are endowed with a contingent amount of points in each round. Its effect is only statistically significant in the non-sanctioning environment. Nonetheless, despite its non-significance in the sanctioning environment, it improves the explanatory power of the model.

In last place, in the category of socio-demographic variables, only *age*

[12] We also tested models with the variable $Round^2$ to examine a possible memory effect, which we discarded.

has a significant impact and only in the non-sanctioning environment. In particular, its effect is positive, which indicates that contributions are increasing with age.

4.4.3.1 Contributors' responsiveness

Once we understand which factors, in general terms, determine the contributors' behaviour, let's study the strategic interaction between contributors and sanctioners from the formers' point of view. In other words, let's examine how do contributors react to sanctions. With this end, and given the structure of the game, we analyse how do contributions change between two periods $t-1$ and t given the sanctioning decision in period $t-1$. As expected, this effect is positive. In particular, receiving a fine increases contributions in 2.04589 points. However, in order to test for H3, we wish to see whether this reactiveness is the same for the pool and the not pool cases. By running the regressions separately we observe that fines increase contributions by 1.8643 points when the sanctioner has a fixed endowment and by 2.1709 points when he has a variable one. Therefore, we can reject H3 and claim that responsiveness differs across treatments.

RESULT 3: *Contributors are not equally responsive to sanctioners' behaviour under the different payoff schemes. When the sanctioners have a fixed endowment, receiving a fine increases contributions in 1.86 points and when the endowment depends on the level of cooperation in 2.17 points.*

4.4.4 The sanctioning strategy

With this experiment we also aim to understand the forces driving the Type B players to behave as they did. Recall these players only took action in the treatments with sanctioning opportunities (SNP and SP), as in the others they were simple observers of what was happening within their group. As we already saw in Figure 4.2, sanctioners make use of their sanctioning power. This action is a costly action with no direct benefit to the sanctioner and does not belong to the Nash equilibrium. Therefore, sanctioning can be interpreted as a signal.

There are two approaches we can follow to study the sanctioning strategy. On the one hand, we can analyse which factors affect the probability that a contribution in period t is fined. Following this first approach, the new contribution variables are now translated to period t. When a sanctioner decides on whether to sanction or not a contribution in period t, he has information about what that contribution in that period t has been as well as the group average in that period t. In the selected model[13] (see appendix 4.B. for the complete set of models), contribution has a negative effect, which implies that the larger a contribution is, the lower the probability of being punished. The group average, however, has a positive effect. That is, the higher the group's contribution is, the higher the probability of that contribution getting punished, result indicating that sanctioners become more demanding as contributions increase. As before, the variable

[13] Generalised linear mixed model fit by maximum likelihood (Laplace approximation). The dependent variable was the probability of receiving a fine at round t. AIC$=$ 984.2 Group has been deleted as a random effect given that the group effect is now collected in other variables.

that reflects deviation from the average contribution is shown to have less explanatory power.

With regard to the fine variables, notice that whether that individual has received a fine in the previous period does not belong to the sanctioner's information set. When sanctioners decide whether or not to sanction contributions there is no identification in order to avoid reputation effects. The total number of group fines in the period, however, indeed is information the sanctioner has. In this line, this variable does not seem anchoring for them, as it has no significant impact.

The round as an experimental setting variable displays a negative impact. Hence, not only contributions fall as time goes by, the same can be said about the probability of sanctioning. The treatment variable has a positive impact. That is, the probability of receiving a fine is larger for contributors whose sanctioner's payoff is contingent to their contributions. This goes is line with what we already observed in Figure 4.2, where we noticed that pool sanctioners implemented a larger number of sanctions than not pool sanctioners.

Concerning the socio-demographic variables, they now refer to the sanctioner: his gender and his age. In this line, results demonstrate that men significantly sanction more than women.

Despite it being a best response no to punish, sanctioners' implement costly punishment with the aim of pushing contributions upwards. We can take notice of this from previous results as well as from the behavior of several groups where sanctioners punish until contributions reach a certain level and stabilize. To study whether the reason behind this is the sanctioner's payoff, we can use the sheriff's endowment as an explanatory variable in the sanctioning behavior. However, given possible multicollineality issues that may arise between *sanctioner's endowment* and *pool*, we use *sanctioner's payoff* as a proxy, where the only difference is that the punishment costs have been deducted. By doing this, we assume that sanctioners evaluate the cost that their decisions imply beforehand. This variable has a significant and negative impact in the probability of being punished. That is, the higher the sanctioner's endowment/payoff is, the lower the probability he will fine. This reassures the idea of sanctioners using their punishment power as a signal of discomfort. As a final comment, notice that, in contrast with contributors, *pool* has a non-significant impact in the explanation of the sanctioner's behaviour. This implies that sanctioners do not sanction because of the fact of having a fixed or a variable endowment, but because of how their sanctions can affect their their future payoff.

	Estimate	Std. Error	Pr(> $\|t$)
Random Effects			
Subject ID	-	0.6051	-
Fixed Effects			
Intercept	2.33421	0.51544	0.0000***
Contribution t	-0.29609	0.03340	0.0000***
Avg. Contribution t	0.47586	0.05151	0.0000***
Round	-0.06849	0.02948	0.0202*
Pool	0.33225	0.21172	0.1166
Sanctioner's gender	0.48472	0.19078	0.0111*
Sanctioner's final payoff	-0.36565	0.04026	0.000***

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.4: Sanctioning behaviour - Probability of receiving a fine

The second approach we can follow in studying the sanctioning strategy is to try to explain which factors affect the probability of a sanctioner punishing. To do so, we implement an ordered probit of the group fines in period t. The selected model is presented in Table 4.5. Notice that the average group contribution and the group fines in the previous period increase the probability of punishment, whereas the round and the sheriff's payoff, on the other side, show a negative impact.

	Estimate	Std. Error	Pr(> $\|t$)
Coefficients			
Avg. Contribution t	0.14182	0.02712	0.0000***
Group fines $t-1$	0.606839	0.069436	0.0000***
Round	-0.097530	0.019219	0.0000***
Sanctioner's final payoff	-0.165772	0.021163	0.0000***
Intercepts			
0|1	-1.025910	0.254009	0.0000***
1|2	-0.203835	0.253279	0.42094
2|3	0.666835	0.266732	0.01242*

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.5: Sanctioning behaviour - Probability of sanctioning

4.4.4.1 Sanctioners' responsiveness

In the analysis of the sanctioners' behaviour, we have already seen how do sanctioners globally respond to contributions with their fines. Nevertheless, in order to contrast H4 we urge to study whether this impact differs in the pool and not pool environments. Carrying out the analysis separately, we see that on the one hand, when the sanctioner has a fixed endowment an additional contribution point decreases the probability of punishing that particular contribution in 0.22. On the other hand, when the sanctioner has a variable endowment an additional contribution point decreases the probability of punishing that particular contribution in 0.4172. Thus, we can reject H4 and assert that responsiveness differs across treatments.

RESULT 5: *Sanctioners are not equally responsive to contributors' behaviour under the different payoff schemes. When the sanctioner has a* *fixed endowment, an additional contribution point decreases the probability of sanctions in 0.22 and when the endowment depends on the level of cooperation in 0.42.*

4.4.5 Social welfare

Having reached this point, an imperative question is: *does this mean that not pool punishment is a better centralized scheme that pool punishment?* In order to answer this question, we must define what we understand for 'better". Following an additive social welfare function approach, we can compute what is the sum of the welfare under each treatment, where the welfare are the net payoffs of each player. As we can appreciate in table 4.6, sanctioning increases social welfare, result in line with previous literature, as in both scenarios where monitors had sanctioning power, total welfare is higher. Additionally, what this study illustrates is how making sanctioners' payoff contingent of the contributions substantially diminishes social welfare. Thus, we can assure that fixed payoffs are better than variable payoffs.

	Not Pool	Pool
Observer	12177	10980.73
Sanctioner	12692	11802.35

Table 4.6: Social Welfare - Total payoffs in points

RESULT 5: *Social welfare is higher when monitors are given sanctioning power. Additionally, social welfare is higher when the monitor's endowment* *is fixed. Therefore, the maximum social welfare is reached in a sanctioning environment with a fixed payoff scheme.*

4.5 Conclusions

Sanctioning can be employed as a mechanism that enhances cooperation between individuals, and that is indeed what we observe when we compare non-sanctioning environments with sanctioning environments. Independently on the monitor's payoff design, the possibility of punishment generates a statistical significant differences in contributions. Sanctioning opportunities are used to signal certain disagreement with the level of cooperation attained.

However, what deserves special attention in this work is the analysis of the impact of the payoff scheme on the behavior of sanctioners and contributors. Studying the sanctioners' behaviour, we notice that punishment is implemented more commonly under pool punishment, that is, when their payoff is contingent to the level of the public good. They do so as an attempt of improving their future endowment, and consequently, their future payoff. Additionally, contributors are more responsive to sanctions in this case, increasing their contributions to a larger extent in the following period if punished by a pool sanctioner. Nevertheless, overall, we observe that the scenario with fixed-endowment and punishment opportunities presents higher contributions and higher social welfare than any other scenario. This reveals that contributors notice the extrinsic motivation of pool punishers and present a lower willingness to cooperate. Hence, the social effect annihilates the sanctioning signal, which becomes noisy.

References

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

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

3. Egas, M., & Riedl, A. (2008). “The Economics of Altruistic Punishment and the Maintenance of Cooperation." *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1637), 871-878.

4. Fehr, E., & Gächter, S. (2000).“Experiments Cooperation and Punishment in Publics Experiments." *American Economic Review*, 90(4), 980-994.

5. Fehr, E., & Williams, T. (2017). “Creating an efficient culture of cooperation." *Mimeo.*

6. Gächter, S., Renner, E., & Sefton, M. (2008). “The Long-Run Benefits of Punishment." *Science*, 322(5907), 1510-1510.

7. Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). “Building the Leviathan?Voluntary centralisation of punishment power sustains cooperation in humans." *Scientific reports,* 6, 20767.

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

9. Kamijo, Y., Nihonsugi, T., Takeuchi, A.,& Funaki, Y. (2014). “Sustaining cooperation in social dilemmas: Comparison of centralized punishment institutions." *Games and Economic Behavior,* 84, 180-195.

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

11. 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.

12. Nikiforakis, N., & Normann, H.T. (2009). “A Comparative Statics Analysis of Punishment in Public-Good Experiments." *Experimental Economics*, 11(4), 358-369.

13. Ozono, H., Jin, N., Watabe, M., & Shimizu, K. (2016). “Solving the second-order free rider problem in a public goods game: An experiment using a leader support system." *Scientific reports,* 6, 38349.

14. Schoenmakers, S., Hilbe, C., Blasius, B., & Traulsen, A. (2014). “Sanctions as honest signals?the evolution of pool punishment by public sanctioning institutions." *Journal of theoretical biology,* 356, 36-46.

15. Sigmund, K., De Silva, H., Traulsen, A., & Hauert, C. (2010). “Social learning promotes institutions for governing the commons." *Nature*, 466(7308), 861-863.

16. Sutter, M., Haigner, S., & Kocher, M. G. (2010). “Choosing the carrot or the stick? Endogenous institutional choice in social dilemma

situations." *The Review of Economic Studies,* 77(4), 1540-1566.

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

Appendix 4.A. Instructions

Instructions for Treatments ONP and OP.

Welcome to the Experiment

The purpose of this experiment is to study how do individuals make decisions in particular contexts. Instructions are simple and if you follow them carefully, you will confidentially receive an amount of cash at the end of the experiment, given that nobody will know the payoffs received by the rest of participants. You can ask at any time the doubts you may have, raising first your hand. Beyond those questions, any type of communication between you is forbidden and subject to the immediate expulsion of the experiment.

What does this experiment consist of?

Este experimento comprises two parts:

Part 1 - Trial

During the first part of the experiment, you will be part of a group of **4 people**, including yourself. The group arrangement will be done **only once** at the beginning of the trial part and in a random way. In other words, you will be part of the same group during all this first part.

The trial part consists of 10 rounds.

In each of the rounds, the 4 Players of each group must make an **investment decision.** To this effect, each one will receive an endowment of 10 points and must decide how many points devote to the investment in a common project and how many points keep as savings. The points that each Player of the group decides to invest in the common project will be multiplied by 2/3.

After making the investment decision, all of you will observe the investment of each group member as shown in the following image.

[Screenshot.]

The profits for each Player in each round, will therefore be:

Profits: Saved points+ 2/3 Sum of the points invested by the group

This first part will take place on a trial bases and the profits will not be part of the experiment earnings.

Part 2 - Experiment

During the second part of the experiment, groups will randomly be arranged again. Now you will be part of a group of a total of **4 people**, including yourself. In each group of 4 people, **3 will be Type A Players and 1 will be Type B Player**. The members of your group will not necessarily be the same as in the trial part. The group arrangement and the types assignment will be done **only once** at the beginning of the experiment part and in a random way. In other words, you will be part of the same group and will be of the same type during all the experiment.

The experiment consists of 10 rounds.

In each round, the 3 **Type A Players must make an investment decision.** To this effect, each one will receive an endowment of 10 points and must decide how many points devote to the investment in a common project and how many points keep as savings. The points that each Player of the group decides to invest in the common project will be multiplied by 2/3.

The **Type B Players** will observe the investment decisions of the Type A Players in each round. The payoff that Type B players will receive will be displayed on screen.

After making the investment decision, all of you (Type A and Type B Players) will observe the investment of each group member as shown in the following image.

[Screenshot.]

The profits for each Type A Player in each round of this stage, will therefore be:

Profits: Saved points + 2/3 Sum of the points invested by the group

Before starting the second part of the experiment, you will be asked to answer some very brief comprehension questions.

The sum of the profits that you accumulate in the 10 rounds of the experiment will determine your earnings. Points will be changed to Euros at the end of the experiment according to the following relation:

15 points = 1 Euro

Once the experiment is over, you will be asked to answer some questionnaires, whose instructions you will observe on screen.

Instructions for Treatments SNP and SP (includes sanctioning stage).

Welcome to the Experiment

The purpose of this experiment is to study how do individuals make decisions in particular contexts. Instructions are simple and if you follow them carefully, you will confidentially receive an amount of cash at the end of the experiment, given that nobody will know the payoffs received by the rest of participants. You can ask at any time the doubts you may have, raising first your hand. Beyond those questions, any type of communication between you is forbidden and subject to the immediate expulsion of the experiment.

What does this experiment consist of?

Este experimento comprises two parts:

Part 1 - Trial

During the first part of the experiment, you will be part of a group of **4 people**, including yourself. The group arrangement will be done **only once** at the beginning of the trial part and in a random way. In other words, you will be part of the same group during all this first part.

The trial part consists of 10 rounds.

In each of the rounds, the 4 Players of each group must make an **in-**

vestment decision. To this effect, each one will receive an endowment of 10 points and must decide how many points devote to the investment in a common project and how many points keep as savings. The points that each Player of the group decides to invest in the common project will be multiplied by 2/3.

After making the investment decision, all of you will observe the investment of each group member as shown in the following image.

[Screenshot.]

The profits for each Player in each round, will therefore be:

Profits: Saved points+ 2/3 Sum of the points invested by the group

This first part will take place on a trial bases and the profits will not be part of the experiment earnings.

Part 2 - Experiment

During the second part of the experiment, groups will randomly be arranged again. Now you will be part of a group of a total of **4 people**, including yourself. In each group of 4 people, **3 will be Type A Players and 1 will be Type B Player**. The members of your group will not necessarily be the same as in the trial part. The group arrangement and the types assignment will be done only once at the beginning of the experiment part and in a random way. In other words, you will be part of the same group and will be of the same type during all the experiment.

The experiment consists of 10 rounds, each one of them composed by two stages.

Stage 1

In this stage, the 3 **Type A Players must make an investment decision**. To this effect, each one will receive an endowment of 10 points and must decide how many points devote to the investment in a common project and how many points keep as savings. The points that each Player of the group decides to invest in the common project will be multiplied by 2/3.

After making the investment decision, all of you (Type A and Type B Players) will observe the investment of each group member as shown in the following image.

[Screenshot]

The profits for each Type A Player in each round of this stage, will therefore be:

Profits Stage 1: Saved points + 2/3 Sum of the points invested by the group

Stage 2

In this stage, the **Type B players will receive an endowment to carry out the sanctions that they consider appropriate.** The endowment that he/she will receive will be displayed on screen. The Type B Player of each group will decide 1) if he/she wants to sanction and 2) which Type A Player or Players to sanction. For each Player he/she decides to sanction, 1 point will be subtracted from his/her endowment. Sanctioned Type A Players will pay a sanction equivalent to 3 points.

After making the sanction decision, everybody (Type A and Type B Players) will observe the decision made by the Type B Player of your group, as shown in the following image.

[Screenshot]

The total profits, which you will observe on screen after each round, will be:

Type B Player = Endowment - number of sanctioned Type A Players

Type A Player = Saved points + 2/3 Sum of the points invested by the group - 3 if the Type B Player has decided to sanction him/her.

Type A Player = Saved points + 2/3 Sum of the points invested by the group

if the Type B Player has decided **not** to sanction him/her.

Before starting the second part of the experiment, you will be asked to answer some very brief comprehension questions.

The sum of the profits that you accumulate in the 10 rounds of the experiment will determine your earnings. Points will be changed to Euros at the end of the experiment according to the following relation:

15 points = 1 Euro

Once the experiment is over, you will be asked to answer some questionnaires, whose instructions you will observe on screen.

Appendix 4.B. Tables

Treatment Effect in contributions

Table 4.7. displays Ordinary Least Squares (OLS) linear regressions. Table 4.8 shows Linear Mixed Effects (LME) models fit by maximum likelihood with Subject ID and Group as random effects. t-tests use Satterthwaite approximations to degrees of freedom. The dependent variable was Contribution$_t$.

	A	**B**	**C**
Treatment	1.4373 ***	-	1.4373 ***
	(0.1257)		(0.1243)
Pool	-	-0.94365***	-0.9437 ***
		(0.12754)	(0.1243)
Adjusted R^2	0.04899	0.02089	**0.0699**
AIC	12946.25	13019.62	**12891.2**

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.7: OLS regressions of treatment effect

	A	**B**	**C**
Treatment	1.4237 ***	-	1.4232 ***
	(0.1205)		(0.1205)
Pool	-	-1.1003*	-1.0889 *
		(0.3919)	(0.3913)
AIC	12728.4	12858.1	**12724.3**

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.8: LME regressions of treatment effect

Contributing behaviour in treatments with punishment

Table 4.9. displays Ordinary Least Squares (OLS) linear regressions. Table 4.10 shows Linear Mixed Effects (LME) models fit by maximum likelihood with Subject ID as a random effect. t-tests use Satterthwaite approximations to degrees of freedom. The dependent variable was $Contribution_t$.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Contribution$_{t-1}$	0.35048 ***	-	0.53087 ***	-	0.35076 ***	0.35010***	0.34957 ***	0.35222***	0.353181***
	(0.027747)		(0.02571)		(0.02771)	(0.02707)	(0.02523)	(0.02511)	(0.025135)
Avg. contribution$_{t-1}$	0.44100***	-	-	0.65065***	0.43932***	0.44520***	0.44573***	0.45057***	0.452194***
	(0.03355)			(0.03095)	(0.03348)	(0.03280)	(0.03134)	(0.03101)	(0.031071)
Diff with avg. contr.$_{t-1}$	-	0.09928**	-	-	-	-	-	-	-
		(0.03187)							
Received fine$_{t-1}$	0.08211	0.12205	1.35603***	-1.37611***	0.07346	0.01318	-	-	-
	(0.28301)	(0.34807)	(0.28356)	(0.27428)	(0.28271)	(0.24311)			
Group Fines$_{t-1}$	0.26644*	-0.04443	-0.26249 .	0.70691***	0.26346*	-	-	-	-
	(0.13354)	(0.16354)	(0.13578)	(0.13683)	(0.13343)				
Group Fines$_{t-1}$R	-	-	-	-	-	0.87377**	0.88210***	0.86336***	0.897706***
						(0.26776)	(0.21908)	(0.021839)	(0.222063)
Round	-0.10684***	-0.13181***	-0.11690***	-0.10915***	-0.10687***	-0.11289***	-0.11294***	-0.11253***	-0.207948 .
	(0.02518)	(0.03096)	(0.02684)	(0.02673)	(0.02517)	(0.02517)	(0.02515)	(0.02515)	(0.114166)
Round^{2}	-	-	-	-	-	-	-	-	0.008656
									(0.010102)
Pool	-0.18007	-1.00126***	-0.42744 **	-0.35295*	-0.16239	-0.15962	-0.15932	-	-
	(0.15090)	(0.18127)	(0.15966)	(0.15951)	(0.14973)	(0.14898)	(0.14881)		
Age	-0.02754	-0.00697	-0.01132	-0.03487	-	-	-	-	-
	(0.02940)	(0.03615)	(0.03133)	(0.03120)					
Gender	-0.03714	0.24398	0.04499	0.02502	-	-	-	-	-
	(0.14583)	(0.17883)	(0.15537)	(0.15470)					
Adjusted R^{2}	0.3673	0.043	0.2805	0.2872	0.3679	0.3712	0.3717	**0.3717**	0.3715
AIC	5960.439	6480.909	6121.439	6109.719	5957.386	5950.637	5948.64	**5947.791**	5949.053

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.9: OLS regressions of contributing behaviour in treatments with punishment

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Contribution$_{t-1}$	0.3146 ***	-	0.4616 ***	-	0.31449 ***	0.31554***	0.31250 ***	0.31696***	0.3180***
	(0.02776)		(0.02643)		(0.02776)	(0.02712)	(0.02540)	(0.02530)	(0.02532)
Avg. contribution$_{t-1}$	0.4353***	-	-	0.59078***	0.43473***	0.44007***	0.44330***	0.45046***	0.4523***
	(0.03407)			(0.03229)	(0.03400)	(0.03340)	(0.03203)	(0.03165)	(0.03171)
Diff with avg. contr.$_{t-1}$	-	0.05789 .	-	-	-	-	-	-	-
		(0.02979)							
Received fine$_{t-1}$	0.1147	0.12850	1.263***	-1.04327***	0.11181	0.08156	-	-	-
	(0.2789)	(0.32101)	(0.2793)	(0.26616)	(0.27884)	(0.24030)			
Group Fines$_{t-1}$	0.2917*	0.07122	-0.1911	0.65900***	0.29244*	-	-	-	-
	(0.1322)	(0.15205)	(0.1340)	(0.13308)	(0.13217)				
Group Fines$_{t-1}$R	-	-	-	-	-	0.89696***	0.94772***	0.92045***	0.9539***
						(0.26433)	(0.21743)	(0.21687)	(0.2206)
Round	-0.1082***	-0.13197***	-0.1190***	-0.11150***	-0.10821***	-0.11432***	-0.11461***	-0.11401***	-0.2044 .
	(0.02462)	(0.02812)	(0.02599)	(0.02548)	(0.02462)	(0.02463)	(0.02462)	(0.02465)	(0.1120)
Round^2	-	-	-	-	-	-	-	-	0.008199
									(0.00907)
Pool	-0.2376	-1.11562***	-0.5421 **	-0.43632*	-0.23492	-0.23174	-0.22987	-	-
	(0.1638)	(0.20071)	(0.1765)	(0.17828)	(0.16257)	(0.16161)	(0.16148)		
Age	-0.0005194	0.09371*	0.03319	0.02024	-	-	-	-	-
	(0.03470)	(0.04733)	(0.03866)	(0.03978)					
Gender	-0.07749	0.05797	-0.008351	-0.06467	-	-	-	-	-
	(0.1673)	(0.22209)	(0.1849)	(0.18866)					
AIC	5950.7	6345.8	6100.7	6063.4	5946.9	5940.3	**5938.4**	5938.5	5939.8

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.10: LME regressions of contributing behaviour in treatments with punishment

Responsiveness to punishment

Table 4.11. displays Ordinary Least Squares (OLS) linear regressions. Table 4.12 shows Linear Mixed Effects (LME) models fit by maximum likelihood with Subject ID as a random effect. t-tests use Satterthwaite approximations to degrees of freedom. The dependent variable was Contribution$_t$ - Contribution$_{t-1}$. The independent variable was Received fine$_{t-1}$ in Models A, Group fines $_{t-1}$ in Models B and Group Fines$_{t-1}$R in Models C.

	A	**B**	**C**
	Received fine$_{t-1}$	Group fines $_{t-1}$	Group Fines$_{t-1}$R
Not Pool	1.8643 ***	0.3108 .	0.7182
	(0.3562)	(0.1800)	(0.4450)
Pool	2.1709***	0.3782*	1.1562 ***
	(0.2958)	(0.1525)	(0.3389)
Global	2.0459***	0.3593**	1.01113***
	(0.2268)	(0.1151)	(0.26813)
Adjusted R^2	**0.0403/0.0776/0.0599**	0.0031/0.0081/0.0069	0.0025/0.01663/0.0104
AIC	**3234.74/3205.32/6434.88**	3258.66/3251.03/6504.12	3259.03/3245.61/6499.67

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.11: OLS regressions of responsiveness to punishment

	A	**B**	**C**
	Received fine$_{t-1}$	Group fines $_{t-1}$	Group Fines$_{t-1}$R
Not Pool	1.8643 ***	0.3108 .	0.7182
	(0.3556)	(0.1797)	(0.4443)
Pool	2.1709***	0.3782*	1.1562 ***
	(0.2953)	(0.1522)	(0.3384)
Global	2.04589***	0.3593**	1.01113***
	(0.22665)	(0.1150)	(0.26792)
AIC	**3236.7/3207.3/6436.9**	3260.7/3253.0/6506.1	3261.0/3247.6/6501.7

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.12: LME regressions of responsiveness to punishment

Contributing behaviour in treatments without punishment

Table 4.13. displays Ordinary Least Squares (OLS) linear regressions. Table 4.14 shows Linear Mixed Effects (LME) models fit by maximum likelihood with Subject ID as a random effect. t-tests use Satterthwaite approximations to degrees of freedom. The dependent variable was Contribution$_t$.

	A	**B**	**C**	**D**	**E**	**F**	**G**
Contribution$_{t-1}$	0.3882 ***	-	0.4514 ***	-	0.3914 ***	0.3887***	0.3889***
	(0.0244)		(0.0243)		(0.0244)	(0.0244)	(0.02436)
Avg. contr. $_{t-1}$	0.3078***	-	-	0.4465***	0.3002***	0.3071***	0.308***
	(0.0322)			(0.034)	(0.032)	(0.0321)	(0.0321)
Diff avg. contr.$_{t-1}$	-	0.1675***	-	-	-	-	-
		(0.025)					
Round	-0.1393***	-0.2953***	-0.1949***	-0.1934***	-0.1403***	-0.1393***	0.002
	(0.0259)	(0.0283)	(0.0261)	(0.0281)	(0.0258)	(0.0258)	(0.1102)
Round^2	-	-	-	-	-	-	-0.0128
	-	-	-	-	-	-	(0.0097)
Pool	-0.321*	-0.878***	-0.5185 **	-0.5169**	-0.322*	-0.3346*	-0.3337*
	(0.1451)	(0.1637)	(0.1488)	(0.1585)	(0.1442)	(0.1442)	(0.1441)
Age	0.0427*	0.0178	0.021	0.0623*	-	0.0423*	0.0424*
	(0.0213)	(0.0244)	(0.0219)	(0.0233)		(0.0212)	(0.0212)
Gender	-0.0399	0.2019	0.0477	0.0409	-	-	-
	(0.1449)	(0.1659)	(0.1497)	(0.1587)			
Adjusted R^2	0.3374	0.1256	0.2896	0.2041	0.3364	**0.338**	0.3384
AIC	5904.784	6253.319	5991.535	6134.794	5903.638	**5901.654**	5901.907

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.13: OLS regressions of contributing behaviour in treatments without punishment

	A	**B**	**C**	**D**	**E**	**F**	**G**
Contribution$_{t-1}$	0.3304 ***	-	0.3815 ***	-	0.3347 ***	0.331***	0.3312***
	(0.025)		(0.0250)		(0.0249)	(0.0249)	(0.0248)
Avg. contr.$_{t-1}$	0.3089***	-	-	0.3954***	0.304***	0.3096***	0.3105***
	(0.0326)			(0.0339)	(0.0325)	(0.0326)	(0.0326)
Diff avg. contr.$_{t-1}$	-	0.0996***	-	-	-	-	-
		(0.0242)					
Round	-0.1518***	-0.2949***	-0.2103***	-0.205***	-0.1519***	-0.1515***	-0.013
	(0.0252)	(0.026)	(0.0253)	(0.0261)	(0.0253)	(0.0252)	(0.1072)
Round^2	-	-	-	-	-	-	-0.0126
	-	-	-	-	-	-	(0.0095)
Pool	-0.4485**	-1.1964***	-0.7092 ***	-0.7498***	-0.4405**	-0.4546**	-0.4535**
	(0.1633)	(0.1895)	(0.169)	(0.1836)	(0.1621)	(0.1623)	(0.1622)
Age	0.0488*	0.0612 .	0.0315	0.0872**	-	0.0481 .	0.0481 .
	(0.0245)	(0.0323)	(0.0259)	(0.03)		(0.0245)	(0.0245)
Gender	-0.0413	0.3772*	0.1399	0.2227	-	-	-
	(0.1624)	(0.192)	(0.1698)	(0.1827)			
AIC	5887.3	6144.4	5971.6	6033.5	5885.7	**5883.8**	5884.1

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.14: LME regressions of contributing behaviour in treatments without punishment

Probability of being punished

Table 4.15. displays Generalized Linear Model (GLM) regressions with a logit link function. Table 4.16 shows linear Generalized Linear Mixed Effects (GLME) models fit by maximum likelihood (Laplace Approximation) with a logit link function and with Subject ID as a random effect. The dependent variable was the probability of Received fine$_t$.

Contribution Effect in the probability of being punished

Table 4.17. displays Generalized Linear Model (GLM) regressions with a logit link function. Table 4.18. shows linear Generalized Linear Mixed Effects (GLME) models fit by maximum likelihood (Laplace Approximation) with a logit link function and with Subject ID as a random effect. The dependent variable was the probability of Received fine$_t$. The independent variable is Contribution$_t$ in Models A, Average contribution$_t$ in Models B and Difference with average contribution$_t$ in Models C.

	A	**B**	**C**
	Contribution$_t$	Avg. contribution$_t$	Diff. with avg. contribution$_t$
Not Pool	-0.19287 ***	0.02265	-0.24717 ***
	(0.03313)	(0.03879)	(0.03783)
Pool	-0.35452***	-0.009505	-0.41924 ***
	(0.03725)	(0.036083)	(0.04444)
AIC	548.63/589.65	584.06/702.41	**536.26/578.15**

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.17: GLM regressions of contribution effect on probability of being punished

	A	**B**	**C**	**D**	**E**	**F**	**G**
Contribution$_t$	-0.29505 ***	-	-0.22690 ***	-	-0.29572 ***	-0.29527***	-0.294758***
	(0.03185)		(0.02887)		(0.03184)	(0.03182)	(0.031837)
Avg. contribution$_t$	0.47772***	-	-	0.39151***	0.47518***	0.47268***	0.473128***
	(0.04907)			(0.04602)	(0.04827)	(0.04842)	(0.048430)
Diff with avg. contr.$_t$	-	-0.339398***	-	-	-	-	-
		(0.029330)					
Group fines$_{t-1}$	0.09693	0.106830	0.14692	0.09535	-	-	-
	(0.10173)	(0.101254)	(0.09554)	(0.09615)	-	-	-
Group fines$_{t-1}$ R	-	-	-	-	-	0.37271	0.391343 .
						(0.23205)	(0.235822)
Round	-0.06543*	-0.079480**	-0.09854***	-0.04794 .	-0.06433*	-0.06855*	-0.122775
	(0.02861)	(0.028194)	(0.02709)	(0.02720)	(0.02847)	(0.02874)	(0.124491)
Round^2	-	-	-	-	-	-	0.004987
							(0.011140)
Pool	0.30546	0.110338	-0.02581	0.65005***	0.32674 .	0.29398	0.293350
	(0.19026)	(0.177526)	(0.16816)	(0.19191)	(0.18846)	(0.19001)	(0.190054)
Sheriff's age	-0.01656	0.007237	0.04776	-0.02927	-	-	-
	(0.03547)	(0.035406)	(0.03356)	(0.03321)			
Sheriff's gender	0.52360**	0.497691**	0.42296**	0.48380**	0.54713***	0.50629**	0.504291**
	(0.16829)	(0.166922)	(0.15652)	(0.15763)	(0.16623)	(0.16849)	(0.168549)
Sheriff's payoff$_t$	-0.32752***	-0.241998***	-0.12239***	-0.40290***	-0.32571***	-0.32578***	-0.326030***
	(0.03549)	(0.026106)	(0.02674)	(0.03344)	(0.03508)	(0.03519)	(0.035205)
AIC	996.27	1008.3	1110.5	1089.7	993.35	**992.82**	994.62

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.15: GLM regressions of probability of being punished

	A	**B**	**C**	**D**	**E**	**F**	**G**
Contribution$_t$	-0.29547 ***	-	-0.23955 ***	-	-0.29609 ***	-0.29628***	-0.295962***
	(0.03342)		(0.03089)		(0.03340)	(0.03336)	(0.03344)
Avg. contribution$_t$	0.47973***	-	-	0.40854***	0.47586***	0.47431***	0.47597***
	(0.05205)			(0.04966)	(0.05151)	(0.05153)	(0.05153)
Diff with avg. contr.$_t$	-	-0.340720***	-	-	-	-	-
		(0.030712)					
Group fines$_{t-1}$	0.01275	0.020838	0.04166	-0.01606	-	-	-
	(0.10899)	(0.108488)	(0.10335)	(0.10389)	-	-	-
Group fines$_{t-1}$ R	-	-	-	-	-	0.26793	-
						(0.24164)	
Round	-0.06822*	-0.084895**	-0.11024***	-0.04762 .	-0.06849*	-0.07164*	-0.077965
	(0.02952)	(0.029071)	(0.02833)	(0.02813)	(0.02948)	(0.02970)	(0.12526)
Round^2	-	-	-	-	-	-	0.0008786
							(0.01127)
Pool	0.33138	0.132957	-0.05309	0.64456**	0.33225	0.30641	0.332391
	(0.21353)	(0.203314)	(0.19983)	(0.21428)	(0.21172)	(0.21248)	(0.21171)
Sheriff's age	-0.02355	-0.001242	0.03014	-0.04015	-	-	-
	(0.04128)	(0.041737)	(0.04131)	(0.03980)			
Sheriff's gender	0.48296*	0.439343*	0.33466 .	0.44728 *	0.48472*	0.45354*	0.4847013*
	(0.19257)	(0.192995)	(0.18790)	(0.18411)	(0.19078)	(0.19238)	(0.19076)
Sheriff's payoff$_t$	-0.36783***	-0.286081***	-0.17349***	-0.45184***	-0.36565***	-0.36449***	-0.365688***
	(0.04064)	(0.031877)	(0.03158)	(0.03881)	(0.04026)	(0.04030)	(0.04026)
AIC	987.9	998.0	1086.9	1071.7	**984.2**	985.0	986.2

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.16: GLME regressions of probability of being punished

	A	**B**	**C**
	Contribution$_t$	Avg. contribution$_t$	Diff. with avg. contribution$_t$
Not Pool	-0.22000 ***	0.009984***	-0.253387 ***
	(0.03765)	(0.001612)	(0.001751)
Pool	-0.417233***	-0.001501	-0.42398 ***
	(0.002144)	(0.042293)	(0.04611)
AIC	533.0/577.9	570.7/693.0	**526.8/577.1**

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.18: GLME regressions of contribution effect on probability of being punished

Probability of the sanctioner punishing

Table 4.19. displays Ordered Probit regressions where the dependent variable was the probability of Group fines$_t$. Table 4.20 shows Multivariate Multiple Regressions where the dependent variable was the Probability of (Group fines$_t$, 3-Group fines$_t$).

	A	**B**	**C**
Avg. contribution$_t$	0.156575 ***	0.141820***	0.148379 ***
	(0.028928)	(0.027121)	(0.026518)
Group fines$_{t-1}$	0.593531***	0.606839***	-
	(0.070017)	(0.069436)	-
Group fines$_{t-1}$ R	-	-	1.037471 ***
			(0.164669)
Round	-0.096756***	-0.097530***	-0.090883 ***
	(0.019287)	(0.019219)	(0.018660)
Pool	0.151712	-	-
	(0.114008)		
Sheriff's age	-0.024479	-	-
	(0.023184)		
Sheriff's gender	0.060799	-	-
	(0.104549)		
Sheriff's payoff	-0.165835***	-0.165772***	-0.165488 ***
	(0.021417)	(0.021163)	(0.020659)
0|1	-1.333918*	-1.025910***	-1.094801 ***
	(0.532077)	(0.254009)	(0.248753)
1|2	-0.509897	-0.203835	-0.306816
	(0.532182)	(0.253279)	(0.247687)
2|3	0.362294	0.666835*	0.539175 *
	(0.537959)	(0.266732)	(0.260840)
AIC	981.1695	**978.2156**	1014.708

. p<0.1; * p < 0.05; ** p<0.01; *** p<0.001

Table 4.19: Ordered probit regressions of the probability of the sanctioner punishing

	A	**B**	**C**
Avg. contribution$_t$	0.056368 **	0.05538**	0.06253 ***
	(0.019173)	(0.01900)	(0.01885)
Group fines$_{t-1}$	0.324934***	0.32858***	-
	(0.058413)	(0.05799)	-
Group fines$_{t-1}$ R	-	-	0.61929 ***
			(0.14318)
Round	-0.033377*	-0.03367*	-0.03479 **
	(0.013402)	(0.01338)	(0.01336)
Pool	0.032930	-	-
	(0.079654)		
Sheriff's age	-0.002716	-	-
	(0.017354)		
Sheriff's gender	0.043212	-	-
	(0.076387)		
Sheriff's payoff	-0.068618***	-0.06945***	-0.07299 ***
	(0.014177)	(0.01398)	(0.01393)
AIC	1775.2	**1769.7**	1783.1

. $p<0.1$; * $p < 0.05$; ** $p<0.01$; *** $p<0.001$

Table 4.20: Multivariate multiple regressions of the probability of the sanctioner punishing

Chapter 5

Coordinated vs. uncoordinated punishment

5.1 Introduction

In the context of the free rider issue present in many daily situations, punishment is the most standard way of smoothing the problem. Punishment can be implemented both formally, with penalty fees, or informally, through social punishment such as ostracism or hostility. Focusing on formal economic sanctions, there are several ways in which the sanctions can have an impact on the punished agent. In this paper, we theoretically compare two different punishment schemes in a team-trust game with information asymmetries: (i) an uncoordinated punishment system where individual punishment destroys the punished agent's payoff and (ii) a coordinated punishment system where it is necessary that the number of individuals willing to carry out punishment exceeds a particular threshold for the punishment to be effective. Our results reveal that only if the proportion of reciprocators in the population is sufficiently high, a coordinated punishment system leads to efficient equilibria in a wider range of cases than uncoordinated punishment. For low and intermediate proportions of reciprocators, coordinated and uncoordinated punishment show no significant differences in terms of efficiency.

The opportunity to sanction, even if it entails a cost, has been proven to have a positive impact on mitigating the free rider issue (Fehr and Gächter, 2000; Gürerk, Irlenbusch and Rockenbach 2006; Gächter, Renner and Sefton, 2008). This happens because society is conformed both by selfish agents who maximize their material payoffs, and by agents with some kind of social concern such as fairness, reciprocity or altruism. The presence of this second type of agents enhances both cooperation and punishment of undesirable behaviour. However, punishment literature has traditionally considered individual uncoordinated punishment. That is, at the end of the game, agents endowed with sanctioning power individually decide whether to punish free riding behaviour and such decisions have a negative impact on the payoff of the deceiver.

Nonetheless, this uncoordinated type of punishment is, at times, questionable for two reasons. In first place, as the number of agents willing to punish a free rider increases, the individual cost of carrying out punishing should fall. This reflects the existence of increasing returns to scale. Consider, for instance, how the cost of a social claim diminishes if someone has already made way before. It is always more costly for the first person to raise a social demand than for the last one to join the cause.

The second dubious aspect of standard uncoordinated punishment is that it assumes it inflicts damage right from the first punisher. Coming back to the social claim example, it is probable that the first person raising his voice was not heard and that a massive movement was needed for it to be effective. Another example are strikes, where a sufficiently large group going on strike is required for the firm to effectively have their profits shrunk. In this line, coordinated punishment captures these two characteristics: increasing returns to scale and a minimum participation in punishment actions for payoffs to be destroyed.

Our aim is to carry out a theoretical comparison between two punishment systems, uncoordinated and coordinated, with increasing returns to scale, where the difference lies on the effectiveness of punishment. To do so, we present a team trust game with two investors and one allocator. Both types of players can either have standard preferences (selfish investors and profit-maximiser allocators) or social concerns (reciprocal investors and fair-minded allocators), being types private information. In the first stage of the game, the two investors individually and simultaneously decide whether or not to invest in a joint project. Only if both decide to invest, a positive surplus is generated and the team's allocator must decide how much of such surplus keep for himself and how much return to the investors. Once these investors observe their return, they can implement punishment. For the game with uncoordinated punishment, the allocator has a proportion of his payoff destroyed as soon as one of the two investors decides to punish him. For the coordinated punishment case, however, it is necessary that both investors decide to punish the allocator for such decision to have a negative impact on the allocator's payoff.

To the best of our knowledge, the first paper dealing with coordinated punishment was by Boyd, Gintis and Bowles (2010), who analyse it in a public goods game. In this work, authors present increasing returns to scale in the sense that he marginal cost of punishment falls as the number of punishers increases. However, punishment is equally effective whatever the number of participants is. In this work, beyond making the marginal cost of punishment depend on the number of punishers, we also consider possible increasing returns to scale in the impact of punishment. In this sense, if both investors punish, coordinated punishment must be as least as effective in reducing inequality as uncoordinated punishment.

In this paper we follow the model by Calabuig and Olcina (2015), a team trust game with increasing returns to scale and an unequal effectiveness of punishment.[1] Their main result is that cooperation only evolves and is maintained if there is enough punishment capacity in society and there are enough individuals willing to implement punishment. The fully cooperative equilibrium is achieved under the threat of effective coordinated punishment, which is supported by the presence of a high proportion of reciprocators in the population. However, authors only provide results for coordinated punishment. Our goal in this work is to compare coordinated punishment with uncoordinated punishment in terms of joint investment, returns and punishment actions, in order to highlight under which conditions is one system better than the other.

[1] In their work, they also introduce a peer punishment stage at the end of the game, which we do not consider here.

Our results emphasize that only when the proportion of reciprocal investors is high enough, coordinated punishment is superior with regard to uncoordinated punishment in enhancing cooperation. This is evidenced with a greater range of equilibria with joint investment under a coordinated punishment system, along with higher rewards from the allocators, even if the proportion of fair-minded allocators in the population is relatively low. The reason for this is that investors under this scheme are more demanding with the rewards they receive, implementing punishment for a wider range of returns than investors under uncoordinated punishment. This, in turn, entails a stronger punishment threat to profit-maximiser allocators. With low and intermediate proportions of reciprocators, however, differences between coordinated and uncoordinated punishment are minimal. As we will later prove, the main reason behind these results is that coordinated punishment is more effective in solving the free rider problem among reciprocators.

The rest of this paper is organized as followed. In section 5.2 we present the model to analyse, describing as well the different preferences considered. Section 5.3 collects the results of the model, stage by stage and, in last place, Section 5.4 presents concluding remarks.

5.2 The model

5.2.1 Team trust game with punishment

Consider a 3-player team trust game with two investors and one allocator.

In the first stage of this game, *the investment phase*, the two investors simultaneously and independently decide whether to invest (I) or not invest (NI) in a project which can generate a surplus of 2 only if both of them decide to carry out the investment. In other words, only the combination of actions (I, I) leads to the project being successful but if at least one investor decides not to invest, the game ends and all players obtain a payoff of 0. Investing is a costly action for the investors. In particular, it entails a cost $c \in (0, 1/2)$ regardless of the success or failure of the project. Not investing, on the other side, is costless.

In the second stage, *the rewarding phase*, the allocator decides how much of the surplus generated he is willing to return to the investors. With this end, the allocator decides the payoff b, the reward to each of the investors ($0 \leq b \leq 1$), being this symmetric to both of them.[2] Recall this stage is only reached if both investors decide to invest in the first stage.

In the third stage, *the punishment phase*, the investors observe their reward b for their investment and simultaneously and independently decide whether to punish (p) or not to punish (np) the allocator. We consider that punishing has a cost of z/n for each punishing investor, where z is the total

[2] In this paper we focus on symmetric equilibria.

cost and n is the number of investors who punish. Not punishing, on the other side, is costless.

The impact of punishment on the allocator's payoff depends on the punishment scheme we consider. Under a coordinated punishment scheme, only if both investors coordinate to punish, the allocator will get his payoff destroyed in a proportion λ_c, where $0 \leq \lambda_c \leq 1$. Under an uncoordinated punishment scheme, however, the individual punishment action has an impact of λ_u on the allocator's payoff, where $0 \leq \lambda_u \leq 1$, regardless of whether the investors coordinate in punishing or not. If they do so, the impact will be $2\lambda_u$ instead. Table 5.1 resumes these details.

	Coordinated	Uncoordinated
p, p	λ_c	$2\ \lambda_u$
np, p	0	λ_u
np, np	0	0

Table 5.1: Impact of punishment

Given that only if both investors invest, the game goes on to the reward and punishment stage, in the following we will only present the final payoff functions for the (I, I) subgame.

The material payoffs for each of the investors will be determined by:

$$\pi_I = b - c - z/n \qquad \text{with punishment} \tag{5.1}$$

The material payoffs of the allocator will depend on the punishment scheme, which can either be coordinated (CP) or uncoordinated (UP):

$$\pi_A(CP) = \begin{cases} 2(1-b)(1-\lambda_c) & \text{if both punish} \\ 2(1-b) & \text{otherwise} \end{cases} \tag{5.2}$$

$$\pi_A(UP) = \begin{cases} 2(1-b)(1-2\lambda_u) & \text{if both punish} \\ 2(1-b)(1-\lambda_u) & \text{only one punishes} \\ 2(1-b) & \text{if nobody punishes} \end{cases} \tag{5.3}$$

Suppose there is complete information and that all players have selfish preferences. By backward induction, selfish investors will never implement costly punishment, np, the selfish allocator will keep all of the surplus for himself and return $b = 0$ and investors will decide not to carry out the costly investment (NI, NI). This, however, will lead to an inefficient Perfect Equilibrium where every player receives a payoff of 0.

5.2.2 Social preferences

Experimental evidence has continuously proved that individual preferences only follow the standard selfishness assumption in a limited number of times (Fehr and Schmidt, 1999; Fehr and Gächter, 2000). For this reason, we consider necessary to introduce social preferences into our model.

On the side of the investors, we consider a certain proportion q of the population from which the two investors are drawn to be reciprocal investors, where we understand as reciprocal the willingness to punish hostile behaviour. That is, if the allocator returns an unfair offer, a reciprocal investor will implement punishment in the last stage. Given that the surplus that each investor generates is of size 1, a reciprocal investor will consider unfair any return below 1/2. We capture the disutility derived from unfair returns with a parameter $\alpha \geq 1$, measuring the proportional distance from the allocator's payoff to the investor's reward. This approach follows Fehr and Schmidt (1999) for disadvantageous inequality.[3] The remaining $1-q$ investors of the population will be considered selfish, with the payoff function formerly proposed.

The payoff function for a reciprocal investors in the subgame (I, I) if he receives a reward $b < 1/2$, will be determined by:

$$\pi_I(CP) = \begin{cases} b - c - z/2 - \alpha[(1-b)(1-\lambda_c) - b] & \text{if both punish} \\ b - c - z - \alpha[(1-b) - b] & \text{if only he punishes} \\ b - c - \alpha[(1-b) - b] & \text{otherwise} \end{cases} \tag{5.4}$$

$$\pi_I(UP) = \begin{cases} b - c - z/2 - \alpha[(1-b)(1-2\lambda_u) - b] & \text{if both punish} \\ b - c - z - \alpha[(1-b)(1-\lambda_u) - b] & \text{if only he punishes} \\ b - c - \alpha[(1-b)(1-\lambda_u) - b] & \text{if only the other one punishes} \\ b - c - \alpha[(1-b) - b] & \text{if nobody punishes} \end{cases} \tag{5.5}$$

On the side of the allocator, we consider a certain proportion m of the population from which the allocator is drawn to be fair-minded allocators,

[3] We assume that investors do not have a disutility for generous rewards above 1/2, which reflects Fehr and Schmidt's (1999) advantageous inequality.

where we understand as fair-minded having as a dominant strategy returning the fair reward $b = 1/2$. The remaining $(1 - m)$ are profit maximisers.

5.2.3 Assumptions

Before moving on to resolving the sequential game by backward induction, we need to make three assumptions on the relationship between the parameters that characterize the punishing institutions. These ensure that punishment is chosen at least under some circumstances and that coordinated punishment has a greater impact than uncoordinated punishment.

Assumption 1: $2\lambda_u \leq \lambda_c$

Assumption 1 states that if both investors punish, coordinated punishment must be at least as destructive for the allocator than uncoordinated punishment. This reflects the nature of coordinated actions, which can lead to synergies and economies of scale.

Assumption 2: $\alpha\lambda_u(\frac{1}{2-2\lambda_u}) \leq z \leq \alpha\lambda_u$

The upper bound is necessary to guarantee that the behaviour of a reciprocal investor differs from the behaviour of a selfish one. Briefly, by punishing, a reciprocal investor reduces his disadvantageous inequality with respect to the allocator in $\alpha\lambda_i$. Thus, this effect must compensate the cost z of punishing the allocator, such that he indeed does so. Notice that, by assumption 1, if $z \leq \alpha\lambda_u$, then $z \leq \alpha\lambda_c$. The lower bound, on the other hand, is a simplifying assumption in order to avoid numerous non-interesting subcases arising. Besides that, it is plausible to impose a lower bound to z, as punishment must be a costly action.

Assumption 3: $\lambda_c \geq 1/2$

Finally, assumption 3 is necessary for coordinated punishment to have a sufficient impact on the behaviour of the allocator. If $\lambda_c < 1/2$, the allocator would prefer to offer a reward of $b = 0$ instead of a fair reward of $b = 1/2$, even if he knew that the investors were going to punish him. Notice that assumptions 1 and 3 imply $\lambda_u \leq 1/2$.

5.3 Results

In this section we derive the Perfect Bayesian Equilibria (PBE) of the two sequential games: the team trust game with coordinated punishment and the team trust game with uncoordinated punishment. To do so, we will characterize behaviour in each subgame, anticipating what will happen in the following stages.

5.3.1 Punishment stage

In this last stage, investors must decide whether to punish the allocator for the reward b received. Given that types are private information, we define μ as the updated probability of facing a reciprocal investor after observing (I, I) in the investment stage. That is, $\mu = Prob(r/(I, I))$, where r stands for a reciprocal type. Any punishment subgame is characterized by a belief μ and a reward b, thus, we denote this subgame by $CP(\mu, b)$ for the coordinated punishment scenario and $UP(\mu, b)$ for the uncoordinated one. We represent the symmetric Bayesian Nash Equilibria (BNE) of this subgame by profiles (x, y), where the first term represents the action of the reciprocal type and the second the action of the selfish type.

First, notice that if the allocator returns a fair or a generous reward $b \geq 1/2$, no type of investor will punish. In the following lemmatas, we present the solution of $CP(\mu, b)$ and $UP(\mu, b)$ for unfair rewards ($b < 1/2$).

***Lemma 1**: The solution of any subgame $CP(\mu, b)$ with unfair rewards is:*

a) If $0 \leq b < b_1^c$, then (p, np) if $\mu > \bar{\mu}$ and (np, np) otherwise

b) If $b_1^c \leq b < b''$, then (p, np) if $\mu > \bar{\bar{\mu}}$ and (np, np) otherwise

c) If $b'' \leq b < 1/2$, then (np, np) $\forall \mu$

where $b_1^c = \frac{1-\lambda_c}{2-\lambda_c}, b'' = \frac{1}{2} - \frac{z}{4\alpha}$, $\bar{\mu} = \frac{z}{z/2+\alpha\lambda_c(1-b)}$ *and* $\bar{\bar{\mu}} = \frac{z}{z/2+\alpha(1-2b)}$

***Lemma 2**: The solution of any subgame $UP(\mu, b)$ with unfair rewards is:*

a) If $0 \leq b < \hat{b}$, then (p, np) $\forall \mu$

b) If $\hat{b} \leq b < b_1^u$, then (p, np) if $\mu > \tilde{\mu}$ and (np, np) otherwise

c) If $b_1^u \leq b < b^{**}$, then (p, np) if $\mu > \tilde{\tilde{\mu}}$ and (np, np) otherwise

d) If $b^{**} \leq b < 1/2$, then (np, np) $\forall \mu$

where $\hat{b} = 1 - \frac{z}{\alpha\lambda_u}$, $b_1^u = \frac{1-2\lambda_u}{2-2\lambda_u}$, $b^{**} = \frac{\alpha(1-\lambda_u)-z/2}{\alpha(2-\lambda_u)}$, $\tilde{\mu} = \frac{2z-2\alpha\lambda_u(1-b)}{z}$ *and* $\tilde{\tilde{\mu}} = \frac{z-\alpha\lambda_u(1-b)}{z/2-2\alpha\lambda_u(1-b)+\alpha(1-2b)}$

Selfish investors have as a dominant strategy not to punish, whereas reciprocal investors may have incentives to do so. Notice that for an intermediate range of values of the reward, reciprocal investors need that beliefs of being paired with another reciprocal investor are high enough in order to punish. Otherwise it does not compensate to undertake the punishment costs by themselves. Additionally, as rewards increase, even though they don't reach the fair reward of $b = 1/2$, reciprocators stop punishing if they receive a high enough reward for their investment. Under uncoordinated punishment, however, there is also a lower reward segment $(0 \leq b < \hat{b})$ which reciprocal investors punish independently of the beliefs. In other words, the reward is so low that they punish it in order to reduce inequality. This, nevertheless, does not happen with coordinated punishment given that individual punishment does not lead to a destruction in the allocator's payoffs. Thresholds on the beliefs are increasing in b and z and decreasing in α and λ_i. Formal proof has been relegated to the appendix (see appendix 5).

If we compare these two situations we can shed light about the frequency of punishment in each context (see Figure 5.1). We can summarize lemma 1 by saying that rewards $0 \leq b < b''$ are punished if there are enough reciprocators, whereas rewards $b'' \leq b \leq 1$ are left unpunished. In the same way, under uncoordinated punishment (lemma 2), if $0 \leq b < \hat{b}$, there is unconditional punishment, rewards $\hat{b} \leq b < b^{**}$ are conditionally punished and rewards $b^{**} \leq b < 1$ are always left unpunished.

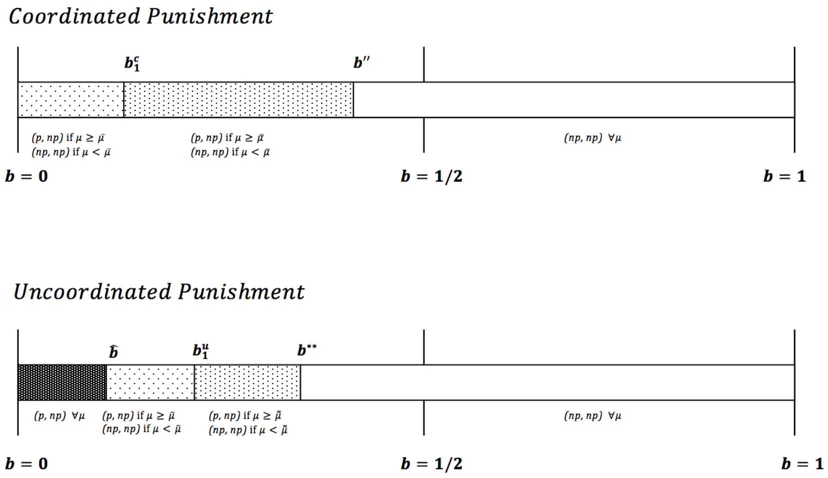
Coordinated Punishment
b_1^c
b''
(p, np) if $\mu \geq \bar{\mu}$
(np, np) if $\mu < \bar{\mu}$
(p, np) if $\mu \geq \bar{\mu}$
(np, np) if $\mu < \bar{\mu}$
$(np, np)\ \forall\mu$
$b = 0$
$b = 1/2$
$b = 1$
Uncoordinated Punishment
$\hat{b}$
b_1^u
b^{**}
$(p, np)\ \forall\mu$
(p, np) if $\mu \geq \tilde{\mu}$
(np, np) if $\mu < \tilde{\mu}$
(p, np) if $\mu \geq \tilde{\tilde{\mu}}$
(np, np) if $\mu < \tilde{\tilde{\mu}}$
$(np, np)\ \forall\mu$
$b = 0$
$b = 1/2$
$b = 1$

Figure 5.1: BNE of the punishment stage

Given that $b^{**} < b''$, the range of values of b that reciprocal investors leave unpunished is larger under uncoordinated punishment. An additional insight is that, for a sufficiently high proportion of reciprocators in the population, punishment occurs more frequently under a coordinated punishment system than under an uncoordinated one.

Example

Let us now suggest a numerical example to provide the reader a better understanding of the results of the punishment stage. Suppose the parameters of the model take the following feasible values: $\alpha = 2$, $c = 0.25$, $z = 0.5$,

$\lambda_c = 0.7$ and $\lambda_u = 0.3$, which fulfill the model's assumptions. With these values, the solution of any subgame $CP(\mu, b)$ with unfair rewards would be as follows:

- If $b < 0.2308$, then (p, np) if $\mu > \bar{\mu} = \frac{0.5}{1.65-1.4b}$ and (np, np) otherwise
- If $0.2308 \leq b < 0.4375$, then (p, np) if $\mu > \bar{\bar{\mu}} = \frac{0.5}{2.25-4b}$ and (np, np) otherwise
- If $0.4375 \leq b < 1/2$, then (np, np) $\forall \mu$

For this set of parameters, the solution of any subgame $UP(\mu, b)$ with unfair rewards would be as follows:

- If $b < 0.1666$, then (p, np) $\forall \mu$
- If $0.1666 \leq b < 0.2857$, then (p, np) if $\mu > \tilde{\mu} = 0.8 + 1.2b$ and (np, np) otherwise
- If $0.2857 \leq b < 0.3382$, then (p, np) if $\mu > \tilde{\tilde{\mu}} = \frac{-0.1+0.6b}{1.05-2.8b}$ and (np, np) otherwise
- If $0.3382 \leq b < 1/2$, then (np, np) $\forall \mu$.

Comparing the two punishment schemes, we can see how for rewards below $b = 0.1666$ there would certainly be uncoordinated punishment whereas coordinated punishment would only happen for high enough values of μ. On the other hand, rewards close enough to the fair reward are always left unpunished. However, with uncoordinated punishment, reciprocal investors are more permissive, as a reward of $b = 0.3382$ is already high enough to leave unpunished whereas with coordinated punishment, reciprocators need

$b = 0.4375$ in order not to punish. Finally, we can also appreciate the higher requirement of investors under coordinated punishment if we take intermediate rewards, such as $b = 0.2$, as the corresponding μ needed to carry out punishment is higher under coordinated than under uncoordinated punishment ($\bar{\mu} = 0.365$ vs. $\tilde{\mu} = 0.08$). Therefore, broadly speaking, we can conclude that coordinated punishers are more demanding with the received returns. This happens due to the necessity of there being two reciprocal investors punishing in order to reduce inequality with respect to the allocator.

5.3.2 Reward policy

This reward stage is only reached in the subgame (I, I) of the investment stage. Thus, after observing that both investors have decided to invest in the project, the allocator chooses how much to return to each investor, $0 \leq b \leq 1$, which we assume symmetric. As the project surplus is of size 2, each investor generates a surplus of 1 with his investment, so we define a reward of $b = 1/2$ as fair. Recall we also consider two possible different types of allocators, fair minded[4] who will return the fair reward $b_f^i = 1/2$ and profit maximisers who will return b_{pm}^i. The super index $i \in \{c, u\}$ represents the punishment scheme: coordinated (c) or uncoordinated (u). Recall that fair rewards are never going to be punished by any type of investor. Unfair rewards, that is, any $b < 1/2$, may be or may not be punished by reciprocators.

When deciding on how much to reward, a profit-maximiser allocator

[4] Notice that a fair-minded allocator does not change his rewarding policy when there is incomplete information.

will compare the expected cost of being punished with the expected cost of avoiding it. On one hand, the expected cost of punishment will depend on the beliefs the allocator has on the the types of the investors. On the other hand, the expected cost of avoiding punishment is the minimum reward an allocator shall return for no investor to punish.

We describe the optimal reward policy of the profit-maximiser allocator with the following set of lemmatas:

Lemma 3: *In the team trust game with coordinated punishment, the reward policy of a profit-maximiser allocator is:*

a) *If* $\mu < \bar{\mu}(0)$ *then* $b^c_{pm} = 0$

b) *If* $\bar{\mu}(0) \leq \mu < \bar{\mu}(b^c_1)$, *then either* $b^c_{pm} = 0$ *or* $b^c_{pm} = b^c$

c) *If* $\bar{\mu}(b^c_1) \leq \mu < 1$, *then either* $b^c_{pm} = 0$ *or* $b^c_{pm} = b^{cc}$

where: $b^c = \frac{\mu(z/2+\alpha\lambda_c)-z}{\mu\alpha\lambda_c} = 0$ *and* $b^{cc} = \frac{\mu(z/2+\alpha)-z}{2\mu\alpha}$

Lemma 4: *In the team trust game with uncoordinated punishment, the reward policy of a profit-maximiser allocator is:*

a) *If* $\mu < \tilde{\mu}(b^u_1)$ *then* $b^u_{pm} = 0$

b) *If* $\tilde{\mu}(b^u_1) \leq \mu < 1$, *then then either* $b^u_{pm} = 0$ *or* $b^u_{pm} = b^{uu}$

where: $b^{uu} = \frac{\mu(z/2+\alpha(1-2\lambda_u))-z+\alpha\lambda_u}{\alpha(\lambda_u+2\mu)}$

If the proportion of reciprocal investors is sufficiently low, a profit-maximiser allocator anticipates that there is going to be no punishment in the last stage, so the expected costs of being punished are zero. Thus, the profit-maximiser allocator prefers to offer the lowest return $b^i_{pm} = 0$. However, if the proportion is sufficiently high and there is conditional punishment, profit-maximiser allocators can either return nothing and face punishment, or, alternatively, offer a minimal reward (b^c, b^{cc} or b^{uu} as the case may be) such that reciprocal investors do not punish them. The expected costs of being punished are $\mu^2\lambda_c$ under coordinated punishment and $\mu\lambda_u$ under uncoordinated punishment, where the difference is that coordinated punishment requires two reciprocators for punishment to be effective. Therefore, the optimal reward policy depends on the critical values derived from the comparison of the expected payoffs by offering $b^i_{pm} = 0$ and $b^i_{pm} > 0$.[5] The characterization of these values is found in appendix 5.

Example

If we go back to the example previously suggested with $\alpha = 2$, $c = 0.25$, $z = 0.5$, $\lambda_c = 0.7$ and $\lambda_u = 0.3$, the reward policy of a profit-maximiser allocator in the team trust game with coordinated punishment will be:

- If $\mu < 0.303$, then $b^c_{pm} = 0$
- If $0.303 \leq \mu < 0.3231$, then $b^c_{pm} = b^c = \frac{-0.357+1.179\mu}{\mu}$

[5] The comparison between the different rewarding policies leads to a cubic equation $f(\mu) = -(\alpha\lambda_c^2)\mu^3 + (\alpha\lambda_c + z/2) - z$ for the comparison between offering $b^c_{pm} = 0$ or $b^c_{pm} = b^c$, a cubic equation $g(\mu) = -(2\alpha\lambda_c)\mu^3 + (\alpha + z/2)\mu - z$ for the comparison between offering $b^c_{pm} = 0$ or $b^c_{pm} = b^{cc}$ and a quadratic equation $h(\mu) = -(2\alpha\lambda_u)\mu^2 + (z/2 + \alpha(1 - 2\lambda_u - \lambda_u^2))\mu - z + \alpha\lambda_u$ for the comparison between offering $b^u_{pm} = 0$ or $b^u_{pm} = b^{uu}$.

- If $0.3231 \leq \mu < 0.7525$, then $b^c_{pm} = 0$

- If $0.7525 \leq \mu < 1$, then $b^c_{pm} = b^{cc} = \frac{-0.5+2.25\mu}{4\mu}$

Likewise, the reward policy of a profit-maximiser allocator in the team trust game with uncoordinated punishment will be:

- If $\mu < 0.8259$, then $b^u_{pm} = 0$

- If $0.8259 \leq \mu < 1$, then $b^u_{pm} = b^{uu} = \frac{0.1+1.05\mu}{0.6+4\mu}$

From these reward policies we can assert that with coordinated punishment, a positive b is returned for a wider range of beliefs. In particular, for $\mu \in [0.303, 0.3231) \cup [0.7525, 1)$ with coordinated punishment and only for $\mu \in [0.8259, 1)$ with uncoordinated punishment. Hence, we can claim that the punishment threat is stronger under a coordinated system, as a $\mu = 0.7525$, is already sufficiently high for profit-maximiser allocators to return the positive reward. Additionally, given any value of μ, the reward returned with coordinated punishment is higher than the reward returned with uncoordinated punishment. For example, for $\mu = 0.85$, the return with coordinated punishment would be $b^{cc} = 0.4154$ and the return with uncoordinated punishment would be $b^{uu} = 0.2481$. Hence, we can conclude that with uncoordinated punishment higher beliefs are needed for a positive reward and, even so, that amount returned will be lower than the amount returned with coordinated punishment.

5.3.3 Investment stage

Finally, in the investment stage, investors foresee the allocator's reward policy and their punishment reaction when deciding on whether to invest or not. There are multiple PBE of the game with coordinated and with uncoordinated punishment. For instance, there is always an inefficient non-cooperative equilibrium where both types of investors choose not to invest for every q and m. However, in this work we are only interested in the pooling equilibria where both types of investors decide to invest.[6] In these pooling equilibria, $q = \mu = Prob(r/I, I)$.

***Proposition 1:* Efficient pooling equilibria with** $b^i_{pm} > 0$

a) *In the team trust game with* **coordinated punishment**, *if either* $\bar{q}(0) \leq q < q_1$ *or* $q_2 \leq q < \bar{q}(b^c_1)$, *and* $m \geq m^c$, *there exists a PBE in which both types of investors decide to invest,* (I, I), *a fair-minded allocator returns the fair reward* $b^c_f = 1/2$, *a profit-maximiser allocator returns a reward sufficiently high to avoid punishment,* $b^c_{pm} = b^c$ *and there is no punishment. Likewise, if either* $\bar{q}(b^c_1) \leq q \leq q^{'}_1$ *or* $q^{'}_2 \leq q < 1$, *and* $m \geq m^{cc}$, *there exists a PBE in which both types of investors decide to invest,* (I, I), *a fair-minded allocator returns the fair reward* $b^c_f = 1/2$, *a profit-maximiser allocator returns a reward sufficiently high to avoid punishment* $b^c_{pm} = b^{cc}$ *and there is no punishment.*

b) *In the team trust game with* **uncoordinated punishment**, *if* $q^{''}_1 \leq q < 1$ *and* $m \geq m^{uu}$, *there exists a PBE in which both types of investors decide*

[6] This game has separating equilibria in which selfish investors invest and reciprocal investors do not invest, but these are not stable in a long-run dynamics, following Calabuig and Olcina (2015).

to invest, (I, I), a fair-minded allocator returns the fair reward $b^u_f = 1/2$, a profit-maximiser allocator returns a reward sufficiently high to avoid punishment, $b^u_{pm} = b^{uu}$ and there is no punishment.

where: $m^c = \frac{c-b^c+\alpha(1-2b^c)}{0.5-b^c+\alpha(1-2b^c)}$, $m^{cc} = \frac{c-b^{cc}+\alpha(1-2b^{cc})}{0.5-b^{cc}+\alpha(1-2b^{cc})}$ *and* $m^{uu} = \frac{c-b^{uu}+\alpha(1-2b^{uu})}{0.5-b^{uu}+\alpha(1-2b^{uu})}$

For these efficient pooling equilibria to happen, the proportion of reciprocal investors must be high enough and the proportion of fair-minded allocators must also be high enough. The first condition ensures that a profit-maximiser allocator sets a positive reward in order to avoid the potential punishment of the reciprocal investors. The values $q_1, q_2, q_1^{'}, q_2^{'}$ and $q_1^{''}$ are the positive roots of the cubic and quadratic equations derived from the comparison of offering $b^i_{pm} = 0$ and being punished or $b^i_{pm} > 0$ and avoiding punishment. The second condition, on the other side, guarantees that both types of investors prefer to invest rather than not to do so. If these two conditions are met, both types of investors invest, regardless of their type. They receive a fair reward in case of being matched with a fair-minded allocator (who has as a dominant action to set these rewards) or a positive reward in case of being matched with a profit-maximiser allocator, following lemmas 3 and 4. If the investor is selfish, he will never punish regardless of the return received, whereas a reciprocal investor would not punish given that he either receives a fair reward or a sufficiently high one, by lemmas 1 and 2.

Notice that, with coordinated punishment, there is a lower segment ($q \in [\bar{q}(0), q_1) \cup [q_2, \bar{q}(b^c_1))$, where with few reciprocal investors, the efficient pooling equilibria is sustained if the amount of fair-minded allocators is sufficiently high. This, however, does not happen with uncoordinated punishment. Nevertheless, as we will appreciate with the numerical example, this segment is minor. More importantly, under both punishment systems, there are efficient pooling equilibria with positive and significantly high rewards, b^{cc} and b^{uu}, for a high proportion of reciprocators. Our main goal is, as previously stated, to compare both punishment systems in terms of achieving greater levels of efficiency. We present our main result concerning this issue in the following proposition.

Proposition 2: *In the team trust game with punishment, for a proportion of reciprocal investors $q > q_2^{'}$, the reward returned by a profit-maximiser allocator is higher under coordinated punishment, $b^{cc} \geq b^{uu}$ and the proportion of fair-minded allocators necessary for this to be an efficient pooling equilibrium is lower, $m^{cc} \leq m^{uu}$.*

This proposition illustrates that in a population of investors with a high proportion of reciprocators, coordinated punishment is substantially superior to uncoordinated punishment. On the one hand, investors obtain higher rewards from allocators and joint investment is achieved with a wider range of the distribution of preferences of the allocator population.

Additionally, there is also a set of efficient pooling equilibria where, if the proportion of fair-minded allocators is remarkably high, profit-maximisers can return nothing and still avoid punishment.

***Proposition 3:* Efficient pooling equilibria with $b^i_{pm} = 0$**

a) *In the team trust game with* **coordinated punishment**, *if $q < \bar{q}(0)$ and $m \geq m^0$, there exists a PBE in which both types of investors decide to invest, (I, I), a fair-minded allocator returns the fair reward $b^c_f = 1/2$, a profit-maximiser allocator returns $b^c_{pm} = 0$ and there is no punishment.*

b) *In the team trust game with* **uncoordinated punishment**, *if $q < \tilde{q}(b^u_1)$ and $m \geq m^0$, there exists a PBE in which both types of investors decide to invest, (I, I), a fair-minded allocator returns the fair reward $b^u_f = 1/2$, a profit-maximiser allocator returns $b^u_{pm} = 0$ and there is no punishment.*

where: $m^0 = \frac{\alpha+c}{\alpha+0.5}$

For sufficiently low proportions of reciprocal investors, profit-maximiser allocators anticipate no future punishment, following lemmas 1 and 2. In this line, they will maximise their payoffs by not returning anything at all, by lemmas 3 and 4. Even in this case there can be investment by both types of investors if the proportion of fair-minded allocators who return them a fair reward is high enough.

Finally, this game also has pooling equilibria which are inefficient, in the sense that reciprocators implement punishment given that profit-maximiser allocators return nothing and the proportion of fair-minded is not sufficiently high.

***Proposition 4:* Inefficient pooling equilibria**

a) *In the team trust game with* ***coordinated punishment****, if either* $q_1 \leq q < q_2^{'}$ *or* $q_1^{'} \leq q < q_2^{'}$*, and* $m \geq m^{cp}$*, there exists a PBE in which both types of investors decide to invest,* (I, I)*, a fair-minded allocator returns the fair reward* $b_f^c = 1/2$*, a profit-maximiser allocator returns* $b_{pm}^c = 0$ *and there is punishment from reciprocal investors.*

b) *In the team trust game with* ***uncoordinated punishment****, if* $\tilde{q}(b_1^u) \leq q < q_1^{''}$ *and* $m \geq m^{up}$*, there exists a PBE in which both types of investors decide to invest,* (I, I)*, a fair-minded allocator returns the fair reward* $b_f^u = 1/2$*, a profit-maximiser allocator returns* $b_{pm}^u = 0$ *and there is punishment from reciprocal investors.*

where: $m^{cp} = \frac{\alpha+c+z-q(z/2+\alpha\lambda_c)}{\alpha+0.5+z-q(z/2+\alpha\lambda_c)}$ *and* $m^{up} = \frac{\alpha(1-\lambda_u)+c+z-q(z/2+\alpha\lambda_u)}{\alpha(1-\lambda_u)+0.5+z-q(z/2+\alpha\lambda_u)}$.

If beliefs about reciprocal investors are sufficiently high such that there is conditional punishment from their part, but profit-maximiser allocators do not return anything to investors, there will be punishment by the reciprocal investors (lemmas 1 and 2). As before, even in this case there could be incentives to invest in the project if the proportion of fair-minded allocators is large enough. This decision is now not only going to depend on the investment cost c, the reward b and the inequality measure α, but also on the punishment cost z and on the probability of being paired with another reciprocal investor, q with who to share the cost. For the case of coordinated punishment, recall that the presence of another reciprocal is needed for punishment to have an a negative impact on the allocator's final payoff.

Example

Let us illustrate the results using the example parameters proposed. With this end, table 5.2 shows, for the different proportions of reciprocal investors, q, the necessary minimal fair-minded allocators proportion, m, for the existence of a pooling PBE, the corresponding reward by the profit-maximiser allocators, b_{pm}, and the punishment strategy of a reciprocator knowing the selfish investor never punishes, (x, np).

q	**m**		b_{pm}		**(x,np)**	
	CP	UP	CP	UP	CP	UP
0.05	0.9	0.9	0	0	np	np
0.1	0.9	0.9	0	0	np	np
0.15	0.9	0.9	0	0	np	np
0.2	0.9	0.9	0	0	np	np
0.25	0.9	0.9	0	0	np	np
0.3	0.9	0.883	0	0	np	p
0.32	0.886	0.883	0.0625	0	np	p
0.4	0.893	0.879	0	0	p	p
0.45	0.889	0.876	0	0	p	p
0.5	0.885	0.873	0	0	p	p
0.55	0.881	0.871	0	0	p	p
0.6	0.876	0.868	0	0	p	p
0.65	0.87	0.865	0	0	p	p
0.7	0.864	0.861	0	0	p	p
0.75	0.858	0.858	0	0	p	p
0.8	0.467	0.855	0.406	0	np	p
0.85	0.409	0.801	0.415	0.248	np	np
0.9	0.345	0.801	0.424	0.249	np	np
0.95	0.276	0.8	0.431	0.249	np	np
1	0.2	0.8	0.438	0.25	np	np

Table 5.2: PBE of the team trust game with (I,I)

This table points out that there is an inverse relationship between q and m. In other words, the higher the proportion of reciprocators in the population is, the lower the proportion of fair-minded allocator needed for there to be a pooling equilibrium in investment, regardless of the punishment scheme. Furthermore, differences between coordinated and uncoordinated punishment can be appreciated when the proportion of reciprocals is high enough $q \geq 0.8$. When this happens, rewards under coordinated punishment are substantially higher and the proportion of fair-minded allocators drastically falls. However, with uncoordinated punishment, both the increase of b^u_{pm} and the fall of m are marginal for these high q. For low and intermediate values of q, however, differences between coordinated and uncoordinated punishment are minimal.

The main difference between coordinated and uncoordinated punishment, which happens for high proportions of reciprocators, is driven by the fact that reciprocators are willing to punish higher offers under coordinated punishment, to which profit-maximiser allocators return higher rewards than under uncoordinated punishment. This, in turn, enhances efficiency through joint investment even with a lower proportion of fair-minded allocators in the population.

At this point, the reader could think that this superiority of coordinated punishment with regard to uncoordinated punishment with many reciprocators could be due to the fact that, for this numerical example, coordinated punishment has been considered more destructive than uncoordinated punishment, i.e. $\lambda_c > 2\lambda_u$. However, we can prove that even if they had equal impact, $\lambda_c = 2\lambda_u$, this effect would prevail for high reciprocators. Table 5.3 shows the PBNE with $\lambda_c = 0.6, \lambda_u = 0.3$.

q	**m**		b_{pm}		**(x,np)**	
	CP	UP	CP	UP	CP	UP
0.7	0.874	0.861	0	0	p	p
0.75	0.869	0.858	0	0	p	p
0.8	0.864	0.855	0	0	p	p
0.85	0.409	0.801	0.415	0.248	np	np
0.9	0.345	0.801	0.424	0.249	np	np
0.95	0.276	0.8	0.431	0.249	np	np
1	0.2	0.8	0.438	0.25	np	np

Table 5.3: PBE of the team trust game with (I,I) for $\lambda_c = 0.6$

With this lower value of λ_c, the reward policy is such that now the positive reward $b^c_{pm} = b^{cc}$ is offered for $\mu \geq 0.8282$ for coordinated punishment and $b^u_{pm} = b^{uu}$ for $\mu \geq 0.8259$ for uncoordinated punishment. This makes that with $q = 0.8$, investors now receive $b^c_{pm} = 0$ and $b^u_{pm} = 0$ and both rewards are punished. However, once profit-maximiser allocators return the positive reward, the combinations of q and m along with the rewards of the efficient pooling equilibrium are still significantly different between the different punishment schemes. Notice that these do not change with respect to the previous example as the rewards b^{cc}, b^{uu} and the thresholds for fair-minded allocators m^{cc}, m^{uu} are independent of λ_c.

The superiority of coordinated punishment for high reciprocators is indeed given by coordination as a mechanism to overcome the free rider issue. Intuitively, differences between the two types of punishment lie on the free riding incentives between reciprocators. With coordinated punishment, this behaviour can be avoided when there are many reciprocators in the population, as if one of the two investors free rides, punishment actions will be costly and ineffective in destroying the allocator's payoff. When the proportion of reciprocators is high, the probability of being paired with another reciprocator is also high, so the chance of there being joint punishment is higher. Uncoordinated punishment, however, is prone to free riding for any value of q, as the allocator's payoff is going to be undermined even if there is only one punisher. Hence, even if the proportion of reciprocators is high, the threat of joint punishment is not as strong as with coordinated punishment.

Notice that, the driving force of reciprocators punishing is the reduction of disadvantageous inequality, which can be understood as a public good achieved with such punishment actions. Nevertheless, the value that the two types of investors give to this public good is not the same: only reciprocators value inequality reduction positively. A way of overcoming the inefficient outcome resulting from a public goods game is to make it become a coordination game, and, for this reason, coordinated punishment proves to be more efficient than uncoordinated punishment when there is a high proportion of reciprocators in the population. With a lower proportion of these, however, the presence of selfish investors dissipates the public good. In this case, the advantage of individual effective uncoordinated punishment prevails and differences between coordinated and uncoordinated punishment are minimal.

5.4 Conclusions

In the theoretical comparison of uncoordinated and coordinated punishment, we have highlighted the superiority of the coordinated system in achieving cooperation when the proportion of reciprocal investors is sufficiently high. This occurs because there are a greater range of cases for which investors decide to jointly invest in the project, the allocator returns a positive reward and there is no punishment. Additionally, the amount returned under a coordinated punishment system is larger than under an uncoordinated one and the proportion of fair-minded allocators required for this equilibrium to happen is lower.

The underlying reason for this behaviour is that the punishment threat exerted by coordinated punishers is stronger than the one exerted by uncoordinated punishers, as under a coordinated scheme, two investors willing to punish are needed for punishment to inflict damage on the allocators' payoffs. This is reflected in the fact that reciprocal investors are more demanding with the allocators' returns under a coordinated punishment system. As allocators foresee this exigence, they will return rewards that are sufficiently high for the investors not to punish them, which enhances joint investment for a wider range of cases.

However, what is particularly noteworthy is that the superiority of coordinated punishment only occurs when there is a high proportion of reciprocators in the population. The reason behind this is that the coordination mechanism can avoid free riding behaviour among reciprocators if many. With a low proportion of reciprocators in the population, however, such advantage disappears.

References

1. Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science,* 328(5978), 617-620.

2. Calabuig, V., Jiménez, N., Olcina, G., & Rodríguez-Lara, I (2018). “United we stand." *Mimeo*

3. Calabuig, V. & Olcina, G. (2015). “Coordinated punishment and the evolution of cooperation." *Journal of Public Economic Theory,* 17(2), 147-173.

4. Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives,* 14(3), 159-181

5. Fehr, E., & Schmidt, K. M. (1999). “A theory of fairness, competition, and cooperation." *The quarterly journal of economics,* 114(3), 817-868.

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

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

Appendix 5. Proofs

Proof of Lemma 1. BNE of the punishment stage under coordinated punishment.

Let us first prove that selfish investors will never punish in the coordinated punishment stage by contradiction. For a selfish investor to implement punishment, the expected utility of punishing, $\mu(p, p) + (1 - \mu)(p, np)$, must be greater than the utility of not punishing, $\mu(np, p) + (1 - \mu)(np, np)$. By substituting the corresponding payoffs, the comparison the selfish investor makes is:

$$\mu(b - c - z/2) + (1 - \mu)(b - c - z) \geq b - c$$

Solving for μ we obtain that a selfish investor will punish if $\mu \geq 2$, which never holds.

Knowing a selfish investor is always going to choose np, a reciprocal investor can incorporate this in his comparison of the expected utility of punishing, $\mu(p, p) + (1 - \mu)(p, np)$ and his expected utility of not punishing $\mu(np, p) + (1 - \mu)(np, np)$. By substituting the corresponding payoffs, the comparison the reciprocal investor makes is:

$$\mu[b - c - z/2 - \alpha[(1 - b)(1 - \lambda_c) - b]] + (1 - \mu)[b - c - z - \alpha[(1 - b) - b]] \geq b - c - \alpha[(1 - b) - b]$$

Notice that there are three inequalities in this comparison, but whether they become active or not depends on the value b takes. In particular, if $b \leq b_1^c$, all inequalities become active. However, if $b > b_1^c$, only inequalities $\alpha[(1-b)-b]$ become active.

Let us start with the case where $b \leq b_1^c$, such that all inequalities become active. Solving for μ we obtain that a reciprocal investor will punish if $\mu > \frac{z}{z/2+\alpha\lambda_c(1-b)}$, from which we define $\bar{\mu}$. This threshold $\bar{\mu}$ is always positive, but for it to be $\bar{\mu} \leq 1$, then $b \leq b^{'} = 1 - \frac{z}{2\alpha\lambda_c}$. As $b^{'} > 1/2$, all values $b \leq b_1^c$ accomplish $\bar{\mu} \leq 1$. Therefore, in the reward segment $b \leq b_1^c$, if $\mu \leq \bar{\mu}$ then reciprocals would not punish, (np, np), because the probability of being paired with another reciprocal investor is low. Otherwise, if $\mu > \bar{\mu}$, reciprocal investors would punish, (p, np).

For the case where $b > b_1^c$, only inequalities $\alpha[(1-b)-b]$ become active. Solving for μ we obtain that a reciprocal investor will punish if $\mu > \frac{z}{z/2+\alpha(1-2b)}$, from which we define $\bar{\bar{\mu}}$. This threshold $\bar{\bar{\mu}}$ is always positive, but for it to be $\bar{\bar{\mu}} < 1$, then $b \leq b^{''} = \frac{1}{2} - \frac{z}{4\alpha}$. It can be shown that $b_1^c < b^{''} < 1/2$, from which we define two segments. If $b_1^c \leq b < b^{''}$ whether there is punishment or not depends on $\bar{\bar{\mu}}$. If $\mu \leq \bar{\bar{\mu}}$ then (np, np). Otherwise, if $\mu > \bar{\bar{\mu}}$, then (p, np). For $b^{''} \leq b < 1/2$, $\bar{\bar{\mu}} > 1$, so (np, np) $\forall \mu$.

■

Proof of Lemma 2. BNE of the punishment stage under uncoordinated punishment.

The proof of not punishing being a dominant strategy for selfish investors in the uncoordinated punishment stage follows from the proof for lemma 1.

For the case of the reciprocal investor, the comparison that he makes is:

$$\mu[b-c-z/2-\alpha[(1-b)(1-2\lambda_u)-b]]+(1-\mu)[b-c-z-\alpha[(1-b)(1-\lambda_u)-b]] \geq \mu[b-c-\alpha[(1-b)(1-\lambda_u)-b]]+(1-\mu)[b-c-\alpha[(1-b)-b]]$$

Notice that there are four inequalities in this comparison and whether they become active or not depends on the value of the reward b:

- If $b < b_1^u = \frac{1-2\lambda_u}{2-2\lambda_u}$, then all inequalities become active.
- If $b_1^u \leq b < b_2^u = \frac{1-\lambda_u}{2-\lambda_u}$, then inequalities $\alpha[(1-b)(1-\lambda_u)-b]$ and $\alpha[(1-b)-b]$ become active.
- If $b_2^u \leq b < 1/2$, then only inequality $\alpha[(1-b)-b]$ becomes active.

If $b \leq b_1^u$, all inequalities become active. Solving for μ we obtain that a reciprocal investor will punish if $\mu > \frac{2z-2\alpha\lambda_u(1-b)}{z}$, from which we define $\tilde{\mu}$. For this threshold to be positive, then $b \geq \hat{b} = 1-\frac{z}{\alpha\lambda_u}$, which holds by assumption 2. For this threshold to be less than 1, $b \leq b^* = 1-\frac{z}{2\alpha\lambda_u}$. It can be shown that $b^* > b_1^u$, therefore, all values below b_1^u satisfy $\tilde{\mu} \leq 1$. Hence, we have two segments: $0 \leq b < \hat{b}$ and $\hat{b} \leq b < b_1^u$. For the first case, as $b < \hat{b}$, then $\tilde{\mu} < 0$ and therefore the unique BNE is (p, np) $\forall \mu$. For the second case, where $\hat{b} \leq b < b_1^u$, whether there is punishment or not depends on the value of μ. If $\mu \leq \tilde{\mu}$ then (np, np). Otherwise, if $\mu > \tilde{\mu}$, then (p, np).

If $b_1^u \leq b < b_2^u$, then inequalities $\alpha[(1-b)(1-\lambda_u)-b]$ and $\alpha[(1-b)-b]$ become active. Solving for μ we obtain that a reciprocal investor will punish if $\mu > \frac{z-\alpha(1-b)\lambda_u}{z/2+\alpha(1-2b)-2\alpha(1-b)}$, from which we define $\tilde{\tilde{\mu}}$. Whether $\tilde{\tilde{\mu}}$ is positive is no longer straightforward and depends on the sign of the numerator and the denominator. For the numerator to be positive $b \geq \hat{b} = 1-\frac{z}{\alpha\lambda_u}$, which holds by assumption 2. For the denominator to be positive $b \leq \hat{\hat{b}} = \frac{1-2\lambda_u}{2-2\lambda_u} + \frac{z}{4\alpha(1-\lambda_u)}$ must hold. Finally, for $\tilde{\tilde{\mu}} \leq 1$, then b must satisfy $b < b^{**} = \frac{1-\lambda_u}{2-\lambda_u} - \frac{z}{2\alpha(2-\lambda_u)}$. By assumption 2, we can claim that $b_1^u < b^{**} < \hat{\hat{b}} < b_2^u$, defining three segments. For any b in the segment $b_1^u \leq b < b^{**}$, both the numerator and the denominator of $\tilde{\tilde{\mu}}$ are positive and $\tilde{\tilde{\mu}} \leq 1$ holds. Hence, whether there is punishment or not depends on $\tilde{\tilde{\mu}}$: if $\mu \leq \tilde{\tilde{\mu}}$ then (np, np). Otherwise, if $\mu > \tilde{\tilde{\mu}}$, then (p, np). For any b in the segment $b^{**} \leq b < \hat{\hat{b}}$, both the numerator and the denominator of $\tilde{\tilde{\mu}}$ are positive, but $\tilde{\tilde{\mu}} > 1$. Therefore, in this case the BNE is (np, np) $\forall\mu$. Finally, for the last segment where $\hat{\hat{b}} \leq b < b_2^u$, the numerator is positive, but the denominator is negative as $b \geq \hat{\hat{b}}$, therefore, $\tilde{\tilde{\mu}}$ must be lower than a negative number, which never holds, so (np, np) is the unique BNE $\forall\mu$.

In last place, let us study the higher segment of b, where $b \geq b_2^u$. Now, only the inequality $\alpha[(1-b)-b]$ becomes active. Solving for μ, we obtain that for the reciprocal investor to punish $\mu \geq \tilde{\tilde{\tilde{\mu}}} = \frac{z-\alpha(1-2b)}{z/2-\alpha(1-2b)}$. As in the previous case, the numerator is positive by assumption 2, but for the denominator to be positive as well, it is necessary that $b \geq \hat{\hat{b}}' = \frac{1}{2} - \frac{z}{4\alpha}$. It can be shown that $b_2^u \leq \hat{\hat{b}} < 1/2$, such that we have two segments. In the first segment, $b_2^u \leq b < \hat{\hat{b}}'$, the numerator of $\tilde{\tilde{\tilde{\mu}}}$ is positive while the denominator is negative. Therefore, $\tilde{\tilde{\tilde{\mu}}}$ must be lower than a negative number, which never holds, so

(np, np) is the unique BNE $\forall\mu$. Alternatively, for $\hat{\hat{b}}' \leq b < 1/2$, both the numerator and the denominator are positive, but $\tilde{\tilde{\bar{\mu}}} > 1$, so (np, np) $\forall\mu$.

■

Proof of Lemma 3: PBE of the reward stage under coordinated punishment:

For the reward policy with coordinated punishment, recall that allocators forward that only reciprocal investors may punish and that if this happens it does so for rewards $0 \leq b < b^{''}$. Whether reciprocal investors punish or not depends on the beliefs of the proportion of reciprocators in the population. In this range of values, profit-maximiser allocators can either set $b^c_{pm} = 0$ and make profits of $\pi^c_{pm}(0) = 2(1 - \mu^2\lambda_c)$, or offer a minimal reward $b^c_{pm} > 0$ to avoid punishment and make profits of $\pi^c_{pm}(b) = 2(1 - b)$. Notice that the critical values $\bar{\mu}(b)$ and $\bar{\bar{\mu}}(b)$ are defined by the range of b where they are relevant. It can be easily shown that the critical values of μ are increasing with b and that they can be ordered such that $\bar{\mu}(0) \leq \bar{\mu}(b^c_1) = \bar{\bar{\mu}}(b^c_1) \leq \bar{\bar{\mu}}(b^{''}) = 1$.

If the beliefs are very low, that is, if $\mu < \bar{\mu}(0)$, then $\mu < \bar{\mu}(b)$ $\forall b$. Given that a profit maximiser faces no risk of being punished with so low beliefs, he will maximise his profits by returning $b^c_{pm} = 0$.

If, instead, $\bar{\mu}(0) \leq b < \bar{\mu}(b^c_1) = \bar{\bar{\mu}}(b^c_1)$ a profit maximiser can either return nothing or set a minimal reward of $b^c = \frac{\mu(z/2 + \alpha\lambda_c) - z}{\mu\alpha\lambda_c}$ to avoid punishment. This minimal reward can be directly obtained from the threshold $\bar{\mu}$. In order to choose between these two options, a profit-maximiser allocator will compare the expected profits of such options. In particular, he will offer $b^c_{pm} = 0$ if $-(\alpha\lambda^2_c)\mu^3 + (\alpha\lambda c + z/2)\mu - z \geq 0$. If the discriminant of this cubic equation is positive, then there is a real root, which is negative and two complex roots. Therefore, setting $b^c_{pm} = 0$ is better. If the discriminant is zero, then all roots are real, one of them negative and the other two would be two equal and positive roots. Therefore, setting $b^c_{pm} = b^c$ is better. However, if the discriminant is negative, then there are three real and unequal roots, one of them negative. In this case, for any $\mu < \mu_1$ and $\mu > \mu_2$, the positive reward $b^c_{pm} = b^c$ would be returned, whereas for any $\mu_1 \leq \mu \leq \mu_2$, the reward would be $b^c_{pm} = 0$, where μ_1 and μ_2 are the positive roots of the cubic equation.

Finally, if $\bar{\mu}(b^c_1) = \bar{\bar{\mu}}(b^c_1) \leq b \leq \bar{\bar{\mu}}(b^{''}) = 1$, the profit-maximiser allocator can, as before, either return nothing or return a positive reward that avoids punishment, $b^{cc} = \frac{\mu(\alpha+z/2)-z}{2\alpha\mu}$, which can be directly obtained from $\bar{\bar{\mu}}$ and it can be shown that it is below the fair return. A profit-maximiser allocator will return $b^c_{pm} = 0$ if $-(2\alpha\lambda_c)\mu^3 + (\alpha + z/2)\mu - z \geq 0$. If the discriminant is positive or equal to zero, the positive reward $b^c_{pm} = b^{cc}$ would always be offered. For the cases where the discriminant is negative, for any $\mu < \mu^{'}_1$ and $\mu > \mu^{'}_2$, the positive reward $b^c_{pm} = b^{cc}$ would be returned, whereas for any $\mu^{'}_1 \leq \mu \leq \mu^{'}_2$, the reward would be $b^c_{pm} = 0$, where $\mu^{'}_1$ and $\mu^{'}_2$ are the positive roots of the cubic equation.

■

Proof of Lemma 4: PBE of the reward stage under uncoordinated punishment:

For the reward policy with uncoordinated punishment, profit-maximiser allocators can either set $b^u_{pm} = 0$ and make profits of $\pi^c_{pm}(0) = 2(1 - \mu\lambda_u)$, or offer a minimal reward $b^u_{pm} > 0$ to avoid punishment and make profits of $\pi^u_{pm}(b) = 2(1 - b)$. Notice that the critical values $\tilde{\mu}(b)$ and $\tilde{\tilde{\mu}}(b)$ are defined by the range of b where they are relevant. It can be shown that $\tilde{\mu}(\hat{b}) \leq \tilde{\mu}(b^u_1) = \tilde{\tilde{\mu}}(b^u_1) \leq \tilde{\tilde{\mu}}(b^{**}) = 1$.

If the beliefs are very low, that is, if $\mu < \tilde{\mu}(\hat{b})$, then $\mu < \tilde{\mu}(b)\ \forall b$. Given that a profit maximiser faces no risk of being punished with so low beliefs, he will maximise his profits by returning $b^u_{pm} = 0$.

If, instead, $\tilde{\mu}(\hat{b}) \leq b < \tilde{\mu}(b^u_1) = \tilde{\tilde{\mu}}(b^u_1)$ a profit maximiser can either return nothing or set a minimal reward of $b^u = \frac{\mu z/2 - z + \alpha\lambda_u}{\alpha\lambda_u}$ to prevent punishment. This minimal reward can be directly obtained from the threshold $\tilde{\mu}$ and it can be shown that it is less than a fair return. In order to choose between these two rewards, a profit-maximiser allocator will compare the expected profits of such options. In particular, he will offer $b^u_{pm} = 0$ if $\mu \geq \frac{z - \alpha\lambda_u}{z/2 - \alpha\lambda_u^2}$. However, this threshold is negative, so a profit-maximiser allocator will always choose $b^u_{pm} = 0$.

Finally, if $\tilde{\mu}(b^u_1) = \tilde{\tilde{\mu}}(b^u_1) \leq b < \tilde{\tilde{\mu}}(b^{**}) = 1$, a profit maximiser can either return nothing or set a minimal reward of $b^{uu} = \frac{\mu(z/2 + \alpha(1 - 2\lambda_u)) - z + \alpha\lambda_u}{\alpha(\lambda_u + 2\mu)}$ to avoid punishment. Such b can be directly obtained from the threshold $\tilde{\tilde{\mu}}$ and it can be shown that it is less than a fair return. A profit-maximiser allocator will return $b^u_{pm} = 0$ if $-(2\alpha\lambda_u)\mu^2 + (z/2 + \alpha(1 - 2\lambda_u - \lambda_u^2)) - z + \alpha\lambda_u \geq 0$. The characteristics of the roots of this quadratic equation depend on the sign of the discriminant. If the discriminant is negative, then there are two complex roots and the allocator will always return $b^{u}_{pm} = b^{uu}$. If the discriminant is equal to zero, then there is only one real root and the profit-maximiser allocator will also return $b^{u}_{pm} = b^{uu}$ always. Finally, if the discriminant positive, then there are two real roots, one negative and one positive. In this case, for any $\mu < \mu_1^{''}$, the positive reward $b^{u}_{pm} = 0$ would be returned, whereas for any $\mu_1^{''} \leq \mu$, the reward would be $b^{u}_{pm} = b^{uu}$, where $\mu_1^{''}$ is the positive root of the quadratic equation.

■

Proof of Proposition 2:

In first place, notice that $\frac{\partial m^{cc}}{\partial b^{cc}}, \frac{\partial m^{uu}}{\partial b^{uu}} < 0$, which implies that m^{cc} and m^{uu} are both decreasing in b^{cc} and b^{uu} respectively. This implies that if $b^{cc} > b^{uu}$, then $m^{cc} < m^{uu}$.

In next place let us show when is $b^{cc} > b^{uu}$, where $b^{cc} = \frac{q(z/2+\alpha)-z}{2\alpha q}$ and $b^{uu} = \frac{q(z/2+\alpha(1-2\lambda_u))-z+\alpha\lambda_u}{\alpha(\lambda_u+2q)}$. Let us simplify these expressions by saying that $b^{cc} = \frac{A}{B}$ and $b^{uu} = \frac{A+C}{B+D}$, where A =q(z/2+α) $- z$, $B = 2\alpha q$, $C = \alpha\lambda_u - 2\alpha q\lambda_u$ and $D = \alpha\lambda_u$.

Notice that $D > 0$, but the sign of C depends on the value that q takes. In particular, if $q > 1/2$, then $C < 0$ and therefore $b^{cc} > b^{uu}$. However, if $q < 1/2$, then $C > 0$. In this case, $b^{cc} > b^{uu}$ if $\frac{A}{B} > \frac{C}{D}$, which holds when $4\alpha q^2 + (z/2 - \alpha)q - z > 0$. This quadratic equation has two real roots, one negative and one positive, being the positive root $q^* = \frac{(\alpha+z/2)+\sqrt{(\alpha-z/2)^2+16\alpha z}}{8\alpha}$. It can be proved that as $\alpha \geq z$, then q$^* < 1/2$.

Therefore, for any $q > q^*$, $b^{cc} > b^{uu}$.

Finally, the highest positive root of the cubic equation we consider in the second part of proposition 1a. can be proved to be greater than this threshold, i.e. $q_2^{'} > q^*$. Thus, for any $q \geq q_2^{'}$, $b^{cc} > b^{uu}$ always holds.

■

Chapter 6

Conclusions

This dissertation has presented three contexts where the free rider issue leads to an inefficient outcome of underprovision of the public good or the underinvestment in a joint project. The three research studies of this work propose different punishment structures, in line with the pool punishment and the coordinated punishment literature, which enhance cooperation in social dilemmas.

Chapter 3 theoretically explores how do centralized sanctioning institutions subject to moral hazard problems emerge in selfish societies that must decide on the provision of a public good. Without the sanctioning institution or if the institution has incorrect incentives, selfish citizens free ride to the public good, leading to an inefficient outcome. However, a high-performance sanctioning institution can achieve a positive public good provision. Societies with a high quality sanctioning institution and high social return of the public good will have higher levels of contribution to the public good under this type of institutions. With respect to the emergence and implementation of a high-performance sanctioning institution, the government representing the social class with the lowest opportunity cost will implement the institution in a wider range of cases than any other.

Chapter 4 experimentally studies sanctioning as a signal by analysing the impact of the payoff scheme on the behaviour of sanctioners and contributors in a centralized sanctioning environment. On the one hand, sanctioners implement punishment more frequently when their payoff depends on the performance of the contributors. Contributors, on the other hand, are more responsive to this kind of sanctions. However, this does not imply that this type of scheme leads to higher contributions and higher social welfare. The underlying reason for this is that contingent payoffs generate a lower willingness to cooperate and, consequently, a noisy sanctioning signal.

Chapter 5, in last place, theoretically compares an uncoordinated and a coordinated punishment scheme in a team trust game. In this work, I highlight the superiority of coordinated punishment in attaining cooperation when the proportion of reciprocal investors in the population is sufficiently high. The reason behind this is that, under coordinated punishment, reciprocal investors are more demanding with the rewards they receive from the allocators. Anticipating this, proft-maximiser allocators return positive rewards that avoid punishment more often which, in turn, persuades joint punishment more frequently. Hence, differences between the two types of punishment proposed lie on the free riding incentives between reciprocators. With coordinated punishment, this behaviour can be avoided when there are many reciprocators in the population, as if one of the two investors free rides, punishment actions will be costly and ineffective in destroying the allocator's payoff.

Conclusiones

Esta tesis presenta tres contextos donde el problema del polizón lleva a un resultado ineficiente de subabastecimiento de un bien público o subinversión en un proyecto conjunto. Los tres trabajos de investigación proponen diferentes estructuras de castigo que promueven la cooperación en dilemas sociales, en línea con la literatura de castigo delegado y de castigo coordinado.

El Capítulo 3 explora de manera teórica cómo instituciones sancionadoras centralizadas sujetas a problemas de riesgo moral emergen en sociedades egoístas que deben decidir sobre la provisión de un bien público. Sin la institución sancionadora o si la institución tiene incentivos inadecuados, los ciudadanos egoístas hacen de polizón al bien público, lo cual lleva a un resultado ineficiente. No obstante, una institución sancionadora de alto rendimiento puede conseguir una provisión positiva del bien. Sociedades con instituciones sancionaras de alta calidad y con un alto rendimiento de los bienes públicos podrán conseguir mayores niveles de contribución a dichos bienes. Con respecto al surgimiento e implementación de una institución sancionadora de alto rendimiento, el gobierno que represente a la clase social con el menor coste de oportunidad implementará la institución en un mayor rango de casos que cualquier otra.

El Capítulo 4 utiliza un enfoque experimental para estudiar las sanciones como una señal. Esto se realiza mediante el análisis del impacto del esquema de pago en el comportamiento de sancionadores y contribuyentes en un entorno de sanciones centralizadas. Por un lado, los sancionadores implementan más frecuentemente castigo cuando su pago depende del desempeño de los contribuyentes. Los contribuyentes, por otro lado, son más sensibles a este tipo de sanciones. Sin embargo, esto no implica que este tipo de esquema conduzca a contribuciones más elevadas y a mayor bienestar social. El motivo subyacente de esto es que los pagos contingentes generan una menor disposición a cooperar y, consecuentemente, una señal de sanción ruidosa.

El Capítulo 5, en último lugar, compara teóricamente un esquema de castigo no coordinado con un esquema coordinado en un juego de confianza en equipo. Este trabajo destaca la superioridad del castigo coordinado en conseguir cooperación cuando la proporción de inversores con nivel de reciprocidad significativa en la población es suficientemente alta. El motivo detrás de esto es que, bajo castigo coordinado, los inversores son más exigentes con las recompensas que reciben de parte de los asignadores de recursos. Anticipando esto, los asignadores de recursos que maximicen sus beneficios devolverán recompensas positivas que evitan el castigo más frecuentemente lo cual, a su vez, persuade la inversión conjunta en más casos. Por tanto, las diferencias entre los dos tipos de castigo propuestos recaen en los diferentes incentivos que los reciprocadores tienen a hacer de polizón.

Con castigo coordinado este comportamiento puede ser evitado cuando hay un número considerable de agentes con nivel de reciprocidad significativo en la población, ya que si uno de los dos inversores hace de polizón, el castigo será costoso e inefectivo en destruir el pago del asignador de recursos.

Investigar es ver lo que todo el mundo ha visto,
y pensar lo que nadie más ha pensado.

Albert Szent-Györgyi