Stable coalition formation through bargaining for the preservation of public goods

Elvio Accinelli* Atefeh Afsar† Filipe Martins‡ José Martins§
Bruno M. P. M. Oliveira¶ Jorge Oviedo‖ Alberto A. Pinto**
Luis Quintas††

August 29, 2024

Abstract

Baliga and Maskin [3] introduced a model of contributions for the provisions of public goods such as contributions for the reduction of air pollution. For an extended version of their model, we consider the formation of stable coalitions which are absorbing states of a bargaining Markov chain, where agents join/leave coalitions according to their cooperation/free-riding incentives. Following Baliga and Maskin [3], we consider heterogeneous agents with quasi-linear utilities of the form $u_j(r_j; r) = \theta_j r^\alpha - r_j$, where r is the aggregate contribution ($\alpha = 1/2$ in [3]) and the exponent α is the elasticity of the benefit function. We show that there is a stable high coalition consisting

*Facultad de Economía, Universidad Autónoma de San Luís Potosí, México, Avenida Pintores s/n Burocratas del Estado, San Luis Potosi 78213, México. E-mail: elvio.accinelli@eco.uaslp.mx

†Mathematics Department, Allen University, Columbia, SC, United States. E-mail: aafsar@allenuniversity.edu

‡Centro de Matemática, Universidade do Porto – CMUP, Rua do Campo Alegre 687, 4169-007, Porto, Portugal. E-mail: luis.f.martins@fc.up.pt

§LIAAD–INESC TEC and Escola Superior de Tecologia e Gestão, Campus 2, Morro do Lena-Alto do Vieiro, 2411-901, Leiria, Portugal. E-mail: jmmartins@ipleiria.pt

¶LIAAD–INESC TEC and Faculdade de Ciências da Nutrição e Alimentação, Universidade do Porto (FCNAUP), 4150-180, Porto, Portugal. E-mail: bmpmo@fcna.up.pt

‖Instituto de Matemática Aplicada San Luis (UNSL-CONICET) and Departamento de Matemática, Universidad Nacional de San Luis, Avenida Italia 1556, D5700, San Luis, Argentina. E-mail: joviedo12@gmail.com

**LIAAD–INESC TEC and Departmento de Matemática, Universidade do Porto, Portugal. E-mail: aapinto@fc.up.pt

††Instituto de Matemática Aplicada San Luis (UNSL-CONICET) and Departamento de Matemática, Universidad Nacional de San Luis, Avenida Italia 1556, D5700, San Luis, Argentina. E-mail: lquintas@unsl.edu.ar

of the set of agents most preferring/valuing the public good. The increase of the elasticity α increases the size of the stable high coalition that changes from a single member (called the competitive coalition as appears in Baliga and Maskin's paper) to the grand coalition involving all agents. However, the utility of members of the stable coalition can be very small when compared to the utility of the free–riders, rendering the formation of stable coalitions difficult. We show that the a variant of the coalition folk theorem holds, meaning that member heterogeneity will tend to decrease the size of stable coalitions. We show that the formation of stable coalitions is subject to the paradox of cooperation, since even when stable coalitions are large and free–riders have not very low preferences for the public good, the utility of the stable coalition may still be low when compared to the full cooperation scenario of the grand coalition. However, the paradox does not hold when the free–riders have a very low preference for the public good, which also facilitates the spontaneous formation of stable coalitions, or when there are no free–riders and the grand coalition is stable, which is always the case when the elasticity α is large enough.

- **Keywords**: public and common goods; free-riding; coalitions; stability; Barrett's paradox of cooperation; Markov chains.
- **JEL classification: C7, D7, H4.**
- **MSC2020 classification: 91-10, 91B18, 91B69, 91B76, 91A12, 91A40.**

1 Introduction

Social dilemmas are situations in which there is an explicit or implicit conflict between self-interest and collective interests. Examples include the free-riding type behaviour such as in the tragedy of the commons [19], where self-interest in the use of a good may lead to depletion of the resource. What are the basics of trust, fairness, or human cooperation? And ultimately, how can incentives work to to act in ways that serve the common interest, now and in the future? These topics and questions are at the heart of theory and research on social dilemmas and collective action (see [11] and `https://socialdilemma.com/`). The key issue around social dilemmas (arising, for instance, in the preservation of public goods and common-pool resources) is that free-riding behavior occurs. The consumption of public and common goods is non-excludable, opening the possibility of over-exploitation of the resource or under-investment and, ultimately, depletion of the common resource to everybody's detriment.

Using game theory, we provide an analysis of strategic choice when facing social dilemmas regarding public goods. The open question for individual consumption or supply of public or common property goods is whether individuals are capable of modifying their individually rational behavior in order to achieve a result providing collective benefits greater than those that individuals would receive if they act independently. The relationship between free-riding problems and the logic of collective action has been recognized in specific

contexts for centuries. In the context of collective action, individuals can form agreements and coalitions where individuals must choose a behavior to preserve or provide a public or common good whose consumption or use brings them together. One of such fundamental problems in the modern world concerns environmental protection and preservation. The signature of international environmental agreements (IEA's) and the development of institutions is of the highest relevance leading to the well-known issue of the formation of stable coalitions (see d'Aspremont *et* al. [10]) for addressing climate change goals. The so-called *paradox of cooperation* (first mentioned in Barrett [4]) and also quoted as *Barrett's paradox of cooperation*) argues that cooperation may be unsuccessful since in situations where cooperation is more important, stable coalition may achieve low results due to the existence of free-riders.

The problem of free-riding has been considered by different thinkers since the origins of social institutions, arriving at different conclusions. For instance, as back as in Plato's *Republic* [26], Glaucon recognizes the problem when he asserts that in some cases people may act against the law if they can escape sanction for the violations. As described by Garrett Hardin [19] on the tragedy of the commons, some reasons for inefficiencies in the consumption or provision of a common good lie in the property of it being a non-excludable good. A non-excludable good is such that it is not possible to effectively exclude any economic agent from its consumption, which may imply the exhaustion or the under-provision of the good. In the economic sciences literature, both public and common goods are non-excludable goods. Elinor Ostrom, in her seminal work on the commons [25], emphasized the importance of coalitions in the governance of the commons and showed examples in which the participation of external regulators was not necessary to get a community to participate in the preservation of public resources through cooperation. Her work challenged the established belief that without external intervention, either governmental or *via* private management, resources will be over-exploited. She emphasizes the difference between a problem of commons, where "(...) people can overuse, they (the sources) can be destroyed, and it is a big challenge to try to figure out how to avoid it. That is a problem, that is real", and a tragedy of commons, where "(...) they cannot, ever, solve it (...) and the only way out was some external government coming in or diving it up into small chunks and everyone owing their own (...)". Another good account and analysis of the role of collective action is provided by R. Hardin in [20], challenging some of the outcomes theorized by public choice theorists.

Quite often, when people are asked how much they value a particular public good, for instance, the value they are willing to contribute for the preservation or provision of a public good, their tendency is to under-report their valuations, or they are even unwilling to cooperate when others do if they can still enjoy the good in question (see Goodstein and Polasky [18] for problems in the context of the environment). In a review of experimental psychological studies, Kopelman *et al.* [21] identify some variables influencing behavior in dilemmas of commons. Different types of experiments (see for instance Fischbacher *et al.*

[16], and Gintis *et al.* [17]) indicate that a substantial part of the subjects facing a situation in which they can choose between cooperating or not to the supply or maintenance of a good, decide to do it even when they can enjoy it for free. Moreover, they are willing to punish those who violate the rules of cooperation, even when this implies some costs. However, the same experiments also reveal that a large part of the subjects choose not to do so. Dannenberg *et al.* [9] analyze experimentally the formation of coalitions for the provision of public goods.

The possibility of forming coalitions for the preservation of a public good calls into question the "zero contribution thesis" of Mancur Olson according to which "unless the number of individuals in a group is quite small, rational self-interested individuals will not act to achieve their common group interests" (see [24]). In a somehow opposite way, the so-called Coase theorem [8] argues that even in the presence of externalities, economic agents should still be able to ensure a Pareto-efficient outcome, provided that there are no significant constraints, such as sufficiently low transaction costs on their ability to bargain and contract. The argument is that if a prospective allocation is inefficient, agents will have the incentive to bargain their way to a Pareto improvement. Thus, even if markets themselves fail, social goals may still be achieved. Andreu Mas-Colell [22] points out that in the paradigm of cooperative game theory, informed economic agents might in principle be able to communicate freely and take joint action, such as the formation of a coalition to achieve joint goals. John E. Roemer [27] and [28] constructs a theory of Kantian optimization. He argues that under negative externalities, such as congestion effects in the use of common property resources, Kantian equilibria can be achieved even when these are threatened by the possibility of non-cooperative equilibria.

Baliga and Maskin [3] propose a model where they show that if communities enjoy the benefits of a public or common good (reduction of air pollution in their case, but can be, with suitable interpretations, other common or public goods) at no cost (because of non-excludability) there will be no agreement on how and who will pay the costs of conservation and provision of the good. They consider mechanism design methods, studying the implementation of social goals. Otherwise, the only possible agreement for the preservation of the good will be a competitive one based on a single agent contributing to the good. This is also the result if one considers an evolutionary framework for a generalized version of Baliga and Maskin's model based on standard adaptive dynamics where agents myopically adjust their contributions in an evolutionary way since individuals change their contributions according to whether a given aggregate subsistence contribution level is attained. In [1] it is shown that the single top agent contribution is the only locally asymptotically stable equilibrium. In [2] the basins of attraction of the equilibria are analyzed and characterized, and it is shown that the single top agent contribution equilibrium is essentially a global attractor.

In this paper, we study the formation of stable coalitions of agents contributing to a public good in a generalized version of Baliga and Maskin's model. Given a coalition,

all agents enjoy a benefit associated to the contributing agents that are members of the coalition, and agents in the coalition share their costs proportionally to their preferences for the public good. Agents not belonging to the coalition are free-riders, *i.e.*, they have the benefits from contributions of the coalition but do not contribute to the maintenance of the good and hence have no cost. The spontaneous formation of coalitions may be exemplified by a bargaining Markov chain that mimics a Coase's type of bargaining. The states are possible coalitions of agents and the transition probabilities are determined according to the following simple negotiation rules according to agents' incentives regarding a coalition: (i) for coalition members, if they stay in the coalition or become free-riders; (ii) for free-riders, if they keep being free-riders or join a coalition. At every moment that the coalition is modified, the new (sub)-Pareto efficient cost of maintenance of the good determined by the new coalition is computed and redistributed among its members. In Theorem 1 we show that this bargaining Markov chain is absorbing, *i.e.* it evolves with probability one to absorbing coalitions. Furthermore, a coalition is absorbing if and only if the coalition is stable, *i.e.* it is both internally and externally stable following the concepts introduced in d'Aspremont *et al.* [10]. However, there is a major drawback to the outcome of this negotiation process: (i) the utility of the free-riders would be much greater than the utility of the members of the stable coalition once the stable coalition is formed; and (ii) there is no uniqueness of the stable coalition which aggravates the problem of defining who would integrate the stable coalition to be formed. In this case, the agents that are randomly selected to decide to enter or not the coalition might prefer to postpone their decision, hindering the formation of a stable coalition. In this case, some sort of mechanism might be needed to select *a priori* the members of a stable coalition ([23]).

Following Baliga and Maskin [3], we consider an extended model of heterogenous agents with quasi-linear utilities, with an additional elasticity parameter $\alpha \in (0, 1)$ associated to the aggregate contribution. For this extended version of the model, it still holds that a single agent paying the preservation costs and all the other agents acting as free-riders is a competitive equilibrium (see [1, 2]). However, stable coalitions can have more than a single member. In Theorem 2 we derive necessary and sufficient conditions for the stability of a coalition, and we thoroughly discuss these conditions in relation to the elasticity α and an associated free-rider threshold. In Theorem 3, we show that there is a unique M for which the coalition formed by the M agents with the highest valuations is stable. If there is another coalition that is stable and differing from the stable high coalition, then this coalition is not quite different from the stable M-high coalition: some of the agents that most prefer the public good have to be part of the coalition and some of the agents that less prefer the good have to be free-riders; only some agents might have been substituted by other agents with similar preferences for the public good. In Theorem 4, we show that the size of stable coalitions differs from that of the stable high coalition at most by one. We call them challenging coalitions when they have one more agent than the high coalition because they have higher welfare than the high coalition. We provide some examples showing the

existence of such stable challenging coalitions.

We also study the effects of homogeneity and heterogeneity of the preferences of the agents on the sizes of the stable coalitions (see Finus and McGinty [15]). We show that the coalition folk theorem holds in the context of our model since a high degree of homogeneity between agents leads to stable coalitions with the highest possible size, which depends on the utility elasticity α. Furthermore, we carefully analyze Barrett's paradox of cooperation, exhibiting the relative minimum utility and welfare of stable coalitions when compared to the grand coalition. We observe that these minima only depend on the number of free-riders and on the utility elasticity. The failure of the paradox of cooperation occurs only when the grand coalition is stabilized or when free-riders have a very low preference for the public good, also making the spontaneous formation of stable coalitions easier.

We also consider a game whose players are a single coalition and all the other agents that do not belong to the coalition (see [6, 12, 13, 14, 29]). The utility of the coalition is the aggregate utility of its members. We prove that for coalitions with more than two agents, the Nash–Cournot and Stackelberg equilibria (when the coalition is the leader) are the (sub)-optimal aggregate contribution of the coalition with all the other players being free-riders. In particular, when the stable coalition has cardinality greater than one, the Nash–Cournot and Stackelberg equilibria are the optimal contribution of the stable coalition with all the other players being free-riders. This shows that the difficulty might be to convince the agents to belong to the stable coalition when the free riders have a higher utility, since there might be another stable coalition where an agent that is in the coalition would be a free–rider and *vice-versa*.

Summarizing, in this paper we offer a unified way to analyze the possibilities of cooperation for the preservation of a common pool resource or a public good depending on the utility elasticity α. We have mainly three scenarios:

1. (Low utility elasticity) all agents are free-riders except the one that most prefers the good and so the Baliga and Maskin competitive equilibrium holds;

2. (Medium utility elasticity) M-high coalitions are stable for some integer $M \geq 2$, signifying some possible cooperation. However, the utility of the members of the stable coalition can be very small compared to the utility of the free-riders. Furthermore, some instances of the paradox of cooperation may occur, with the exception occurring when the free-riders have very low preferences for the public good relative to the preferences of the members of the high coalition, which also facilitates the formation of the relatively high-welfare stable coalitions, and so an Olson's type of argument holds;

3. (High utility elasticity) the grand coalition is the unique stable coalition and so a Coasian type of argument holds. In this case, arguments such as those presented by E. Ostrom about the possibility of "spontaneous cooperation for the preservation of the common good" would be justified.

This paper is organized in the following way. In section 2 we consider a generalization of Baliga and Maskin's model of contributions for a public good and we characterize Nash–Cournot and Stackelberg equilibria. In section 3 we consider coalitions and their cost distributions, and we discuss the formation of stable coalitions through a bargaining Markov chain. In section 4 we study the coalition stability. In section 5 we study stable high coalitions and challenging coalitions. In section 6, we discuss the coalition folk theorem and the paradox of cooperation in the context of the model. We finish with some conclusions and final remarks in section 7. We present the proofs of the results in the appendix A.

2 Coalitions for a generalized Baliga–Maskin model

In this section, we will introduce a generalization of Baliga and Maskin's model of contributions to a public/common good model ([3]) and we introduce coalitions of agents contributing to the good. Following current literature (see [6, 12, 13, 14, 29]), we consider a *public goods games* where one of the players is a coalition, with utility being the aggregate utility of the coalition members, and the other players are all the other agents that do not belong to the coalition. In particular, when the coalition is formed by a single agent, it consists of a game where all the players are the agents.

We consider two variants of the public goods game: simultaneous (Nash–Cournot) and sequential (Stackelberg). In the *simultaneous* Nash-Cournot game, all the players choose their contributions simultaneously. In the *sequential* Stackelberg game the coalition chooses its contribution first, *i.e.* the coalition is the leader, and all the other agents will follow by choosing their contributions afterward. We will characterize equilibria of the two variants of the game.

2.1 A generalized Baliga–Maskin model

As in Baliga and Maskin's model, we consider that there are N agents (or communities according to the nomenclature used in their paper), that can be countries, individuals, or in game theory terms, players. They are indexed by $j \in \mathcal{N} = \{1, 2, \ldots, N\}$ and they are users or consumers of a public good (the public good is the reduction of pollution itself in Baliga and Maskin's model). They have different preferences over the quality or preservation of this good. These differences in preferences for the social alternatives are characterized by the *preference parameter* $\theta_j > 0$ for each $j \in \mathcal{N}$, in such way that a higher θ_j indicates a higher preference for the good. Without loss of generality, we consider

$$\theta_1 \geq \theta_2 \geq \cdots \geq \theta_N > 0.$$

The contribution of agent j for the preservation of the public good is denoted by $r_j \geq 0$. The quantity $r = \sum_{j=1}^{N} r_j$ is the *aggregate effort or contribution* for the public good. The

aggregate contribution of all agents except j is denoted by $r_{-j} = \sum_{i \in \mathcal{N} \setminus \{j\}} r_i$. Hence, $r = r_j + r_{-j}$.

The *benefit* of agent j is

$$v_j(r) \equiv v(r; \theta_j) \equiv \theta_j r^\alpha \ .$$

Note that all agents benefit from the public good, even the ones that eventually do not contribute. We observe that the benefit functions are strictly concave since $\alpha \in (0, 1)$ ($\alpha = 1/2$ in Baliga and Maskin [3]). We observe that

$$\alpha = \lim_{\Delta r \to 0} \frac{\Delta v_j}{v_j} / \frac{\Delta r}{r} = \frac{dv_j}{dr} \frac{r}{v_j(r)} \ .$$

and so the parameter α is the ratio of the percentage change in the benefit to the percentage change in the aggregate effort, that is, it is the elasticity of the benefit/utility with respect to the aggregate contribution.

The *net utility* of agent j is the quasi-linear utility function

$$u_j(r_j; r_{-j}) \equiv v_j(r) - r_j = \theta_j (r_j + r_{-j})^\alpha - r_j \ .$$

We have that $\partial u_j / \partial r_j = 0$ if and only if

$$r = \overline{r}_j \equiv (\alpha \theta_j)^{\frac{1}{1-\alpha}} \ ,$$

and so $\overline{r}_j$ is the aggregate effort maximizing the net utility of agent j. We call $\overline{r}_j$ the *stand-alone effort* of agent j since it is the optimal effort $r_j = \overline{r}_j$ of agent j when all the other agents do not contribute, *i.e.* $r_{-j} = 0$. Clearly, we have that

$$\overline{r}_1 \geq \overline{r}_2 \geq \cdots \geq \overline{r}_N > 0 \ .$$

2.2 Coalitions

A *coalition* $\mathcal{A} \subset \mathcal{N}$ is a subset of agents that are willing to participate in a cooperation agreement for the conservation or provision of a public good. The agents in $\mathcal{N} \setminus \mathcal{A}$ are called the *free-riders* relative to $\mathcal{A}$. Two examples are the *grand coalition* $\mathcal{N}$ where all agents participate, and the *singleton coalitions* $\{i\}$ constituted by a single agent. We will use the notation $\#\mathcal{A}$ to denote the number of agents of a coalition $\mathcal{A}$.

The *effort of coalition* $\mathcal{A}$ is the sum $r_\mathcal{A} = \sum_{j \in \mathcal{A}} r_j$ of the efforts of all the agents in coalition $\mathcal{A}$. The quantity $r_{-\mathcal{A}} = r_\mathcal{N} - r_\mathcal{A}$ is the contribution of all the free-riders relative to $\mathcal{A}$. In particular, we have that

$$r = r_\mathcal{N} = r_\mathcal{A} + r_{-\mathcal{A}}.$$

The *preference of coalition* $\mathcal{A}$ is given by $\Theta_\mathcal{A} = \sum_{j \in \mathcal{A}} \theta_j$, and the $\mathcal{A}$-*conditional preference* of agent $j \in \mathcal{N}$ relative to coalition $\mathcal{A}$ is $\theta_{j/\mathcal{A}} = \theta_j / \Theta_\mathcal{A}$.

The *net utility* of coalition $\mathcal{A}$

$$u_{\mathcal{A}}(r_{\mathcal{A}}; r_{-\mathcal{A}}) \equiv \sum_{j \in \mathcal{A}} u_j(r_j; r_{-j}) = \Theta_{\mathcal{A}}(r_{\mathcal{A}} + r_{-\mathcal{A}})^{\alpha} - r_{\mathcal{A}} \tag{1}$$

is the sum of the net utilities of all the agents of coalition $\mathcal{A}$. We have that $\partial u_{\mathcal{A}}/\partial r_{\mathcal{A}} = 0$ if and only if

$$r = \overline{r}_{\mathcal{A}} \equiv (\alpha \Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}} \ .$$

and so $\overline{r}_{\mathcal{A}}$ is the aggregate effort maximizing the net utility of coalition $\mathcal{A}$. We call $\overline{r}_{\mathcal{A}}$ the *stand-alone effort* of coalition $\mathcal{A}$ since it is the optimal effort $r_{\mathcal{A}} = \overline{r}_{\mathcal{A}}$ of coalition $\mathcal{A}$ when all free-riders do not contribute $r_{-\mathcal{A}} = 0$. Clearly, $\overline{r}_{\mathcal{A}} \leq \overline{r}_{\mathcal{N}}$. Observe that the stand-alone effort $\overline{r}_{\mathcal{A}}$ is a sub-optimal effort for a given coalition, being optimal when $\mathcal{A} = \mathcal{N}$.

2.3 Nash–Cournot and Stackelberg equilibria of the public goods game

We now consider simultaneous and sequential games where the players are the following: (i) a *formed* coalition $\mathcal{F}$, with utility being the net utility of the coalition as defined above; (ii) all the free-riders $j \in \mathcal{N} \setminus \mathcal{F}$ relative to $\mathcal{F}$. We will determine Nash–Cournot equilibria when all players choose their efforts simultaneously, and Stackelberg equilibria when the coalition chooses its effort first, and all the other players move simultaneously after the coalition player.

We first define two notions that will be useful to analyze the equilibria of the public goods game. A *low-cooperation* strategy of a coalition $\mathcal{A}$ is a vector of individual efforts with the following two properties: i) the effort $r_{\mathcal{A}}$ of coalition $\mathcal{A}$ equals the stand-alone effort of the top agent, *i.e.*, agent 1:

$$r_{\mathcal{A}} = \overline{r}_1 = (\alpha \theta_1)^{\frac{1}{1-\alpha}} \ ;$$

ii) all the free-riders relative to $\mathcal{A}$ do not contribute:

$$r_{-\mathcal{A}} = 0.$$

A *stand-alone* strategy of a coalition $\mathcal{A}$ is a vector of individual efforts with the following two properties: i) the effort $r_{\mathcal{A}}$ of coalition $\mathcal{A}$ equal its stand-alone effort $\overline{r}_{\mathcal{A}}$:

$$r_{\mathcal{A}} = \overline{r}_{\mathcal{A}} = (\alpha \Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}};$$

ii) all the free-riders relative to $\mathcal{A}$ do not contribute:

$$r_{-\mathcal{A}} = 0.$$

For low-cooperation and stand-alone strategies, the individual efforts of free-riders are determined (equal to zero), while for coalition members they are not. Indeed, it is possible that coalition members $i \in \mathcal{A}$ do not contribute, *i.e.*, $r_i = 0$.

We begin by characterizing the best response functions of both types of players in the following lemma.

Lemma 1. *The best response function of the coalition player $\mathcal{F}$ is*

$$r^*_{\mathcal{F}}(r_{-\mathcal{F}}) = \begin{cases} 0, & \text{if } r_{-\mathcal{F}} \geq \overline{r}_{\mathcal{F}} \\ \overline{r}_{\mathcal{F}} - r_{-\mathcal{F}}, & \text{if } r_{-\mathcal{F}} < \overline{r}_{\mathcal{F}} \end{cases} .$$

The best response functions of free-riders $j \in \mathcal{N} \setminus \mathcal{F}$ are

$$r^*_j(r_{-j}) = \begin{cases} 0, & \text{if } r_{-j} \geq \overline{r}_j \\ \overline{r}_j - r_{-j}, & \text{if } r_{-j} < \overline{r}_j \end{cases} .$$

We now analyze the equilibria of the Nash–Cournot game. Let $\mathcal{T}$ be the coalition of all the (top) agents i that most prefer the good, *i.e.*, $i \in \mathcal{T}$ if and only if $\theta_i = \theta_1$.

Lemma 2. *The Nash–Cournot equilibria are the following:*

1. *if $\Theta_{\mathcal{F}} < \theta_1$, every low-cooperation strategy of the top agents coalition $\mathcal{T}$;*
2. *if $\Theta_{\mathcal{F}} = \theta_1$, every low-cooperation strategy of the coalition $\mathcal{T} \cup \mathcal{F}$;*
3. *if $\Theta_{\mathcal{F}} > \theta_1$, every stand-alone strategy of coalition $\mathcal{F}$.*

When $\mathcal{F}$ is a singleton coalition, the game consists of a competitive (simultaneous) game played by all the agents. In this scenario, only cases 1 or 2 of Lemma 2 occur. So competitive equilibria consist of low-cooperation strategies. In the case where all the preferences θ_i are distinct, we obtain the competitive equilibria derived by Baliga and Maskin in this context, where only the top agent contributes. Observe that in case 2, a stand-alone strategy of $\mathcal{F}$ is a particular example of a low-cooperation strategy of $\mathcal{T} \cup \mathcal{F}$. So we conclude that if the top agent is a member of $\mathcal{F}$, then a stand-alone strategy of $\mathcal{F}$ is always Nash–Cournot equilibrium.

We now analyze the equilibria of the Stackelberg game. The case where $\mathcal{F} = \mathcal{N}$ is trivial, since there are no free-riders, and the equilibria are stand-alone strategies of the grand coalition $\mathcal{N}$ by the previous lemma. So in case 1 below we assume that $\mathcal{F} \neq \mathcal{N}$, and so there are free-riders playing the second stage of the Stackelberg game. Take the smallest $k \in \mathcal{N} \setminus \mathcal{F}$ and let $\theta(\mathcal{F}) = \theta_k$. Let $\mathcal{T}(\mathcal{F})$ be the set of agents $j \in \mathcal{N} \setminus \mathcal{F}$ such that $\theta_j = \theta(\mathcal{F})$. In other words, $\mathcal{T}(\mathcal{F})$ is the coalition of the top free-riders relative to $\mathcal{F}$, *i.e.* the free-riders relative to $\mathcal{F}$ having the highest valuation, which is given by $\theta(\mathcal{F})$.

Lemma 3. *The Stackelberg equilibria are the following:*

1. *if*
$$\Theta_{\mathcal{F}}(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq \theta(\mathcal{F}),$$
every low-cooperation strategy of the top free-riders coalition $\mathcal{T}(\mathcal{F})$;

2. *if*
$$\theta(\mathcal{F}) \leq \Theta_{\mathcal{F}}(1-\alpha)^{\frac{1-\alpha}{\alpha}},$$
every stand-alone strategy of coalition $\mathcal{F}$.

We observe that $e^{-1} < (1-\alpha)^{\frac{1-\alpha}{\alpha}} < 1$ and is an increasing function of $\alpha \in (0,1)$. As a result, we have that lower values of α may generate a situation that falls in case 1 of the previous lemma, in which case the burden of the contributions for the good lies on (top) free-riders relative to $\mathcal{F}$, with the coalition using its leader advantage in not contributing. Moreover, if $\Theta_{\mathcal{F}} < \theta(\mathcal{F})$, then case 1 of Lemma 3 holds and also $\mathcal{T} \cap \mathcal{F} = \emptyset$. This occurs, for instance when $\mathcal{F}$ is a singleton coalition of a non-top agent. So we have that $\theta(\mathcal{F}) = \theta_1$ and $\mathcal{T}(\mathcal{F}) = \mathcal{T}$ and the Stackelberg equilibria are low-cooperation strategies of the coalition of top agents $\mathcal{T}$.

3 Coalition stability and bargaining

We will present a Markov chain providing a Coasian type of bargaining for the formation of stable coalitions. This Coasian type of bargaining consists of random successive negotiations among the agents without transaction costs to improve their net utilities by joining or leaving a coalition. We prove that this Markov chain is absorbing and that the absorbing states are stable coalitions. When the stable coalition is neither the grand coalition nor the singleton coalition, then the argument does not lead to the Pareto optimum of the grand coalition but can be much better than the result obtained by the competitive equilibrium attained by the singleton coalition formed only by the agent that most prefers the public good.

From now on we restrict our analysis to *focal stand-alone* strategies of coalitions that we will describe in the next subsection. We recall that these strategies are Nash–Cournot and Stackelberg equilibria except in the cases pointed out in Lemmas 2 and 3.

3.1 Distribution of costs and utilities

As we have mentioned above, stand-alone strategies of a coalition do not determine the individual efforts of coalition members, with only the aggregate effort of the coalition being determined and equal to the stand-alone effort of the coalition. In other words, it does not determine how costs are shared among coalition members. Following Baliga and Maskin

([3]) we consider that the cooperation agreement is that the distribution of costs is proportional to the preferences of each agent, *i.e.*, according to the $\mathcal{A}$-conditional preferences. This is sometimes called a *focal* coalition cost structure in the related literature.

The contribution $r_{j/\mathcal{A}}$ of agent $j \in \mathcal{A}$ is

$$r_{j/\mathcal{A}} \equiv \theta_{j/\mathcal{A}} \overline{r}_{\mathcal{A}} = (\theta_j / \Theta_{\mathcal{A}}) \, (\alpha \Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}} = \alpha \theta_j \, (\alpha \Theta_{\mathcal{A}})^{\frac{\alpha}{1-\alpha}} = \alpha \theta_j \overline{r}_{\mathcal{A}}^{\alpha} \, .$$

Since we are considering stand-alone strategies, the contribution of each free-rider $j \in \mathcal{N} \setminus \mathcal{A}$ is $r_{j/\mathcal{A}} = 0$. Given that coalition $\mathcal{A}$ is formed, the utility of each agent j derived from the coalition is

$$u_{j/\mathcal{A}} \equiv u(j/\mathcal{A}) \equiv \begin{cases} \theta_j \overline{r}_{\mathcal{A}}^{\alpha} (1-\alpha) & \text{if } j \in \mathcal{A} \\ \theta_j \overline{r}_{\mathcal{A}}^{\alpha} & \text{if } j \notin \mathcal{A} \end{cases} . \tag{2}$$

These determine the valuation functions of each agent for a given coalition $\mathcal{A}$.

Given a coalition $\mathcal{A}$ we observe that the utility ratio between a free-rider $i \notin \mathcal{A}$ and a member of the coalition $j \in \mathcal{A}$ is

$$\frac{u(i/\mathcal{A})}{u(j/\mathcal{A})} = \frac{\theta_i}{\theta_j} \, (1-\alpha)^{-1} \, .$$

Hence, when α goes to one the above ratio grows without bound, which *a priori* may make it more difficult in making free-riders join the agreements and may hinder the formation of larger stable coalitions.

The utility attained by a stand-alone strategy of coalition $\mathcal{A}$ is $U(\alpha; \mathcal{A}) \equiv u_{\mathcal{A}}(\overline{r}_{\mathcal{A}}; 0)$ which is the *aggregate utility of coalition* $\mathcal{A}$ and which equals

$$U(\alpha; \mathcal{A}) = \Theta_{\mathcal{A}} \overline{r}_{\mathcal{A}}^{\alpha} - \overline{r}_{\mathcal{A}} = r_{\mathcal{A}}^{\alpha} \Theta_{\mathcal{A}} \, (1-\alpha) = \overline{r}_{\mathcal{A}} \left(\frac{1}{\alpha} - 1 \right) = (\alpha \Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}} \left(\frac{1}{\alpha} - 1 \right) . \tag{3}$$

We observe that the aggregate utility of the coalition increases when the preference of the coalition increases. We observe that: (i) $\lim_{\alpha \to 0} \overline{r}_{\mathcal{A}} = 0$; and (ii) $\lim_{\alpha \to 1} \overline{r}_{\mathcal{A}}$ is equal to: (a) $+\infty$, if $\Theta_{\mathcal{A}} > 1$; (b) 0, if $\Theta_{\mathcal{A}} < 1$; (c) $1/e$, if $\Theta_{\mathcal{A}} = 1$. Furthermore, we also have that: (i) $\lim_{\alpha \to 0} u(\alpha; \mathcal{A}) = \Theta_{\mathcal{A}}$; and (ii) $\lim_{\alpha \to 1} u(\mathcal{A}; \alpha)$ is equal to: (a) $+\infty$, if $\Theta_{\mathcal{A}} > 1$; (b) 0, if $\Theta_{\mathcal{A}} \leq 1$. Hence, the elasticity α has different consequences for the stand-alone effort of the coalition and for the aggregate utility of the coalition when it is close to 1, depending on the aggregate preference of the coalition. To address this we will consider *relative utilities* of coalitions $\mathcal{A}$ and $\mathcal{B}$:

$$U(\alpha; \mathcal{A}/\mathcal{B}) \equiv \frac{U(\alpha; \mathcal{A})}{U(\alpha; \mathcal{B})} = \left(\frac{\Theta_{\mathcal{A}}}{\Theta_{\mathcal{B}}} \right)^{\frac{1}{1-\alpha}} . \tag{4}$$

Relative utilities (instead of the utilities themselves) are scale invariant with respect to the preferences, and so can be used as an indicator to measure the efficiency of coalitions to

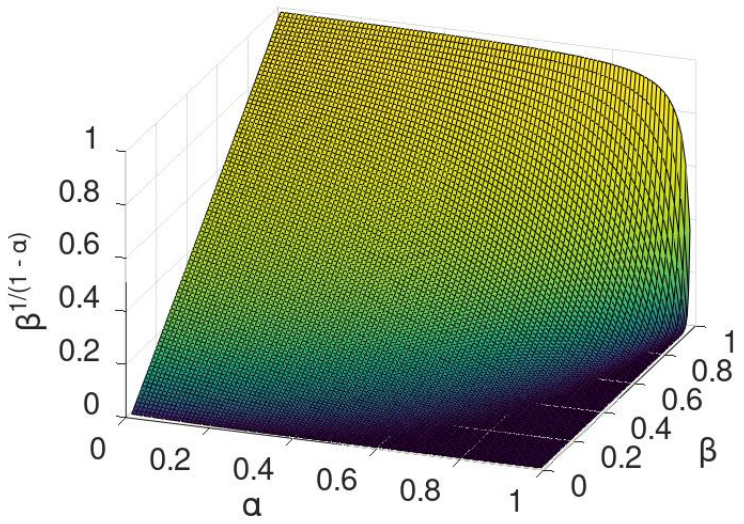

Figure 1: The relative utility $U(\alpha; \mathcal{A}/\mathcal{B})$, where $0 < \beta = \Theta_{\mathcal{A}}/\Theta_{\mathcal{B}} < 1$.

preserve the public good. In Figure 1, we plot the relative utility as a function of α and $\beta = \Theta_{\mathcal{A}}/\Theta_{\mathcal{B}} \in (0, 1)$. We observe that the relative utility decreases with α.

We will also consider the *total welfare* associated to a stand-alone strategy of a coalition $\mathcal{A}$ given by the total net utility of the coalition $\mathcal{A}$ and the sum of the utilities of the free-riders $j \notin \mathcal{A}$. Using (2) and (3) we get

$$W(\alpha; \mathcal{A}) \equiv U(\alpha; \mathcal{A}) + \sum_{j \notin \mathcal{A}} u_{j/\mathcal{A}} = \Theta_{\mathcal{N}} \overline{r}_{\mathcal{A}}^{\alpha} - \overline{r}_{\mathcal{A}}^{\alpha} = \overline{r}_{\mathcal{A}}^{\alpha} \left(\Theta_{\mathcal{N}} - \alpha \Theta_{\mathcal{A}}\right) \ . \tag{5}$$

The *relative welfare* $W(\alpha; \mathcal{A}/\mathcal{B})$ between coalitions $\mathcal{A}$ and $\mathcal{B}$ is

$$W(\alpha; \mathcal{A}/\mathcal{B}) \equiv \frac{W(\alpha; \mathcal{A})}{W(\alpha; \mathcal{B})} = \left(\frac{\Theta_{\mathcal{A}}}{\Theta_{\mathcal{B}}}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{\Theta_{\mathcal{N}} - \alpha \Theta_{\mathcal{A}}}{\Theta_{\mathcal{N}} - \alpha \Theta_{\mathcal{B}}}\right) \ . \tag{6}$$

3.2 Stability of coalitions

Given a coalition, it is essential to know if agents are able to increase their utilities by free-riding on the behavior of the cooperators in the coalition, or if agents would like to join the coalition.

After d'Aspremont *et al.* [10], a coalition $\mathcal{A}$ is *internally stable* if all members $j \in \mathcal{A}$ of the coalition $\mathcal{A}$ prefer not to become free-riders, *i.e.*, if

$$u(j/\mathcal{A}) > u(j/(\mathcal{A} \setminus \{j\}));$$

and a coalition $\mathcal{A}$ is *externally stable*, if all free-riders $j \notin \mathcal{A}$ prefer not to become members of $\mathcal{A}$, *i.e.* if

$$u(j/\mathcal{A}) \geq u(j/(\mathcal{A} \cup \{j\})).$$

A coalition $\mathcal{A}$ is *stable* if and only if it is both internally and externally stable.

3.3 A bargaining Markov chain model for Coase's argument

To illustrate a bargaining type of argument, we will make use of a Markov chain modeling bargaining and membership of a coalition. The state space of the *bargaining* Markov chain is the set of all possible coalitions $\mathcal{A} \subset \mathcal{N}$. Suppose that for every coalition $\mathcal{A}$, every agent $j \in \mathcal{N}$ has a positive probability $p_{j/\mathcal{A}} > 0$ to be selected and so $\sum_{j \in \mathcal{N}} p_{j/\mathcal{A}} = 1$. Suppose the Markov chain is at state $\mathcal{A}$: (i) the transition probability from $\mathcal{A}$ to $\mathcal{A} \cup \{j\}$ is $p_{j/\mathcal{A}}$ if $j \notin \mathcal{A}$ does not want to continue to be a free-rider, or in other words wants to join the coalition, *i.e.*, if

$$u(j/\mathcal{A} \cup \{j\}) > u(j/\mathcal{A});$$

(ii) the transition probability from $\mathcal{A}$ to $\mathcal{A} \setminus \{j\}$ is $p_{j/\mathcal{A}}$, if $j \in \mathcal{A}$ wants to become a free-rider, or in other words wants to leave the coalition, *i.e.*, if

$$u(j/\mathcal{A} \setminus \{j\}) \geq u(j/\mathcal{A});$$

(iii) otherwise, the Markov chain remains at state $\mathcal{A}$.

An *absorbing state* of a Markov chain is a state that is forward invariant for the (stochastic) dynamics of the Markov chain. For the bargaining Markov chain an absorbing state is a coalition $\mathcal{A}$ that has transition probability 1 to itself, and so a coalition $\mathcal{A}$ is an absorbing state if and only if $\mathcal{A}$ is (internally and externally) stable. A Markov chain is said to be an *absorbing Markov chain* if it has some absorbing state, and from every state the Markov chain reaches an absorbing state with positive probability. In such case, it is proven ([7]) that the Markov chain reaches some absorbing state with probability 1, and non-absorbing states are called *transient* states, *i.e.*, the process only spends a finite time in such states.

Theorem 1. *The bargaining Markov chain is absorbing. Moreover, for $\alpha \geq 1/2$, all the absorbing states are also stable Nash–Cournot equilibria of the public goods game, and for $\alpha \geq 0.732\ldots$, all the absorbing states are stable Stackelberg equilibria of the public goods game.*

For $\alpha < 1/2$, a coalition $\mathcal{A}$ can be stable, with $\Theta_{\mathcal{A}} < \theta_1$, but by Lemma 2, it is not a Nash–Cournot equilibrium. For a stable coalition $\mathcal{A}$, we prove that for $\alpha \geq 1/2$ one necessarily has $\Theta_{\mathcal{A}} \geq \theta_1$, and so it is a Nash–Cournot equilibrium by Lemma 2, and that for $\alpha \geq 0.732\ldots$ one necessarily has $\theta(\mathcal{A}) \leq \Theta_{\mathcal{A}}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$, and so it is a Stackelberg equilibrium by Lemma 3.

One important step in the proof of the above theorem is the existence of a stable M-high coalition $\mathcal{H}_M$ (see section 5, Theorem 3) consisting of the M agents with the highest valuations. When the stable M-high coalition $\mathcal{H}_M$ is the unique stable coalition (for instance, when the condition in Theorem 3 holds), then it is the unique absorbing state of the Markov chain. However, as we will explore in section 5, there may be more than one stable coalition.

So we conclude that the bargaining process stops with probability one, resulting in the formation of a stable coalition. When the grand coalition is stable, it is the unique stable coalition and bargaining leads to the formation of the grand coalition. When the grand coalition is not stable bargaining leads to the formation of a stable coalition which is not the grand coalition.

4 Free-rider threshold and stability

In this section, we study and characterize the stability of coalitions.

We define the following two quantities:

i) the *free-rider threshold* $F : (0,1) \to (0, 1-1/e)$ given by

$$F \equiv F(\alpha) \equiv 1 - (1-\alpha)^{\frac{1-\alpha}{\alpha}}.$$

ii) the *agent cap* $\ell(\alpha) \equiv \ell(F) \equiv \ell(F(\alpha)) \equiv \lceil 1/F - 1 \rceil \in \mathbb{N}$ is the unique integer such that

$$\ell(F) \in \left[\frac{1}{F} - 1, \frac{1}{F}\right).$$

The terminology adopted will be clear from the following theorem. We observe that $F(\alpha)$ is a decreasing function whose range is $(0, 1-1/e)$, where $1 - 1/e \approx 0.63 > 0.5$ (see Figure 3). We observe that: (i) the agent cap $\ell(F(\alpha))$ is an increasing step function of the elasticity parameter α; (ii) it is left-continuous, and in particular, $\ell(F(\alpha)) = 1$ for $\alpha \leq 0.5$ (recall that 0.5 is the value used by Baliga and Maskin [3]); (iii) $\ell(F(\alpha))$ tends to infinity when α tends to 1; and (iv) from the definition, we see that $\ell(F(\alpha))$ is bounded by the functions $1/F(\alpha)$ and $(1 - F(\alpha))/F(\alpha)$, and the latter coincides with the values of $\ell(F(\alpha))$ at its discontinuity points (see Figure 2).

Theorem 2 (Stability and cardinality)**.** *If $\mathcal{A}$ is stable, then $\#\mathcal{A} \leq \ell(F)$. Furthermore, the coalition $\mathcal{A}$ is stable if and only if: (i) $\mathcal{A}$ is internally stable:*

$$\theta_{j/\mathcal{A}} > F$$

for every $j \in \mathcal{A}$; and (ii) $\mathcal{A}$ is externally stable:

$$\theta_{j/\mathcal{A}} \leq \frac{F}{1-F}$$

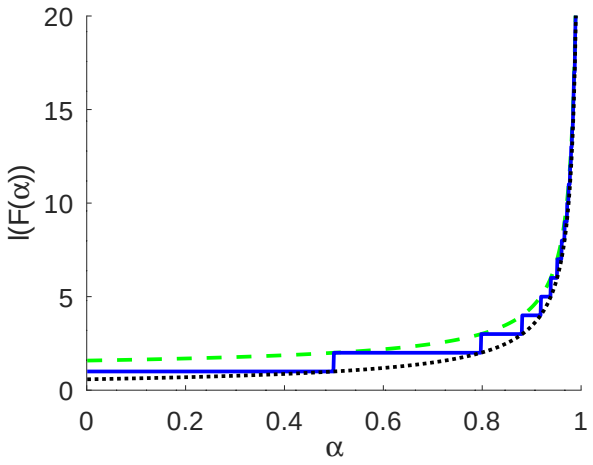

Figure 2: The agent cap $\ell(F(\alpha))$ (blue line), with its associated bounds: $1/F(\alpha)$ (green dashes) and $(1-F(\alpha))/F(\alpha)$ (black dots).

for every $j \notin \mathcal{A}$.

We observe that coalition stability is a scale-invariant property, *i.e.* it is invariant if preferences are multiplied by a positive constant. The crucial elements that determine the stability of the coalition $\mathcal{A}$ are: (i) the heterogeneity of the agents measured by the $\mathcal{A}$-*conditional preferences* $\theta_{j/\mathcal{A}}$; and (ii) the value of the elasticity α measured by the free-rider threshold $F(\alpha)$. The conclusions of Theorem 2 may be rewritten as in the following corollary.

Corollary 1. *The heterogeneity of the agents has a strong impact on the formation of stable coalitions* $\mathcal{A}$*: (i) agents with* small preferences *are free-riders: if* j *is such that*

$$\theta_{j/\mathcal{A}} \leq F$$

then $j \notin \mathcal{A}$*; (ii) agents with* intermediate preferences *are indifferent between being free-riders or belonging to the stable coalition* $\mathcal{A}$*: if* j *is such that*

$$F < \theta_{j/\mathcal{A}} \leq \frac{F}{1-F},$$

then j *can be a free-rider or belong to the stable coalition* $\mathcal{A}$*; and (iii) agents with* high preferences *belong to the stable coalition* $\mathcal{A}$*: if* j *is such that*

$$\theta_{j/\mathcal{A}} > \frac{F}{1-F}$$

then $j \in \mathcal{A}$.

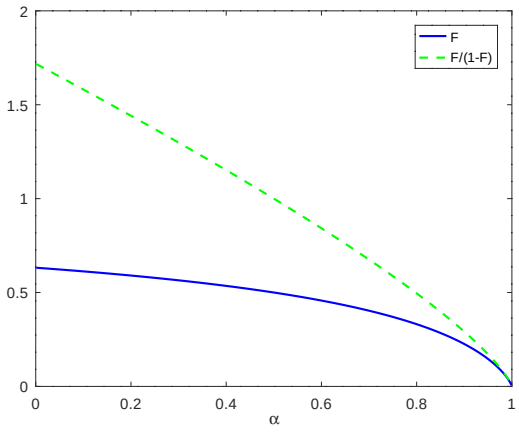

Figure 3: Plots of the bounds $F(\alpha)$ (blue line) and $F(\alpha)/(1-F(\alpha))$ (green dashes) of the small, intermediate and high preferences regions of Corollary 1.

In Figure 3, we exhibit regions of: (i) small preferences; (ii) intermediate preferences; and (iii) high preferences as functions of the parameter α. In particular, we observe that the size of the regions with intermediate and small preferences decreases when the elasticity α increases. Furthermore, the grand coalition is stable for sufficiently high values of the elasticity α.

5 Stable high and challenging coalitions

In this section, we show that there is a unique stable high coalition $\mathcal{H}_M$, consisting of the M agents with the highest preferences, for some integer M. The value of M increases with the elasticity α. Furthermore, we show that all the other stable coalitions have either M or $M+1$ agents. In the case where the stable coalition $\mathcal{A}$ has $M+1$ agents, we call $\mathcal{A}$ a *challenging* coalition of the stable high coalition $\mathcal{H}_M$, since we also prove that $\Theta_{\mathcal{A}} > \Theta_{\mathcal{H}_M}$ and so the challenging stable coalition $\mathcal{A}$ has a greater utility and welfare than the stable high coalition $\mathcal{H}_M$. We also show that a challenging coalition $\mathcal{A}$ does not contain the stable high coalition $\mathcal{H}_M$.

We show that the Baliga–Maskin competitive singleton equilibrium is a stable coalition when α is low; and the grand coalition is the unique stable coalition when α is high.

5.1 Stable high coalitions

The *M-high coalition* $\mathcal{H}_M$ is the coalition with agents $\{1, 2, \ldots, M\}$, *i.e.* those M agents with the highest preferences for the good. The grand coalition $\mathcal{H}_N = \mathcal{N}$ is the coalition

where all agents participate, and the singleton coalition $\mathcal{H}_1 = \{1\}$ is the coalition formed by the single agent that most prefers or values the good (the competitive coalition).

The observation following Lemma 2 shows that $\mathcal{H}_M$ is always a Nash–Cournot equilibrium since it contains a top agent. Easy computations show that if $M = 2$ and $\alpha > 1/2$ then $\mathcal{H}_M$ is a Stackelberg equilibrium, and if $M \geq 3$ then $\mathcal{H}_M$ is a Stackelberg equilibrium for every α. We observe that if $M = 1$ and there is another agent with preference equal to θ_1 then case 1 of Lemma 3 always holds and so $\mathcal{H}_1$ is not a Stackelberg equilibrium.

Theorem 3 (Stable high coalitions)**.** *There is a unique M for which the M-high coalition $\mathcal{H}_M$ is stable. Furthermore, if $\theta_{M+1} \leq (1-F)\theta_M$ then $\mathcal{H}_M$ is the only stable coalition.*

Noting that $\theta_{2/\mathcal{H}_2} \leq 0.5 < 0.63 \approx 1 - 1/e$ and putting together Theorems 2 and 3, we obtain that the $\mathcal{H}_M$-conditional preferences $\theta_{M/\mathcal{H}_M}$ determine a partition of the free-rider threshold range, and so they determine a partition of the elasticity α characterizing the cardinality of the stable high coalition.

Corollary 2. *Let $\mathcal{N} = \{1, \ldots, N\}$ be the grand coalition. The $\mathcal{H}_M$-conditional preferences $\theta_{M/\mathcal{H}_M}$ form a N-partition $0 < \theta_{N/\mathcal{H}_N} < \ldots < \theta_{2/\mathcal{H}_2} < 1 - 1/e \approx 0.63$ of the range of the free-rider threshold $F = F(\alpha)$ with the following properties:*

(i) the grand coalition $\mathcal{N} = \mathcal{H}_N$ is stable if and only if $F < \theta_{N/\mathcal{H}_N}$;

(ii) the high coalition $\mathcal{H}_M$ is stable if and only if $\theta_{M+1/\mathcal{H}_{M+1}} \leq F < \theta_{M/\mathcal{H}_M}$;

(iii) the singleton coalition $\mathcal{H}_1$ is stable if and only if $F \geq \theta_{2/\mathcal{H}_2}$.

For lower elasticity α (*i.e.* high values of F), $\mathcal{H}_1$ is the only stable high coalition, reasserting that the Baliga and Maskin competitive equilibrium is a stable coalition. However, for high elasticity α (*i.e.* small values of F), the grand coalition $\mathcal{N} = \mathcal{H}_N$ is stable, reasserting a Coasian-type bargaining argument.

5.2 Challenging coalitions

We now study the existence and properties of the stable coalitions differing from the stable high coalition $\mathcal{H}_M$. In Theorem 4, we prove that stable coalitions either have either M or $M + 1$ agents. When the stable coalition $\mathcal{A}$ has M agents, we have that $\Theta_{\mathcal{A}} \leq \Theta_{\mathcal{H}_M}$ and so the stable coalition $\mathcal{A}$ has a lower utility and welfare than the stable high coalition $\mathcal{H}_M$. However, when the stable coalition $\mathcal{A}$ has $M + 1$ agents, we prove that $\Theta_{\mathcal{A}} > \Theta_{\mathcal{H}_M}$ and so the stable coalition $\mathcal{A}$ has a greater utility and welfare than the stable high coalition $\mathcal{H}_M$. For this reason, when the stable coalition $\mathcal{A}$ has $M + 1$ agents, we call it a stable *challenging* coalition. In Examples 1 and 2 we show that stable challenging coalition may indeed exist in the model and we determine the values of the elasticity α for which they exist in Corollary 3.

The next result shows the possibilities of a stable coalition when $\mathcal{H}_M$ is the stable high coalition.

Theorem 4. *If $\mathcal{A}$ is a stable coalition, then $M \leq \#\mathcal{A} \leq M+1$, where $\mathcal{H}_M$ is the stable high coalition, and*

$$1 - F < \frac{\Theta_{\mathcal{A}}}{\Theta_{\mathcal{H}_M}} < \frac{1}{1-F} \ .$$

Furthermore, $\Theta_{\mathcal{H}_M} \geq \Theta_{\mathcal{A}}$ if $\#\mathcal{A} = M$; and $\Theta_{\mathcal{H}_M} < \Theta_{\mathcal{A}}$ if $\#\mathcal{A} = M+1$.

Since stable challenging coalitions have a higher preference than the stable high coalition, we have that stable challenging coalitions are always Nash–Cournot equilibria by Lemma 2. They are also Stackleberg equilibria if $M = 2$ and $\alpha > 1/2$, or when $M \geq 3$.

Using the bounds from Theorem 4, the relative utility of stable coalitions $\mathcal{A}$ and the stable high coalition $\mathcal{H}_M$ satisfies the following inequalities:

$$(1-\alpha)^{\frac{1}{\alpha}} = (1-F)^{\frac{1}{1-\alpha}} < U(\alpha; \mathcal{A}/\mathcal{H}_M) = \left(\frac{\Theta_{\mathcal{A}}}{\Theta_{\mathcal{H}_M}}\right)^{\frac{1}{1-\alpha}} < (1-F)^{-\frac{1}{1-\alpha}} = (1-\alpha)^{-\frac{1}{\alpha}}.$$

We observe that $\lim_{\alpha \to 1}(1-F) = 1$, and so when α tends to 1, the high coalition tends to the grand coalition and the bounds of the relative aggregate preferences $\Theta_{\mathcal{A}}/\Theta_{\mathcal{H}_M}$ tend to 1. However, we have $\lim_{\alpha \to 1}(1-F)^{\frac{1}{1-\alpha}} = 0$, and so the information provided by these bounds becomes irrelevant.

Stable challenging coalitions may not always exist. For instance, if $F \geq 1/2$ (*i.e.* $\alpha \leq 1/2$), we have $\ell(F) = 1$, and so only singleton coalitions may be stable. Hence $\mathcal{H}_1 = \{1\}$ is stable and there are no challenging coalitions. However, there can be more than one stable singleton coalition. For instance, consider $\theta_1 = 1$, $\theta_2 = 0.9$ and $F = 0.6$. An application of Theorem 2 shows that both singleton coalitions $\mathcal{H}_1 = \{1\}$ and $\mathcal{A} = \{2\}$ are stable.

For $1/3 \leq F < 1/2$ stable coalitions have cardinality 1 or 2 since $\ell(F) = 2$. Hence, the only way to have a stable challenging coalition is when the stable high coalition is $\mathcal{H}_1 = \{1\}$. Moreover, another application of Theorem 2 implies that stable challenging coalitions of size 2 have to be disjoint from $\mathcal{H}_1$. Indeed, this is also a consequence of the following more general observation.

Remark 1. *If $\mathcal{H}_M$ is the stable high coalition and $\mathcal{A}$ is a stable challenging coalition, then $\mathcal{A}$ cannot be of the form $\mathcal{A} = \mathcal{H}_M \cup \{i\}$ where $i \geq M+1$. Otherwise, we would get $\theta_{i/\mathcal{A}} \leq \theta_{M+1/\mathcal{H}_{M+1}} \leq F$ by Corollary 2, which is a contradiction by Corollary 1. So challenging coalitions $\mathcal{A}$ do not contain the stable high coalition $\mathcal{H}_M$ and so cannot be an extension of the stable high coalition by including some free-rider. So, at least one agent in the stable high coalition $\mathcal{H}_M$ must be 'changed' by a lower agent $i \geq M+1$ while forming a stable challenging coalition.*

The next example shows that there are stable challenging coalitions for F in the interval $1/3 \leq F < 1/2$.

Example 1. *Let $\mathcal{N}_3$ be a grand coalition with three agents. Both coalitions $\mathcal{H}_1 = \{1\}$ and $\mathcal{A}_2 = \{2, 3\}$ are stable if and only if the following inequalities hold:*

1. $\sqrt{2} - 1 \leq F < 1/2$;
2. *the preferences θ_1 and θ_2 satisfy*
$$\frac{\theta_1(1-F)}{2F} \leq \theta_2 \leq \frac{\theta_1 F}{1-F};$$
3. *if $\theta_2 < \theta_1(1-F)^2/F$, the preference θ_3 satisfies*
$$\frac{\theta_1(1-F)}{F} - \theta_2 \leq \theta_3 \leq \theta_2,$$
otherwise, the preference θ_3 satisfies
$$\frac{\theta_2 F}{1-F} < \theta_3 \leq \theta_2.$$

Hence, for $1/3 \leq F < \sqrt{2} - 1 \approx 0.414$, there is no stable challenging coalition with two agents when the stable high coalition is $\mathcal{H}_1$. Note that increasing the number of elements of the grand coalition does not help since for F in this interval stable coalitions always have size 1 or 2. Moreover, for F in this interval, when the stable high coalition is $\mathcal{H}_2$ there are no challenging coalitions since the stable coalition size is at most 2. So for F in this range we conclude that there are no stable challenging coalitions.

The next example shows that there are stable challenging conditions for $F < 1/3$. Let $\mathcal{N}_{l+1} = \{1, \ldots, l+1\}$ be the grand coalition. We define the *candidate* challenging coalition of the stable high coalition $\mathcal{H}_{l-1}$ as $\mathcal{A}_l = \mathcal{N}_{l+1} \setminus \{l-1\}$. Let us consider the case where the stable high coalition has at least two agents. Hence, for $3 \leq l \leq \ell(F)$, let:

1. $\Theta_{\mathcal{H}_{l-1}} = (1 + (l-2)b)\theta_{l-1} = \theta_{l-1} + (l-2)b\theta_{l-1}$ be the aggregate preference of the high coalition $\mathcal{H}_{l-1}$. Observe that $b = (\Theta_{\mathcal{H}_{l-1}} - \theta_{l-1})/((l-2)\theta_{l-1}) \geq 1$.
2. $\theta_{l+1} = \theta_l = C\theta_{l-1}$, for some $C < 1$. Hence $\Theta_{\mathcal{A}_l} = (2C + (l-2)b)\theta_{l-1}$ is the aggregate preference of the candidate challenging coalition.

Hence, to have $\Theta_{\mathcal{A}_l} > \Theta_{\mathcal{H}_{l-1}}$ it is necessary that $C > 1/2$.

Example 2. *Let $\mathcal{N}_{l+1}$ be the grand coalition, $\mathcal{H}_{l-1}$ be the $(l-1)$-high coalition, and $\mathcal{A}_l$ be the challenging coalition. For $F < 1/3$ and $3 \leq l \leq \ell(F)$, both coalitions $\mathcal{H}_{l-1}$ and $\mathcal{A}_l$ are stable if and only if the following inequalities hold:*

1. *the parameter C satisfies*

$$\underline{C} \equiv \max\left\{1-F, \frac{F(l-2)}{1-2F}\right\} < C < 1;$$

2. *the parameter b satisfies*

$$\underline{b}(C) \leq b < \frac{C(1-2F)}{F(l-2)},$$

where

$$\underline{b}(C) \equiv \begin{cases} \frac{1-F(1+2C)}{F(l-2)}, & \text{if } C \leq \frac{1}{1+F} \\ \\ \frac{C(1-F)-F}{F(l-2)}, & \text{if } \frac{1}{1+F} \leq C < 1 \end{cases}.$$

We observe that $1-F<1$ and that $F(l-2)/(1-2F)<1$ for every value of α, and so $\underline{C}<1$. Also, $1-F>2/3$, and so $C>1/2$ is always satisfied. Moreover, $1/(1+F)<1$ and $1-F<1/(1+F)$.

Putting together the previous observations we get the following.

Corollary 3. *For $F \in [1/3, \sqrt{2}-1) \cup [1/2, 1-1/e)$, all stable coalitions have the same cardinality as the stable high coalition. For $F \notin [1/3, \sqrt{2}-1) \cup [1/2, 1-1/e)$ there are examples of stable challenging coalitions with one additional agent compared to the stable high coalition. They do not contain the stable high coalition, but they do achieve greater aggregate preference and hence greater utility than the stable high coalition.*

In particular, letting $\alpha_L = F^{-1}(\theta_{2/\mathcal{H}_2})$ and $\alpha_R = F^{-1}(\theta_{N/\mathcal{H}_N})$, we obtain $1/2 \leq \alpha_L \leq \alpha_R < 1$. Hence: (i) the single coalition $\mathcal{H}_1$ is a stable coalition for $\alpha \leq \alpha_L$, and moreover there are no challenging stable coalitions for $\alpha \leq 1/2$; and (ii) the grand coalition is the unique stable coalition for $\alpha_R < \alpha < 1$, where the Coasian bargaining argument holds.

An interesting point is how the utility of stable challenging coalitions compares to that of stable high-coalitions. We already know that it is larger, but it is interesting to study how they are relative to each other using the relative utility defined before, particularly when the number of agents is large. This question is addressed in the following remark.

Remark 2. *In appendix A.7, for $b=1$ and $C = F(l-1)/(1-F)$, we prove that there is a sequence γ_l tending to 1 when l tends to $+\infty$, such that*

$$\lim_{l\to\infty} U(\gamma_l; \mathcal{A}_l/\mathcal{H}_{l-1}) = \lim_{l\to\infty} \left(\frac{1}{1-F(\gamma_l)}\right)^{\frac{1}{1-\gamma_l}} = +\infty\ .$$

Therefore, in some cases, challenging coalitions can have a much higher utility than the stable high coalition when the number of agents is large.

6 Coalition folk theorem and Barrett's paradox of cooperation

In this section, we will discuss two well-known relevant topics in the stable coalition literature, the so-called coalition folk theorem and Barrett's paradox of cooperation.

6.1 Coalition folk theorem

We now discuss the so-called "coalition folk theorem". The coalition folk theorem (see for instance [15]) is a statement mentioned in the related literature asserting that when there is no transfer mechanism, stable coalitions will tend to be smaller with heterogeneous players than with homogeneous agents. This conclusion holds in our model since in the homogeneous case stable coalitions will have the maximum possible size $\max\{\ell(F), N\}$, which can only decrease (recall Theorem 2 and Corollary 1) when the degree of heterogeneity between agents increases.

Furthermore, we will show that some degree of homogeneity in the preferences of the agents is sufficient for the stability of the coalitions with the highest possible cardinality $\ell(F)$, according to Theorem 2. Let L and R be such that

$$F < L/(R\ell(F)) \leq L \leq 1/\ell(F) \leq R \leq R/(L\ell(F)) \leq F/(1-F)\ . \tag{7}$$

In particular, $L = R = 1/\ell(F)$ is a solution to inequality (7), which still holds if L and R are changed slightly. The following remark results from Theorem 2.

Remark 3. *Let $\beta > 0$ and L and R satisfy inequalities (7), for some $F = F(\alpha)$. Let $\mathcal{N}$ be a grand coalition, with cardinality $N \geq \ell(F)$, such that $L \leq \beta\theta_i \leq R$, for every $1 \leq i \leq N$. A coalition $\mathcal{A}$ in $\mathcal{N}$ is stable if and only if $\mathcal{A}$ has cardinality $\ell(F)$.*

6.2 Barrett's paradox of cooperation with heterogeneous agents

The usual approach to Barrett's paradox of cooperation is under the assumption of homogeneous agents. However, some works consider heterogeneous agents, such as for example [15], where diversity between agents is not an obstacle to cooperation, both at the level of coalition size (coalition folk theorem, as we have explored in the previous subsection) and at the level of the gains obtained from cooperation.

Let us consider a stable coalition $\mathcal{A}$ with $k+1$ free-riders and let $\mathcal{F}$ be the set of all free-riders. Let $\Theta_{\mathcal{F}}$ be the aggregate preference of the $\#\mathcal{F} = k+1$ free-riders and $\Theta_{\mathcal{F}/\mathcal{A}} = \Theta_{\mathcal{F}}/\Theta_{\mathcal{A}}$ the $\mathcal{A}$-conditional preference of the set of free-riders $\mathcal{F}$. Hence, from Theorem 2 the coalition $\mathcal{A}$ is externally stable if and only if for every $j \in \mathcal{F}$, $\theta_{j/\mathcal{A}} \leq F/(1-F)$, and so we have

$$1 + \Theta_{\mathcal{F}/\mathcal{A}} \leq (1+kF)/(1-F).$$

Therefore, we have that the relative utility between coalition $\mathcal{A}$ and the grand coalition

$$U(\alpha;\mathcal{A}/\mathcal{N}) = \left(\frac{1}{1+\Theta_{\mathcal{F}/\mathcal{A}}}\right)^{\frac{1}{1-\alpha}} \geq \left(\frac{1-F}{1+kF}\right)^{\frac{1}{1-\alpha}} .$$

Hence, the *minimum* U_m that the relative utility can achieve is

$$U_m(\alpha;k) \equiv \left(\frac{1-F}{1+kF}\right)^{\frac{1}{1-\alpha}} > 0.$$

We observe that this minimum is attained when the $k+1$ free-riders have the maximal admissible $\mathcal{A}$-conditional preference $\theta_{j/\mathcal{A}} = F/(1-F)$ for every $j \in \mathcal{F}$. Hence, $1+\Theta_{\mathcal{F}/\mathcal{A}} = (1+kF)/(1-F)$ and the minimum relative utility is attained. We observe that $U_m(\alpha;k)$ is decreasing in both variables. Hence, we have the following bound

$$U_m(\alpha;k) \leq U_m(\alpha;0) = (1-F)^{\frac{1}{1-\alpha}} = (1-\alpha)^{\frac{1}{\alpha}}.$$

Remark 4 (Barrett's paradox of cooperation)**.** *We have that the minimum relative utility $U_m(\alpha;k)$ tends to 0 when α tends to 1 or when k tends to $+\infty$ (see Figure 4). In particular, even when the number of free-riders $k+1$ is small, a stable coalition may achieve very low utility when compared to the grand coalition. This goes in the direction of the paradox of cooperation.*

Using (6), the relative welfare between a coalition $\mathcal{A}$ and the grand coalition is given by

$$W(\alpha;\mathcal{A}/\mathcal{N}) = \left(\frac{\Theta_{\mathcal{A}}}{\Theta_{\mathcal{N}}}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{\Theta_{\mathcal{N}} - \alpha\Theta_{\mathcal{A}}}{\Theta_{\mathcal{N}}(1-\alpha)}\right) = \left(\frac{\Theta_{\mathcal{A}}}{\Theta_{\mathcal{N}}}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{1-\alpha\Theta_{\mathcal{A}}/\Theta_{\mathcal{N}}}{(1-\alpha)}\right).$$

Noting that $\Theta_{\mathcal{N}} = \Theta_{\mathcal{A}} + \Theta_{\mathcal{F}}$, we get

$$\frac{1-F}{1+kF} \leq (1+\Theta_{\mathcal{F}/\mathcal{A}})^{-1} = \Theta_{\mathcal{A}}/\Theta_{\mathcal{N}} \leq 1.$$

Hence, the *minimum* W_m of the relative welfare is

$$\begin{aligned}
W_m(\alpha;k) &\equiv \left(\frac{1-F}{1+kF}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{1-\alpha(1-F)/(1+kF)}{(1-\alpha)}\right) \\
&= \left(\frac{1-F}{1+kF}\right)^{\frac{\alpha}{1-\alpha}} \left(\frac{1-\alpha+(k+\alpha)F}{(1+kF)(1-\alpha)}\right) \\
&= (1+kF)^{-\frac{1}{1-\alpha}} (1-F)^{\frac{\alpha}{1-\alpha}} \left(1+\frac{(k+\alpha)F}{1-\alpha}\right) \\
&= \frac{1-\alpha+(k+\alpha)F}{(1+kF)^{\frac{1}{1-\alpha}}}.
\end{aligned}$$

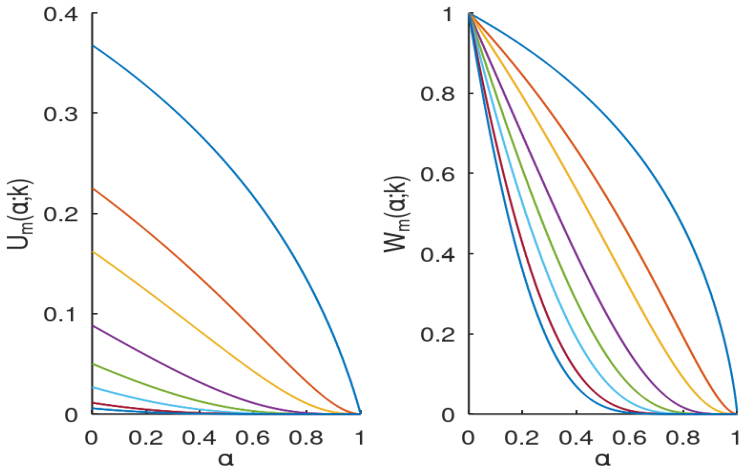

Figure 4: The minimum relative utility $U_m(\alpha;k)$ (left) and the minimum relative welfare $W_m(\alpha;k)$ (right) for different values of $k = 0, 1, 2, 5, 10, 20, 50, 100$. Recall that there are $k+1$ free-riders.

Remark 5 (Barrett's paradox of cooperation)**.** *We have that the minimum the relative welfare $W_m(\alpha;k)$ tends to 0 when α tends to 1 or when k tends to $+\infty$ (see Figure 4).*

Observe that this minimum is attained when all the free-riders $j \in \mathcal{F}$ have the same maximal admissible $\mathcal{A}$-conditional preference $\theta_{j/\mathcal{A}} = F/(1-F)$, and so $1+\Theta_{\mathcal{F}/\mathcal{A}} = (1+kF)/(1-F)$. This case corresponds to the most extreme case in which the paradox of cooperation occurs: when in a stable coalition the free-riders have the highest possible preferences relative to that coalition (such that they still prefer to be free-riders). In this extreme case, when α is high or the number of free-riders is large, the relative welfare of the stable coalition compared to the grand coalition will be very small.

In opposition to the paradox of cooperation, and in the spirit of [15], we present the following example which gives a kind of opposite to the paradox of cooperation. For given α, take: i) $\ell(\alpha)$ agents with equal high preferences $\theta_H = (1-\theta_L)/\ell(\alpha)$; and ii) $N-\ell(\alpha) = 1+J$ agents with equal low preferences θ_L. Hence, $\Theta_{\mathcal{H}_{\ell(\alpha)}} = (1-\theta_L)$ and $\Theta_{\mathcal{N}} = 1 + J\theta_L$.

Example 3 (Anti-paradox of cooperation)**.** *For given α and $0 < \eta < 1$, take the low preference of the free-riders $\theta_L = \theta_L(\alpha)$ with*

$$\theta_L \leq \min\left\{F(\alpha), \frac{1-\eta^{1-\alpha}}{1+J\eta^{1-\alpha}}\right\}.$$

Then, there is a stable high coalition $\mathcal{H}_{\ell(\alpha)}$ achieving relative utility at least η when compared to the grand coalition, i.e., such that

$$U(\alpha;\mathcal{H}_{\ell(\alpha)}/\mathcal{N}) \geq \eta\ ,$$

and achieving relative welfare when compared to the grand coalition such that

$$W(\alpha; \mathcal{H}_{\ell(\alpha)}/\mathcal{N}) \geq \eta - \eta \log(\eta) \ .$$

We have that $\eta - \eta \log(\eta) \approx 1$ *when* $\eta \approx 1$. *The relative utility and welfare between the singleton competitive coalition formed by a top agent and the grand coalition are such that*

$$U(\alpha; \mathcal{H}_1/\mathcal{N}) < \ell(\alpha)^{-\frac{1}{1-\alpha}} \quad \textit{and} \quad W(\alpha; \mathcal{H}_1/\mathcal{N}) < \frac{\ell(\alpha)^{-\frac{\alpha}{1-\alpha}}}{1-\alpha}.$$

We observe that

$$\lim_{\alpha \to 1} \ell(\alpha)^{-\frac{1}{1-\alpha}} = 0 \quad \textit{and} \quad \lim_{\alpha \to 1} \frac{\ell(\alpha)^{-\frac{\alpha}{1-\alpha}}}{1-\alpha} = 0.$$

The conclusion is that when there is an *elite* formed by M agents ($1 < M << N$) that captures a great part of the aggregate preferences of the grand coalition, *i.e.* $\Theta_{\mathcal{H}_M} \approx \Theta_{\mathcal{N}}$ (or in other words when the preference of free-riders is small) then they can form a stable high coalition $\mathcal{H}_M$ that attains a utility and welfare close to those of the grand coalition. Moreover, there is a large gap between the singleton competitive coalition formed by a top agent and the full cooperation scenario of the grand coalition, which means that the effects of eventual cooperation are meaningful. So this case opposes Barrett's paradox of cooperation in the presence of large heterogeneity between agents since the competitive equilibrium coalition $\mathcal{H}_1$ might achieve little, but the stable coalition might achieve quite a lot when compared to the grand coalition.

7 Conclusions and final remarks

In this paper, we have developed a model to study the formation of coalitions in a problem of consumption or provision of a public good or common pool. We have shown the emergence of stable coalitions composed of individuals willing to contribute to the maintenance of the good through a dynamical bargaining Markov chain based on simple rules regarding their incentives and preferences. The dynamics of the Markov chain converges to absorbing states which are stable coalitions. Hence, the bargaining Markov chain gives a path for the emergence of (internally and externally) stable coalitions. The Markov chain is so a form of forming stable coalitions in a decentralized way through bargaining.

We proved the existence of a unique M such that the M-high coalition formed by the M agents that most prefer or value the good is stable. We have shown that any other stable coalitions have either M or $M+1$ members. Depending on the elasticity α: the grand coalition is stable, and so a Coasian type of argument holds; or M-high coalitions are stable, and so Olson's type of argument holds; or all agents are free-riders, except the one that most prefer the good, and so Baliga and Maskin's arguments hold. Hence, our

results give a new insight into the literature, unifying these different points of view under the same model as the elasticity varies.

We have studied the effects of the model parameters on stability by means of a free-rider threshold and the effects of homogeneity and heterogeneity among the agents. We have proven that a high degree of homogeneity among agents is enough to guarantee that stable coalitions have the highest possible size and hence, asymmetry will tend to lead to smaller stable coalitions reasserting the so-called "coalition folk theorem" in our framework.

We have observed some features along the lines of Barrett's paradox of cooperation. We exhibit the minima that the relative welfare of a stable coalition can attain. These minima only depend on the number of free-riders and on the elasticity parameter. However, when the are not too many free-riders and they have a very small preference for the public good, then the paradox of cooperation might not hold, as stable coalitions might achieve welfare comparable to the welfare of the grand coalition. For instance, in the presence of dichotomous heterogeneity of agents, if there is an "elite" formed by a subgroup with a high preference for the public good, and all the other agents have a low preference, then this elite might be enough to form a stable coalition attaining a welfare level close to the welfare of the full-cooperation scenario of the grand coalition.

Agent heterogeneity opens the possibility of studying mechanisms, such as transfers, rewards, or cost allocation which takes into account agent heterogeneity with the goal of stabilizing coalitions. Other possibility of future work can be the study of environmental problems and thresholds under uncertainty such as in [5].

A Proofs of the results

A.1 Proof of Lemmas 1, 2 and 3

For the coalition player $\mathcal{A}$, observe that $\partial u_{\mathcal{A}}(r_{\mathcal{A}}; r_{-\mathcal{A}})/\partial r_{\mathcal{A}} = 0$ if and only if $r_{\mathcal{A}} + r_{-\mathcal{A}} = \overline{r}_{\mathcal{A}} = (\alpha\Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}}$, and that $\partial u_{\mathcal{A}}(r_{\mathcal{A}}; r_{-\mathcal{A}})/\partial r_{\mathcal{A}} < 0$ if and only if $r_{\mathcal{A}} + r_{-\mathcal{A}} > \overline{r}_{\mathcal{A}}$.

For the free-riders $j \in \mathcal{N} \setminus \mathcal{A}$, observe that $\partial u_j(r_j; r_{-j})/\partial r_j = 0$ if and only if $r_j + r_{-j} = \overline{r}_j = (\alpha\theta_j)^{\frac{1}{1-\alpha}}$, and that $\partial u_j(r_j; r_{-j})/\partial r_j = 0$ if and only if $r_j + r_{-j} = \overline{r}_j = (\alpha\theta_j)^{\frac{1}{1-\alpha}}$.

Hence, the best response functions are as claimed in Lemma 1. □

Let us now find the Nash–Cournot equilibria. Recall that $\mathcal{T}$ is the coalition of all (top) agents i that most prefer the good, *i.e.*, agents $i \in \mathcal{T}$ if and only if $\theta_i = \theta_1$. We separate the analysis into three cases.

Case $\overline{r}_{\mathcal{A}} < \overline{r}_1$. This means that $\Theta_{\mathcal{A}} < \theta_1$ and hence $\mathcal{T} \cap \mathcal{A} = \emptyset$. The best response of all top players $j \in \mathcal{T}$ is optimal if and only if

$$\sum_{j \in \mathcal{T}} r_j = \overline{r}_1,$$

and the best response of all other players $j \in \mathcal{N} \setminus \mathcal{T}$ is optimal if and only if

$$\sum_{j \in \mathcal{N} \setminus \mathcal{T}} r_j = 0.$$

Furthermore, any other contributions are not optimal.

Case $\overline{r}_{\mathcal{A}} = \overline{r}_1$. This means that $\Theta_{\mathcal{A}} = \theta_1$. The best response of all top players $j \in \mathcal{T}$ and player $\mathcal{A}$ is optimal if and only if

$$\sum_{j \in \mathcal{T} \cup \mathcal{A}} r_j = \overline{r}_1,$$

and the best response of all other players $j \in \mathcal{N} \setminus (\mathcal{T} \cup \mathcal{A})$ is optimal if and only if

$$\sum_{j \in \mathcal{N} \setminus (\mathcal{T} \cup \mathcal{A})} r_j = 0.$$

Furthermore, any other contributions are not optimal.

Case $\overline{r}_{\mathcal{A}} > \overline{r}_1$. This means that $\Theta_{\mathcal{A}} > \theta_1$. The best response of player $\mathcal{A}$ is optimal if and only if

$$\sum_{j \in \mathcal{A}} r_j = \overline{r}_{\mathcal{A}},$$

and the best response of all other players $j \in \mathcal{N} \setminus \mathcal{A}$ is optimal if and only if

$$\sum_{j \in \mathcal{N} \setminus \mathcal{A}} r_j = 0.$$

Furthermore, any other contributions are not optimal. □

Let us now find the Stackelberg equilibria. Take the smallest $k \in \mathcal{N} \setminus \mathcal{A}$ and recall that $\theta(\mathcal{A}) = \theta_k$ and that $\mathcal{T}(\mathcal{A})$ is the set of agents in $j \in \mathcal{N} \setminus \mathcal{A}$ such that $\theta_j = \theta_k$, *i.e.*, the set of top free-riders.

Case: Leader chooses $r_{\mathcal{A}} \geq \overline{r}_k$. In this case, the best response r_j of the followers $j \in \mathcal{N} \setminus \mathcal{A}$ is 0 and so

$$\sum_{j \in \mathcal{N} \setminus \mathcal{A}} r_j = 0.$$

Hence, the best choice for the leader is

$$r_{\mathcal{A}} = \begin{cases} \overline{r}_k & \text{if } \Theta_{\mathcal{A}} \leq \theta_k \\ \overline{r}_{\mathcal{A}} & \text{if } \Theta_{\mathcal{A}} \geq \theta_k \end{cases}.$$

Furthermore,

$$u_{\mathcal{A}}(\overline{r}_k; 0) = \Theta_{\mathcal{A}} \left(\alpha \theta_k\right)^{\frac{\alpha}{1-\alpha}} - \left(\alpha \theta_k\right)^{\frac{1}{1-\alpha}},$$

and

$$u_{\mathcal{A}}(\overline{r}_{\mathcal{A}};0) = (\alpha\Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}}\left(\frac{1}{\alpha}-1\right).$$

Case: Leader chooses $r_{\mathcal{A}} \leq \overline{r}_k$. In this case, the best response r_j of the followers $j \in \mathcal{T}(\mathcal{A})$ is such that

$$r_{\mathcal{A}} + \sum_{j\in\mathcal{T}(\mathcal{A})} r_j = \overline{r}_k;$$

and the best response r_j of the followers $j \in \mathcal{N}\setminus(\mathcal{A}\cup\mathcal{T}(\mathcal{A}))$ is 0 and so

$$\sum_{j\in\mathcal{N}\setminus(\mathcal{A}\cup\mathcal{T}(\mathcal{A}))} r_j = 0.$$

Hence, the best choice for the leader is $r_{\mathcal{A}} = 0$. Furthermore,

$$u_{\mathcal{A}}(0;\overline{r}_k) = \Theta_{\mathcal{A}}\left(\alpha\theta_k\right)^{\frac{\alpha}{1-\alpha}}.$$

Now, let us compare the utility of the leader for both of its choices.
Case $\Theta_{\mathcal{A}} \geq \theta_k$. We have that

$$u_{\mathcal{A}}(0;\overline{r}_k) = \Theta_{\mathcal{A}}\left(\alpha\theta_k\right)^{\frac{\alpha}{1-\alpha}} \leq (\alpha\Theta_{\mathcal{A}})^{\frac{1}{1-\alpha}}\left(\frac{1}{\alpha}-1\right) = u_{\mathcal{A}}(\overline{r}_{\mathcal{A}};0)$$

if and only if

$$\theta_k \leq \Theta_{\mathcal{A}}\left(1-\alpha\right)^{\frac{1-\alpha}{\alpha}}.$$

which gives the two cases in Lemma 3.
Case $\Theta_{\mathcal{A}} \leq \theta_k$. Hence, $\theta(\mathcal{A}) = \theta_k = \theta_1$, and the following inequality always holds for every α:

$$u_{\mathcal{A}}(0;\overline{r}_1) = \Theta_{\mathcal{A}}\left(\alpha\theta_1\right)^{\frac{\alpha}{1-\alpha}} > \Theta_{\mathcal{A}}\left(\alpha\theta_1\right)^{\frac{\alpha}{1-\alpha}} - \left(\alpha\theta_1\right)^{\frac{1}{1-\alpha}} = u_{\mathcal{A}}(\overline{r}_1;0),$$

and so the Stackelberg equilibrium is a low-cooperation of $\mathcal{T}(\mathcal{A})$. Since $\Theta_{\mathcal{A}}(1-\alpha)^{\frac{1-\alpha}{\alpha}} < \Theta_{\mathcal{A}} \leq \theta_k$, this case falls in case 1 of Lemma 3, and the lemma is proved. □

A.2 Proof of Theorem 2

Theorem 2 follows from Lemmas 4 and 5.

Lemma 4. *The coalition* $\mathcal{A}$ *is stable if and only if: (i)* $\mathcal{A}$ *is internally stable:* $\theta_{j/\mathcal{A}} > F$ *for every* $j \in \mathcal{A}$*; and (ii)* $\mathcal{A}$ *is externally stable:* $\theta_{j/\mathcal{A}} \leq F/(1-F)$ *for every* $j \notin \mathcal{A}$.

Proof. First observe that

$$\theta_j \overline{r}_{\mathcal{A}}^{\alpha} = \theta_j (\alpha \Theta_{\mathcal{A}})^{\frac{\alpha}{1-\alpha}}$$

and that

$$r_{j/\mathcal{A}} = \alpha \theta_j \overline{r}_{\mathcal{A}}^{\alpha} \ .$$

Let us first consider the external stability of $\mathcal{A}$. Take $j \notin \mathcal{A}$. The external stability condition is

$$\theta_j \overline{r}_{\mathcal{A}}^{\alpha} \geq \theta_j \overline{r}_{\mathcal{A} \cup \{j\}}^{\alpha} - r_{j/\mathcal{A} \cup \{j\}}.$$

which is equivalent to

$$\theta_j \leq \left((1-\alpha)^{-\frac{1-\alpha}{\alpha}} - 1 \right) \Theta_{\mathcal{A}} = \frac{F}{1-F} \Theta_{\mathcal{A}}.$$

Now let us consider the internal stability of $\mathcal{A}$. Take $j \in \mathcal{A}$. The internal stability condition is

$$\theta_j \overline{r}_{\mathcal{A} \setminus \{j\}}^{\alpha} \leq \theta_j \overline{r}_{\mathcal{A}}^{\alpha} - r_{j/\mathcal{A}}$$

which is equivalent to

$$\theta_j \geq \frac{\left((1-\alpha)^{-\frac{1-\alpha}{\alpha}} - 1 \right)}{\left((1-\alpha)^{-\frac{1-\alpha}{\alpha}} \right)} \Theta_{\mathcal{A}} = F \Theta_{\mathcal{A}}.$$

□

Lemma 5. *If $\mathcal{A}$ is a stable coalition, then $\#\mathcal{A} \leq \ell(F)$.*

Proof. For the free-rider threshold F, let $\mathcal{A}$ be a stable coalition with the highest possible cardinality $L = \#\mathcal{A}$. Let $i_L \in \mathcal{A}$ be the agent in $\mathcal{A}$ with the smallest preference for the public good. Suppose that $L > \ell(F)$ and so

$$\theta_{i_L/\mathcal{A}} = \frac{\theta_{i_L}}{\Theta_{\mathcal{A}}} \leq \frac{1}{\ell(F)+1} \leq F \ .$$

By Lemma 4, agent i_L prefers to be a free-rider and so $\mathcal{A}$ is not internally stable, which is absurd. Hence, $L \leq \ell(F)$. □

A.3 Proof of Theorem 3

Theorem 3 follows from the following two lemmas.

Lemma 6. *There is a unique M such that the high coalition $\mathcal{H}_M$ is stable.*

Proof. We consider the following algorithm: (i) start with $M = N$; (ii) if the high coalition $\mathcal{H}_M$ is not internally stable, then decrease M by one and repeat this step; (iii) otherwise, *i.e.* if the stable high coalition $\mathcal{H}_M$ is internally stable, stop.

We observe that agent $j = 1$ will prefer to be in the singleton high coalition $\mathcal{H}_1 = \{1\}$ than the case where all agents free ride since $u(1/\{1\}) = \theta_1 \overline{r}^{\alpha}_{\{1\}} - r_{1/\{1\}}$. Hence the algorithm stops for some $M \geq 1$.

Suppose that by applying the above algorithm we obtain the high-coalition $\mathcal{H}_M$. It follows that $\mathcal{H}_{M+k}$ is not internally stable for every $k \geq 1$, and $\mathcal{H}_M$ is internally stable. Hence, Theorem 2 implies that agent $M+1$ prefers to be a free-rider for the high coalition $\mathcal{H}_M$. Therefore, again by Theorem 2, all agents $j \geq M+1$ also prefer to be free-riders for the high coalition $\mathcal{H}_M$ because $\theta_j \leq \theta_{M+1}$. Hence, $\mathcal{H}_M$ is also externally stable and hence a stable coalition.

It remains to prove uniqueness. By Theorem 2, since the high coalition $\mathcal{H}_M$ is internally stable then agent M prefers not to be a free-rider for the high coalition $\mathcal{H}_{M-1}$. Hence, $\mathcal{H}_{M-1}$ is not externally stable. So by Theorem 2, agent M also prefers not to be a free-rider for the high coalition $\mathcal{H}_{M-k}$. So $\mathcal{H}_{M-k}$ is also not externally stable for all $1 \leq k < M$. Therefore, $\mathcal{H}_M$ is the only high coalition that is stable. □

Lemma 7. *If $\mathcal{H}_M$ is stable and $\theta_{M+1} \leq (1-F)\theta_M$ then $\mathcal{H}_M$ is the only stable coalition.*

Proof. By contradiction, suppose that there is $\mathcal{A} \neq \mathcal{H}_M$ that is stable. Since $\mathcal{H}_{M-1}$ is not externally stable, as a consequence of Theorem 2 any coalition with less than M agents is also not externally stable. Hence, the cardinality of $\mathcal{A}$ is at least $\#\mathcal{H}_M = M$. Hence, there is an agent $k \geq M+1$ such that $k \in \mathcal{A}$. By hypothesis and Theorem 2, we have that

$$(1-F)\theta_{M/\mathcal{A}} \geq \theta_{M+1/\mathcal{A}} \geq \theta_{k/\mathcal{A}} > F.$$

Hence, $\theta_{M/\mathcal{A}} > F/(1-F)$. Again, by Theorem 2, agent $M \in \mathcal{A}$ and so $l \in \mathcal{A}$ for all agents $l \leq M$. Thus $\mathcal{A} \supset \mathcal{H}_M$. Since $\mathcal{H}_{M+1}$ is not internally stable, then $\mathcal{A}$ is also not internally stable which is absurd. □

A.4 Proof of Theorem 4

Theorem 4 follows from Lemmas 8 and 9.

Lemma 8. *If $\mathcal{H}_M$ is the stable high coalition and $\mathcal{A}$ is a stable coalition then $M \leq \#\mathcal{A}$. Furthermore,*

$$1 - F < \lambda = \frac{\Theta_{\mathcal{H}_M}}{\Theta_{\mathcal{A}}} < \frac{1}{1-F}.$$

Proof. The results are clear when $\mathcal{A} = \mathcal{H}_M$. Let $\mathcal{A}$ be a stable coalition such that $\mathcal{A} \neq \mathcal{H}_M$. We have that $\#\mathcal{A} \geq \#\mathcal{H}_M = M$, since $\mathcal{H}_{M-1}$ is not externally stable and so by Theorem

2 we obtain that any other coalition with less than M agents is also not externally stable. For every agent $i \in \mathcal{N}$,

$$\frac{\theta_{i/\mathcal{A}}}{\theta_{i/\mathcal{H}_M}} = \frac{\Theta_{\mathcal{H}_M}}{\Theta_{\mathcal{A}}} = \lambda.$$

Since $\#\mathcal{A} \geq M$ and $\mathcal{A} \neq \mathcal{H}_M$ we obtain that $\mathcal{A} \nsubseteq \mathcal{H}_M$. Hence, there is an agent $i \in \mathcal{A}$ and $i \notin \mathcal{H}_M$. Since $i \in \mathcal{A}$ prefers not to be a free-rider, by Theorem 2, $F < \theta_{i/\mathcal{A}}$. Since $i \notin \mathcal{H}_M$ does not belong to the stable coalition $\mathcal{H}_M$, by Theorem 2, $\theta_{i/\mathcal{H}_M} \leq F/(1-F)$. Thus

$$\lambda = \frac{\theta_{i/\mathcal{A}}}{\theta_{i/\mathcal{H}_M}} > 1 - F \ .$$

Since $\mathcal{H}_{M+1}$ is not internally stable, we obtain that $\mathcal{H}_M \nsubseteq \mathcal{A}$. Hence, there is an agent $j \in \mathcal{H}_M$ such that $j \notin \mathcal{A}$. Since $j \in \mathcal{H}_M$ prefers not to be a free-rider, by Theorem 2, $F < \theta_{i/\mathcal{H}_M}$ Since $j \notin \mathcal{A}$ does not belong to the stable coalition $\mathcal{A}$, by Theorem 2, $\theta_{i/\mathcal{A}} \leq F/(1-F)$. Thus,

$$\lambda = \frac{\theta_{j/\mathcal{A}}}{\theta_{j/\mathcal{H}_M}} < \frac{1}{1-F} \ .$$

Therefore, $1 - F < \lambda < 1/(1-F)$. □

Lemma 9. *If $\mathcal{H}_M$ is the stable high coalition and $\mathcal{A} \neq \mathcal{H}_M$ is a stable coalition, then $\#\mathcal{A} \leq M+1$. Furthermore,*

$$\lambda = \frac{\Theta_{\mathcal{H}_M}}{\Theta_{\mathcal{A}}} < 1 + \frac{\#\left(\mathcal{H}_M \setminus \mathcal{A}\right) F^2}{1-F} - (\#\mathcal{A} - M)F \ ,$$

and if $\#\mathcal{A} = M+1$

$$\Theta_{\mathcal{H}_M} < \Theta_{\mathcal{A}} \ .$$

If $\#\mathcal{A} = M$, then

$$\Theta_{\mathcal{H}_M} > \Theta_{\mathcal{A}} \ .$$

Proof. Since $\#\mathcal{A} \geq M$ from the previous lemma we have that there is an injective function j that associates to each agent $k \in \mathcal{H}_M \setminus \mathcal{A}$ an agent $j(k) \in \mathcal{A} \setminus \mathcal{H}_M$ with the following property $\theta_{k/\mathcal{A}} - \theta_{j(k)/\mathcal{A}} > 0$. Let $\mathcal{B}$ be the set of agents $m \in \mathcal{A}$ such that: (i) $m \notin \mathcal{A} \cap \mathcal{H}_M$; and (ii) $m \neq j(k)$, for every $k \in \mathcal{H}_M \setminus \mathcal{A}$. We observe that $\#\mathcal{B} = \#\mathcal{A} - M \geq 0$. Therefore,

$$\sum_{i \in \mathcal{A}} \theta_{i/\mathcal{A}} = \sum_{i \in \mathcal{A} \cap \mathcal{H}_M} \theta_{i/\mathcal{A}} + \sum_{k \in \mathcal{H}_M \setminus \mathcal{A}} \theta_{j(k)/\mathcal{A}} + \sum_{m \in \mathcal{B}} \theta_{m/\mathcal{A}} = 1 \ .$$

For each agent $k \in \mathcal{H}_M \setminus \mathcal{A}$, by Theorem 2, $\theta_{k/\mathcal{A}}(\mathcal{A}) \leq F/(1-F)$ and $\theta_{j(k)/\mathcal{A}}(\mathcal{A}) > F$. Hence,

$$\theta_{k/\mathcal{A}} - \theta_{j(k)/\mathcal{A}} < F/(1-F) - F = \frac{F^2}{1-F} \ ,$$

and so

$$-\sum_{k\in\mathcal{H}_M\setminus\mathcal{A}}\theta_{j(k)/\mathcal{A}} < -\sum_{k\in\mathcal{H}_M\setminus\mathcal{A}}\theta_{k/\mathcal{A}} + \frac{\#\left(\mathcal{H}_M\setminus\mathcal{A}\right)F^2}{1-F} \ .$$

Furthermore, $\sum_{i\in\mathcal{A}\cap\mathcal{H}_M}\theta_{i/\mathcal{A}} + \sum_{k\in\mathcal{H}_M\setminus\mathcal{A}}\theta_{k/\mathcal{A}} = \Theta_{\mathcal{H}_M}/\Theta_{\mathcal{A}} = \lambda$. Hence,

$$\sum_{m\in\mathcal{B}}\theta_{m/\mathcal{A}} = 1 - \sum_{i\in A\cap\mathcal{H}_M}\theta_{i/\mathcal{A}} - \sum_{k\in\mathcal{H}_M\setminus\mathcal{A}}\theta_{j(k)/\mathcal{A}} < 1-\lambda+\frac{\#\left(\mathcal{H}_M\setminus\mathcal{A}\right)F^2}{1-F} \ .$$

Thus, $\lambda < 1 + \frac{\#(\mathcal{H}_M\setminus\mathcal{A})F^2}{1-F} - \sum_{m\in\mathcal{B}}\theta_{m/\mathcal{A}}$. Since by Theorem 2, for every $m\in\mathcal{B}$, $\theta_{m/\mathcal{A}} > F$, then

$$\lambda < 1 + \frac{\#\left(\mathcal{H}_M\setminus\mathcal{A}\right)F^2}{1-F} - (\#\mathcal{B})F \ ,$$

and the main inequality of the lemma is proved. Since by Lemma 5, $\#\left(\mathcal{H}_M\setminus\mathcal{A}\right) \le \#\mathcal{H}_M \le \ell(F) < 1/F$, we also have $\lambda < 1 + F/(1-F) - (\#\mathcal{B})F$.

Now let us suppose that $\#\mathcal{B} = 1$, *i.e.* $\#\mathcal{A} = M+1$. Again by Lemma 5, $\#\left(\mathcal{H}_M\setminus\mathcal{A}\right) \le \#\mathcal{H}_M \le l(F) - 1 < \frac{1-F}{F}$, and so

$$\lambda = \frac{\Theta_{\mathcal{H}_M}}{\Theta_{\mathcal{A}}} < 1 + \frac{\#\left(\mathcal{H}_M\setminus\mathcal{A}\right)F^2}{1-F} - (\#\mathcal{B})F < 1 \ ,$$

and the second part of the lemma is proved.

Now let us prove that $\#\mathcal{B} \le 1$, which would imply $\#\mathcal{A} \le M+1$. Suppose by contradiction that $\#\mathcal{B} \ge 2$. Again by Lemma 5, $\#\left(\mathcal{H}_M\setminus\mathcal{A}\right) \le \#\mathcal{H}_M \le \ell(F) - 2 < \frac{1-2F}{F}$. So we have

$$\lambda < 1 + \frac{\#\left(\mathcal{H}_M\setminus\mathcal{A}\right)F^2}{1-F} - (\#\mathcal{B})F < 1 + \frac{(1-2F)F}{1-F} - 2F = 1 - \frac{F}{1-F} \ ,$$

which is a contradiction, since by Lemma 8, $\lambda > 1-F$. So $\#\mathcal{A} \le M+1$.

Finally, observe that if $\#\mathcal{A} = M$ and since $\mathcal{H}_M$ is a coalition consisting of the M highest agents, then necessarily $\Theta_{\mathcal{H}_M} > \Theta_{\mathcal{A}}$, and the lemma is proved. □

A.5 Proof of Theorem 1

We must show that there is some absorbing state, and that every state can reach an absorbing state with positive probability. We have from Theorem 3 that there is at least one stable coalition and hence the Markov chain has at least one absorbing state. Let N be the cardinality of the grand coalition. The following path from every coalition $\mathcal{A}(0)$ to the stable high coalition $\mathcal{A}(N) = \mathcal{H}_M$ has positive probability in time N, unless the path follows in another absorbing state before time N: at each time step K, form the the coalition $\mathcal{A}(K+1)$ either by: (i) adding to the coalition $\mathcal{A}(K)$ the agent with the highest

preference who is a free-rider; or (ii) excluding the agent in the coalition $\mathcal{A}(K)$ with the smallest preference. If both these moves have transition probability zero in the Markov chain for some step $K < N$, then the Markov chain transitions from $\mathcal{A}(K)$ to $\mathcal{A}(K)$ with probability 1, meaning that $\mathcal{A}(K)$ is an absorbing state and hence stable. Otherwise, $\mathcal{A}(N)$ is an absorbing state.

For $\alpha \geq 1/2$, if $\mathcal{A}$ is stable, then by Theorem 2 we have $\theta_{1/\mathcal{A}} \leq F/(1-F) \leq 1$ and so stand-alone strategies of $\mathcal{A}$ are Nash–Cournot equilibria by Lemma 2.

We have that $F/(1-F) \leq 1-F$ if and only if $\alpha \geq 0.732\ldots$. Let k be the top free-rider relative to $\mathcal{A}$, *i.e.*, $\theta_k = \theta(\mathcal{A})$. If $\mathcal{A}$ is stable, then, by Theorem 2, $\theta_{k/\mathcal{A}} \leq F/(1-F) \leq 1-F = (1-\alpha)^{\frac{1-\alpha}{\alpha}}$, and so stand-alone strategies of $\mathcal{A}$ are Stackelberg equilibria by Lemma 3. □

A.6 Construction of Example 1

The high coalition $\mathcal{H}_1 = \{1\}$ is stable if and only if agent 2 prefers to be a free-rider for the high coalition $\mathcal{H}_1 = \{1\}$: $\theta_2/\theta_1 \leq F/(1-F)$. Hence,

$$\theta_2 \leq \theta_1 F/(1-F). \tag{8}$$

Coalition $\mathcal{A}_2 = \{2,3\}$ is internally stable if and only if agent 3 prefers not to be a free-rider for coalition $\mathcal{A}_2 = \{2,3\}$: $\theta_3/(\theta_2+\theta_3) > F$. Hence,

$$b_1 = \theta_2 F/(1-F) < \theta_3 \leq \theta_2, \tag{9}$$

and so $F < 1/2$ is a necessary condition. Coalition $\mathcal{A}_2 = \{2,3\}$ is externally stable if and only if agent 1 prefers to be a free-rider for coalition $\mathcal{A}_2 = \{2,3\}$: $\theta_1/(\theta_2+\theta_3) \leq F/(1-F)$.

$$b_2 = \theta_1(1-F)/F - \theta_2 \leq \theta_3 \leq \theta_2, \tag{10}$$

and so $\theta_1(1-F)/(2F) \leq \theta_2$ is a necessary condition.

Hence, coalitions $\mathcal{H}_1$ and $\mathcal{A}_2$ are stable if and only if the triples $(\theta_1, \theta_2, \theta_3)$ satisfy the aforementioned inequalities (8), (9) and (10). Let us show that these are non-empty by showing that there are indeed triples $(\theta_1, \theta_2, \theta_3)$ satisfying them. For every $\theta_1 > 0$, choose $\theta_2 > 0$ satisfying the above inequality (8) and the necessary condition implied by inequality (10):

$$\theta_1(1-F)/(2F) \leq \theta_2 \leq F/(1-F)\theta_1, \tag{11}$$

and this is possible when the necessary condition $(1-F)/(2F) \leq F/(1-F)$ holds, *i.e.*, when $F \geq \sqrt{2}-1$. Note that $b_1 < b_2$ is equivalent to $\theta_2 < \theta_1(1-F)^2/F$. Hence, if $\theta_2 < \theta_1(1-F)^2/F$, choose $\theta_3 > 0$ satisfying inequality (10):

$$b_2 = \theta_1(1-F)/F - \theta_2 \leq \theta_3 \leq \theta_2;$$

otherwise, choose $\theta_3 > 0$ satisfying inequality (9),

$$b_1 = \theta_2 F/(1-F) < \theta_3 \leq \theta_2.$$

Such choices are possible since $(1-F)/(2F) < F/(1-F) < 1$.

A.7 Construction of Example 2 and Remark 2

Let $F < 1/3$ and write $3 \leq l = a/F \leq \ell(F)$. Hence, $3F \leq a \leq \ell(F)F < 1$. Let $C\theta_{l-1} = \theta_l = \theta_{l+1}$, with $1 - F < C < 1$, and write $\Theta_{\mathcal{H}_{l-1}} = (1 + (l-2)b)\theta_{l-1}$, for some $b \geq 1$.

By Theorem 2, the high coalition $\mathcal{H}_{l-1}$ is stable if: (i) agent $l \notin \mathcal{H}_{l-1}$ does not want to become a member of $\mathcal{H}_{l-1}$:

$$\theta_l/(\theta_{l-1}(1 + (l-2)b)) \leq F/(1-F);$$

and (ii) agent $l - 1 \in \mathcal{M}_{l-1}$ does not want to become a free-rider:

$$\theta_{l-1}/(\theta_{l-1}(1 + (l-2)b)) > F.$$

Noting that $\theta_{l-1}C = \theta_l$, and putting together the above two inequalities, we get

$$C(1-F)\theta_{l-1} \leq F(1 + (l-2)b)\theta_{l-1} < \theta_{l-1}.$$

Hence,

$$\underline{b}_1 \equiv (C(1-F) - F)/(a - 2F) \leq b < (1-F)/(a-2F) \equiv \overline{b}_1. \tag{12}$$

By Theorem 2, the coalition $\mathcal{A}_l$ is stable if: (i) agent $l - 1 \notin \mathcal{A}_l$ does not want to become a member of $\mathcal{A}_l$:

$$\theta_{l-1}/(2\theta_l + (l-2)b\theta_{l-1}) \leq F/(1-F);$$

and (ii) agent $l + 1 \in \mathcal{A}_l$ does not want to become a free-rider:

$$\theta_{l+1}/(2\theta_l + (l-2)b\theta_{l-1}) > F.$$

Noting that $C\theta_{l-1} = \theta_l = \theta_{l+1}$, and putting together the above two inequalities, we get

$$(1-F)\theta_{l-1} \leq F(2C + (l-2)b)\theta_{l-1} < C\theta_{l-1}$$

Hence,

$$\underline{b}_2 \equiv (1 - F(1+2C))/(a-2F) \leq b < C(1-2F)/(a-2F) \equiv \overline{b}_2 \tag{13}$$

Since $C < 1 \leq (1-F)/(1-2F)$, by simple algebra we get $\overline{b}_2 \leq \overline{b}_1$ (recall that $1 - 2F > 1/3 > 0$ in the interval of F that we are considering). We also get that $\overline{b}_2 > 1$ is equivalent to $C > F(l-2)/(1-2F) \equiv \tilde{C}$.

Regarding the lower bounds, we have that $C < 1$ gives $\underline{b}_1 < \bar{b}_2$. Furthermore, $1 - F < C$ gives $\underline{b}_2 < \bar{b}_2$. We have that $\underline{b}_2 \leq \underline{b}_1$ is equivalent to $C \geq 1/(1+F)$ in which case we obtain

$$\underline{b}_2 \leq \underline{b}_1 \leq b < \bar{b}_2 \leq \bar{b}_1.$$

Otherwise, if $C \leq 1/(1+F)$ we have $\underline{b}_2 \geq \underline{b}_1$ and so we obtain

$$\underline{b}_1 \leq \underline{b}_2 \leq b < \bar{b}_2 \leq \bar{b}_1.$$

Hence, $\mathcal{H}_{l-1}$ and $\mathcal{A}_l$ are stable if and only if $\underline{b} \equiv \max(\underline{b}_1, \underline{b}_2) \leq b < \bar{b}_2$ and $\max(1-F, \tilde{C}) < C < 1$. □

Let α_l and β_l be implicitly determined by

$$l = \frac{1}{F(\alpha_l)} \quad \text{and} \quad l = \frac{1 + F(\beta_l)^2}{F(\beta_l)(1 + F(\beta_l))}.$$

The proof of the lemma below will imply that $\alpha_l < \beta_l < \alpha_{l+1}$.

Lemma 10. *$\underline{b} \leq 1$ if and only if the following holds:*

$$l = \ell(F), \quad \ell(F) \geq \frac{1+F^2}{F(1+F)} \quad \textit{and} \quad \frac{1 - F(\ell(F)-1)}{2F} \leq C \leq \frac{F(\ell(F)-1)}{1-F} < 1.$$

Furthermore, $\underline{b} \leq 1$ if and only if α satisfies $\alpha_l < \alpha \leq \beta_l < \alpha_{l+1}$.

Proof. According to the definition of $\underline{b}$ there are two cases.
Case 1. $C \leq 1/(1+F)$:

$$\underline{b} = \frac{1 - F(1+2C)}{F(l-2)} \leq 1$$

if and only if

$$\frac{1 - F(l-1)}{2F} \leq C.$$

However, for this inequality to have a solution we must have

$$\frac{1 - F(l-1)}{2F} \leq \frac{1}{1+F},$$

which is equivalent to

$$\frac{1+F^2}{F(1+F)} \leq l.$$

Case 2. $1/(1+F) \leq C < 1$:

$$\underline{b} = \frac{C(1-F) - F}{F(l-2)} \leq 1$$

if and only if

$$C \leq \frac{F(l-1)}{1-F}.$$

However, for this inequality to have a solution we must have

$$\frac{1}{1+F} \leq \frac{F(l-1)}{1-F},$$

which is equivalent to

$$\frac{1+F^2}{F(1+F)} \leq l.$$

We now analyse when this necessary condition is satisfied. Recall that by definition of $\ell(F)$ we have that $1/F - 1 \leq \ell(F) < 1/F$. We have that:

(a) when $l < \ell(F)$, there is no solution since we would have $(1+F^2)/(F(1+F)) \leq l \leq \ell(F) - 1 < (1-F)/F$, which is equivalent to $1 + F^2 < 1 - F^2$, which is a contradiction;

(b) when $l = \ell(F)$ then there is a solution since we have that

$$\frac{1}{F} - 1 < \frac{1+F^2}{F(1+F)} < \frac{1}{F},$$

and this is an increasing function of α. So we can find F (or α) such that $(1+F)/(F(1+F)) \leq l = \ell(F)$. In particular, we have that for $\alpha = \beta_l$ this holds with equality. Observe that for $\alpha \leq \alpha_l$ we have that $l = \ell(F) < 1/F(\alpha) \leq 1/F(\alpha_l) = l$, which is a contradiction. Hence, there are solutions if and only if $\alpha_l < \alpha \leq \beta_l$. We also observe that if $\alpha_{l+1} \leq \beta_l$ we would have $l = 1/F(\alpha_{l+1}) - 1 \leq 1/F(\beta_l) - 1 < l$, which is a contradiction. So we have that $\beta_l < \alpha_{l+1}$. □

Proof. (Remark 2) Recall that

$$\frac{\Theta_{\mathcal{A}_l}}{\Theta_{\mathcal{H}_{l-1}}} = \frac{2C + (l-2)b}{1 + (l-2)b}.$$

Let $l = \ell(F)$ and $C = F(\ell(F) - 1)/(1-F)$. The conditions of the previous lemma are satisfied, and so let $b = 1$. We obtain that

$$\frac{\Theta_{\mathcal{A}_l}}{\Theta_{\mathcal{H}_{l-1}}} = \frac{2F(\ell(F)-1) + (1-F)(\ell(F)-2)}{(1-F)(\ell(F)-1)} = \frac{1}{1-F}\left(1 + \frac{F\ell(F)-1}{\ell(F)-1}\right).$$

So the relative utility between the two coalitions is

$$U(\alpha; \mathcal{A}_l/\mathcal{H}_{l-1}) = \left(\frac{\Theta_{\mathcal{A}_l}}{\Theta_{\mathcal{H}_{l-1}}}\right)^{\frac{1}{1-\alpha}} = \left(\frac{1}{1-F}\right)^{\frac{1}{1-\alpha}} \left(1 + \frac{F\ell(F)-1}{\ell(F)-1}\right)^{\frac{1}{1-\alpha}}.$$

Clearly, since $F\ell(F) < 1$ we have that

$$\left(1 + \frac{F\ell(F) - 1}{\ell(F) - 1}\right)^{\frac{1}{1-\alpha}} < 1.$$

Recall that in the interval $(\alpha_l, \beta_l]$ we have the bound $F\ell(F) \geq (1 + F^2)/(1 + F) > 1 - F$, and that $\ell(F) < 1/F$, which implies $1/(\ell(F) - 1) > F/(1 - F)$. So we have

$$\left(1 + \frac{F\ell(F) - 1}{\ell(F) - 1}\right)^{\frac{1}{1-\alpha}} > \left(1 - \frac{F^2}{1 - F}\right)^{\frac{1}{1-\alpha}}.$$

and it may be seen graphically that the right hand side approaches 1 when α approaches 1. So we may take a sequence $\gamma_l \in (\alpha_l, \beta_l]$ such that

$$\lim_{l\to\infty} \left(1 + \frac{F(\gamma_l)\ell(F(\gamma_l)) - 1}{\ell(F(\gamma_l)) - 1}\right)^{\frac{1}{1-\gamma_l}} = 1.$$

Hence

$$\lim_{l\to\infty} U(\gamma_l; \mathcal{A}_l/\mathcal{H}_{l-1}) = \lim_{l\to\infty} \left(\frac{1}{1 - F(\gamma_l)}\right)^{\frac{1}{1-\gamma_l}} = +\infty.$$

□

A.8 Proof of Remark 3

By construction of the interval $[L, R]$, it follows from Theorem 2 that all coalitions $\mathcal{A}$ with cardinality $l(F)$ are stable. Now, let us suppose that a coalition $\mathcal{B}$ has cardinality $\ell(F) - k$, for some $k \geq 1$. Hence, the aggregate preference satisfies the inequality $\Theta_{\mathcal{B}} \leq (\ell(F) - k)R/\beta$. For every $i \notin \mathcal{B}$, $\beta\theta_i \geq L$, and so

$$\theta_i/\Theta_{\mathcal{B}} \geq L/(\ell(F) - k)R > F\ell(F)/(\ell(F) - k) \geq F/(1 - 1/\ell(F)) > F/(1 - F)\ .$$

Hence, i prefers to be a member of $\mathcal{B}$, and so, by Theorem 2, $\mathcal{B}$ is not stable. □

A.9 Construction of Example 3

Let: i) M agents with $\theta_H = (1 - \theta_L)/M$. This puts a bound on M since in order for $\mathcal{H}_M$ to be stable we need $M < 1/F$, hence we may take $M = \ell(\alpha) < \ell(F)$; ii) $N - M = 1 + J$ agents with $\theta_L \leq F$. Hence, $\Theta_{\mathcal{H}_M} = (1 - \theta_L)$ and $\Theta_{\mathcal{N}} = 1 + J\theta_L$. In these conditions, $\mathcal{H}_M$ is stable by Corollary 2. Also observe that $\theta_L \leq F < F/(1 - F) < 1/(1 + M)$ which means that $\theta_H > \theta_L$. Therefore,

$$\tilde{U} = U(\alpha; \mathcal{H}_M/\mathcal{N}) = \left(\frac{1 - \theta_L}{1 + J\theta_L}\right)^{\frac{1}{1-\alpha}} \geq \eta,$$

which is equivalent to

$$\theta_L \leq \frac{1-\eta^{1-\alpha}}{1+J\eta^{1-\alpha}}.$$

Also, for the welfare we have

$$\begin{aligned} W(\alpha;\mathcal{H}_M/\mathcal{N}) &= \frac{(1+J\theta_L)\left(\alpha(1-\theta_L)\right)^{\frac{\alpha}{1-\alpha}} - \left(\alpha(1-\theta_L)\right)^{\frac{1}{1-\alpha}}}{\left(\alpha(1+J\theta_L)\right)^{\frac{1}{1-\alpha}}(1/\alpha-1)} \\ &= \left(\frac{1-\theta_L}{1+J\theta_L}\right)^{\frac{\alpha}{1-\alpha}}\left(\frac{1}{1/\alpha-1}\right)\frac{(1+J\theta_L-\alpha(1-\theta_L))}{\alpha(1+J\theta_L)} \\ &= \frac{\tilde{U}^\alpha-\alpha\tilde{U}}{1-\alpha} \geq \frac{\eta^\alpha-\alpha\eta}{1-\alpha} \geq \eta-\eta\log(\eta), \end{aligned}$$

since this is an increasing function of $\tilde{U}$ and a decreasing function of α, allowing us to let α go to 1.

Furthermore, we have that $\theta_1 = (1-\theta_L)/\ell(\alpha)$, and so

$$U(\alpha;\mathcal{H}_1/\mathcal{N}) = \left(\frac{(1-\theta_L)/\ell(\alpha)}{1+J\theta_L}\right)^{\frac{1}{1-\alpha}} = \ell(\alpha)^{-\frac{1}{1-\alpha}}U(\alpha;\mathcal{H}_M/\mathcal{N}) < \ell(\alpha)^{-\frac{1}{1-\alpha}},$$

which tends to 0 when α goes to 1.

Regarding the welfare, we have that

$$W(\alpha;\mathcal{H}_1/\mathcal{N}) = \frac{(1+J\theta_L)\left(\alpha(1-\theta_L)/\ell(\alpha)\right)^{\frac{\alpha}{1-\alpha}} - \left(\alpha(1-\theta_L)/\ell(\alpha)\right)^{\frac{1}{1-\alpha}}}{\left(\alpha(1+J\theta_L)\right)^{\frac{1}{1-\alpha}}(1/\alpha-1)}.$$

Hence

$$\begin{aligned} W(\alpha;\mathcal{H}_1/\mathcal{N}) &< \frac{\left(\alpha(1-\theta_L)/\ell(\alpha)\right)^{\frac{\alpha}{1-\alpha}}/\alpha}{\left(\alpha(1+J\theta_L)\right)^{\frac{\alpha}{1-\alpha}}(1/\alpha-1)} = \left(\frac{(1-\theta_L)/\ell(\alpha)}{1+J\theta_L}\right)^{\frac{\alpha}{1-\alpha}}\left(\frac{1}{1-\alpha}\right) \\ &= \frac{\ell(\alpha)^{-\frac{\alpha}{1-\alpha}}}{1-\alpha}U(\alpha;\mathcal{H}_M/\mathcal{N})^\alpha < \frac{\ell(\alpha)^{-\frac{\alpha}{1-\alpha}}}{1-\alpha}, \end{aligned}$$

which tends to 0 when α goes to 1. □

B A note on the stability of low-cooperation Nash–Cournot and Stackelberg equilibria

In this paper, we have only considered (focal) stand-alone strategies of coalitions, which are Nash–Cournot and Stackelberg equilibria according to the conditions in Lemmas 2 and

3. We have observed that for stable coalitions indeed these conditions hold for large α (see Theorem 3), and for stable high coalitions H_M with enough members. In this appendix, we present some observations on the external and internal stability of coalitions when Nash–Cournot and Stackelberg equilibria may also be low-cooperation strategies and such that when members enter or leave a coalition the subsequent strategy is still a Nash–Cournot or Stackelberg equilibria. When stand-alone coalitions are formed we always assume that the strategy is focal.

We first analyze external stability. Observe that when Nash–Cournot equilibria are stand-alone strategies of $\mathcal{F}$ (case 3 of Lemma 2), then when a free-rider $i \notin \mathcal{F}$ joins $\mathcal{F}$ we remain at case 3 of Lemma 2, and so Nash–Cournot equilibrium is still a stand-alone strategy of $\mathcal{F}$, which is the case covered in this paper. The same observation holds for Stackelberg equilibria since for $i \notin \mathcal{F}$ we have

$$\frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}} \leq \frac{\Theta_{\mathcal{F}}}{\theta(\mathcal{F})} < \frac{\Theta_{\mathcal{F}\cup\{i\}}}{\theta(\mathcal{F}\cup\{i\})},$$

and so Stackelberg equilibrium is still a stand-alone strategy of $\mathcal{F}$.

Low-cooperation strategies are Nash–Cournot equilibria in cases 1 and 2 of Lemma 2. In the following, it is useful to define $\beta = \Theta_{\mathcal{F}}/\theta_1$.

We first consider case 1 of Lemma 2 where $\Theta_{\mathcal{F}} < \theta_1$, *i.e.*, $\beta < 1$, and equilibria are low-cooperation strategies of $\mathcal{T}$. Assuming that the strategy of the newly formed coalition $\mathcal{F}\cup\{i\}$ (where $i \in \mathcal{N}\setminus\mathcal{F}$ is a free-rider relative to $\mathcal{F}$) remains a Nash–Cournot equilibrium, we want to compare the gains of i from joining coalition $\mathcal{F}$. As a free-rider, i gets $\theta_i \overline{r}_1^{\alpha} - r_i$, where r_i is his/her contribution in the low-cooperation (equilibrium) strategy of $\mathcal{T}$. We have that either $i \notin \mathcal{T}$ or $i \in \mathcal{T}$.

Let us consider first the case $i \notin \mathcal{T}$, so that $\theta_i < \theta_1$ and $r_i = 0$. There are three possibilities:

1. $\Theta_{\mathcal{F}\cup\{i\}} < \theta_1$, *i.e.*, $\beta < 1 - \theta_i/\theta_1$. In this case i gets the same value $\theta_i \overline{r}_1^{\alpha}$ as a free-rider or by joining $\mathcal{F}$, and so is indifferent;

2. $\Theta_{\mathcal{F}\cup\{i\}} = \theta_1$, *i.e.*, $\beta = 1 - \theta_i/\theta_1$. After joining $\mathcal{F}$, i would get $\theta_i \overline{r}_1^{\alpha} - r_i'$, where $r_i' \geq 0$ is the contribution of i after joining. So i prefers not to join $\mathcal{F}$, and so $\mathcal{F}$ is externally stable or indifferent if $r_i' = 0$;

3. $\Theta_{\mathcal{F}\cup\{i\}} > \theta_1$, *i.e.*, $\beta > 1 - \theta_i/\theta_1$. After joining $\mathcal{F}$, i would get $\theta_i \overline{r}_{\mathcal{F}\cup\{i\}}^{\alpha}(1-\alpha)$. So we obtain that i prefers not to join $\mathcal{F}$ if and only if

$$\frac{\theta_i}{\theta_1} \leq \frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}} - \beta.$$

   The expression on the right-hand side is always positive for $\beta < 1$, and so defines a non-empty region for θ_i/θ_1 where $\mathcal{F}$ is externally stable. It is easy to see that when α grows towards 1 this region of externally stability becomes smaller.

Let us now consider the case $i \in \mathcal{T}$. Then we necessarily have that $\Theta_{\mathcal{F}\cup\{i\}} > \theta_1$, and so equilibria are stand-alone strategies of $\mathcal{F}\cup\{i\}$, and so, after joining $\mathcal{F}$, i would get $\theta_1 \overline{r}^{\alpha}_{\mathcal{F}\cup\{i\}}(1-\alpha)$. So i prefers not to join coalition $\mathcal{F}$ if and only if

$$\frac{r_i}{\overline{r}_1^{\alpha}} \leq \theta_1 \left(1 - (\beta+1)^{\frac{\alpha}{1-\alpha}}(1-\alpha)\right).$$

For $\alpha \leq 1/2$ the right-hand side is always positive, since otherwise we would have $1 > \beta > 1/(1-\alpha)^{\frac{1-\alpha}{\alpha}} - 1 \geq 1$, and so it defines a non-empty region for the contribution r_i where $\mathcal{F}$ is externally stable. For $\alpha \geq 1/2$, $\mathcal{F}$ is not externally stable when β is large.

In case 2 of Lemma 2, $\Theta_{\mathcal{F}} = \theta_1$, *i.e.*, $\beta = 1$, equilibria are low-cooperation strategies of $\mathcal{T}\cup\mathcal{F}$. For any free-rider $i \in \mathcal{N}\setminus\mathcal{F}$ relative to $\mathcal{F}$ we have that $\Theta_{\mathcal{F}\cup\{i\}} > \theta_1$, and so equilibria are stand-alone strategies of $\mathcal{F}\cup\{i\}$, and so, after joining $\mathcal{F}$, i would get $\theta_i \overline{r}^{\alpha}_{\mathcal{F}\cup\{i\}}(1-\alpha)$. Again, either $i \in \mathcal{T}$ or $i \notin \mathcal{T}$.

If $i \notin \mathcal{T}$, then $r_i = 0$, and so i prefers not to join $\mathcal{F}$ if and only if

$$\frac{\theta_i}{\theta_1} \leq \frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}} - 1.$$

This condition holds for every $\alpha \leq 1/2$, since the right-hand side is greater than 1, and clearly, $\theta_i/\theta_1 \leq 1$, implying that $\mathcal{F}$ is externally stable. However, for $\alpha \geq 1/2$ the right-hand side is less than 1 and so the condition may fail to hold if θ_i is close to θ_1, in which case i prefers to join $\mathcal{F}$.

If $i \in \mathcal{T}$, then $\theta_i = \theta_1$ and i prefers not to join $\mathcal{F}$ if and only if

$$\frac{r_i}{\overline{r}_1^{\alpha}} \leq \theta_1 \left(1 - 2^{\frac{\alpha}{1-\alpha}}(1-\alpha)\right).$$

The right-hand side is negative for $\alpha \geq 1/2$, and so $\mathcal{F}$ is not externally stable. For $\alpha \leq 1/2$, it can be stable depending on whether the contribution r_i is relatively small.

We now turn to low-cooperation strategies that are Stackelberg equilibria. Define $\gamma = \Theta_{\mathcal{F}}/\theta(\mathcal{F})$. We have that $\beta \leq \gamma \leq \Theta_{\mathcal{F}}/\theta(\mathcal{F}\cup\{i\})$. From case 1 of Lemma 3 we have that $\gamma(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq 1$, and equilibria are low-cooperation strategies of the top free-riders $\mathcal{T}(\mathcal{F})$. Consider a free-rider $i \in \mathcal{N}\setminus\mathcal{F}$ relative to $\mathcal{F}$. Then, as a free-rider, i gets $\theta_i \overline{r}_1^{\alpha} - r_i$, where r_i is his/her contribution as part of the low-cooperation (equilibrium) strategy of $\mathcal{T}(\mathcal{F})$. We have that either $i \in \mathcal{T}(\mathcal{F})$ or $i \notin \mathcal{T}(\mathcal{F})$, and in the latter case $r_i = 0$.

Let us first analyse the case $i \notin \mathcal{T}(\mathcal{F})$. Then $\theta_i < \theta(\mathcal{F}) = \theta(\mathcal{F}\cup\{i\})$. There are two possibilities:

1. $\Theta_{\mathcal{F}\cup\{i\}}(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq \theta(\mathcal{F}\cup\{i\})$. In this case, i gets the same value $\theta_i \overline{r}_1^{\alpha}$ as a free-rider or by joining $\mathcal{F}$, and so is indifferent;

2. $\theta(\mathcal{F}\cup\{i\}) \leq \Theta_{\mathcal{F}\cup\{i\}}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$. After joining $\mathcal{F}$, i would get $\theta_i \overline{r}^{\alpha}_{\mathcal{F}\cup\{i\}}(1-\alpha)$. So we obtain that i prefers not to join $\mathcal{F}$ if and only if

$$\frac{\theta_i}{\theta_1} \leq \frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}} - \beta.$$

The right-hand side is always non-negative, since otherwise we would have $1 < \beta(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq \gamma(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq 1$, which is absurd. So the previous condition defines a non-empty region for θ_i/θ_1 where $\mathcal{F}$ is externally stable, occurring when the free-rider preference θ_i is small. When α grows towards 1 or when β grows this region of externally stability becomes smaller.

Let us now analyse the case $i \in \mathcal{T}(\mathcal{F})$. Then $\theta_i = \theta(\mathcal{F})$ and $\theta(\mathcal{F}\cup\{i\}) \leq \theta(\mathcal{F})$ since the preference of the top free-riders cannot increase if i joins $\mathcal{F}$. There are two possibilities:

1. $\Theta_{\mathcal{F}\cup\{i\}}(1-\alpha)^{\frac{1-\alpha}{\alpha}} \leq \theta(\mathcal{F}\cup\{i\})$. After joining $\mathcal{F}$, i would get $\theta_i \overline{r}^{\alpha}_1$, so i always prefers to join $\mathcal{F}$, and $\mathcal{F}$ is not externally stable;

2. $\theta(\mathcal{F}\cup\{i\}) \leq \Theta_{\mathcal{F}\cup\{i\}}(1-\alpha)^{\frac{1-\alpha}{\alpha}}$. After joining $\mathcal{F}$, i would get $\theta_i \overline{r}^{\alpha}_{\mathcal{F}\cup\{i\}}(1-\alpha)$. So we obtain that i prefers not to join $\mathcal{F}$ if and only if

$$\frac{r_i}{\overline{r}^{\alpha}_1} \leq \theta_i \left(1 - \left(\beta + \frac{\theta_i}{\theta_1}\right)^{\frac{\alpha}{1-\alpha}} (1-\alpha)\right).$$

The right-hand side is non-negative if and only if $\theta_i/\theta_1 \leq 1/(1-\alpha)^{\frac{1-\alpha}{\alpha}} - \beta$. Hence, when θ_i is relatively small, then it is possible for $\mathcal{F}$ to be stable if r_i is relatively small. If θ_i is not relatively small then $\mathcal{F}$ is not externally stable since i prefers to enter.

We now turn to internal stability, assuming that $\#\mathcal{F} \geq 2$. We first observe that in cases 1 and 2 of Lemma 2 we have $\Theta_{\mathcal{F}} \leq \theta_1$, and equilibria are low-cooperation strategies of $\mathcal{T}$ or of $\mathcal{F}\cup\mathcal{T}$. Let $j \in \mathcal{F}$ be a coalition member. Since $j \notin \mathcal{T}$, *i.e.*, $\theta_j < \theta_1$, j gets $\theta_j \overline{r}^{\alpha}_1$ as a coalition member or as a free-rider by leaving $\mathcal{F}$, and so is indifferent. For low-cooperation Stackelberg equilibria in case 1 of Lemma 3, as a coalition member, j gets $\theta_j \overline{r}^{\alpha}_1$. If j leaves $\mathcal{F}$, then since

$$\frac{\Theta_{\mathcal{F}\setminus\{j\}}}{\theta(\mathcal{F}\setminus\{j\})} < \frac{\Theta_{\mathcal{F}}}{\theta(\mathcal{F})} \leq \frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}},$$

we remain in case 1 of Lemma 3, and so j gets $\theta_j \overline{r}^{\alpha}_1 - r_j$, where $r_j \geq 0$ is the contribution of j if j is a top free-rider, *i.e.*, $j \in \mathcal{T}(\mathcal{F}\setminus\{j\})$, and 0 otherwise. In both cases, j is either indifferent or prefers to remain in coalition $\mathcal{F}$ if $r_i > 0$.

It remains to characterize internal stability of stand-alone strategies when the leaving of an agent results in a Nash–Cournot or Stackelberg equilibrium which is a low-cooperation

strategy. Let $j \in \mathcal{F}$ be a coalition member, which gets $\theta_j \overline{r}_{\mathcal{F}}^{\alpha}(1-\alpha)$. We first consider the Nash–Cournot case: $\Theta_{\mathcal{F}} > \theta_1$, *i.e.*, $\beta > 1$. Then if j leaves $\mathcal{F}$, Nash–Cournot equilibria are low-cooperation strategies when $\Theta_{\mathcal{F}\setminus\{j\}} \leq \theta_1$, *i.e.*, $\beta \leq 1+\theta_j/\theta_1$. There are two possibilities:

1. $j \notin \mathcal{T}$, in which case, after leaving $\mathcal{F}$, j would get $\theta_j \overline{r}_1^{\alpha}$. So we obtain that j prefers to remain in $\mathcal{F}$ if and only if
$$\beta > \frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}}.$$
If $\alpha \leq 1/2$ then this implies $\beta > 2$ which is impossible since $\beta \leq 1 + \theta_j/\theta_1 < 2$, and so $\mathcal{F}$ is not internally stable. For every β, the condition holds if α is large enough, and so $\mathcal{F}$ is internally stable;

2. $j \in \mathcal{T}$, in which case, after leaving $\mathcal{F}$, j would get $\theta_1 \overline{r}_1^{\alpha} - r_j$, where r_j is his/her contribution as part of the low-cooperation strategy of $\mathcal{T}$. So we obtain that j prefers not to leave $\mathcal{F}$ if and only if
$$\frac{r_j}{\overline{r}_1^{\alpha}} > \theta_1 \left(1 - \beta^{\frac{\alpha}{1-\alpha}}(1-\alpha)\right).$$
For $\alpha \geq 1/2$ the right-hand side is non-negative and the condition defines a region of small contributions r_j such that $\mathcal{F}$ is not internally stable. For given β if α is sufficiently large then the right-hand side is negative and so $\mathcal{F}$ is internally stable regardless of the contribution r_j.

Now for the Stackelberg case we have that $\gamma(1-\alpha)^{\frac{1-\alpha}{\alpha}} \geq 1$. We want to analyse the case where j leaves $\mathcal{F}$ and Stackelberg equilibria become low-cooperation strategies of $\mathcal{T}(\mathcal{F}\setminus\{j\})$. There are two possibilities:

1. $j \notin \mathcal{T}(\mathcal{F}\setminus\{j\})$, in which case, after leaving $\mathcal{F}$, j would get $\theta_j \overline{r}_1^{\alpha}$. So we obtain that j prefers to remain in $\mathcal{F}$ if and only if
$$\beta > \frac{1}{(1-\alpha)^{\frac{1-\alpha}{\alpha}}}.$$

2. $j \in \mathcal{T}(\mathcal{F}\setminus\{j\})$, in which case, after leaving $\mathcal{F}$, j would get $\theta_j \overline{r}_1^{\alpha} - r_j$, where r_j is his/her contribution as part of the low-cooperation strategy of $\mathcal{T}(\mathcal{F}\setminus\{j\})$. So we obtain that j prefers not to leave $\mathcal{F}$ if and only if
$$\frac{r_j}{\overline{r}_1^{\alpha}} > \theta_j \left(1 - \beta^{\frac{\alpha}{1-\alpha}}(1-\alpha)\right).$$

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020, and within project "Modelling, Dynamics and Games" with reference PTDC/MAT-APL/31753/2017.

Elvio Accinelli wish to thank CUMEX (Consortium of Mexican Universities) and the Ibero-American Postgraduate University Association (AUIP) for the support granted for his stay in Portugal between July 10 and August 24, 2021, and to thank the dean of the Faculty of Economic Sciences of the UASLP for the support provided for the academic stay at Porto, and also the Department of Mathematics of University of Porto for their hospitality.

Atefeh Afsar thanks the financial support of FCT through a PhD. grant of the MAP-PDMA program with the reference PD/BD/142886/2018.

Filipe Martins was partially supported by CMUP, member of LASI, which is financed by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the project with reference UIDB/00144/2020.

Jorge Oviedo was partially supported by UNSL through grants 03-2016 and 03-1323, and from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) through grant PIP 112-200801-00655, and from Agencia Nacional de Promoción Científica y Tecnológica through grant PICT 2017-2355.

References

[1] Elvio Accinelli, Filipe Martins, and Alberto A. Pinto. Evolutionary dynamics for the generalized Baliga–Maskin public good model. *Chaos, Solitons & Fractals*, page 109496, 2019. URL: http://www.sciencedirect.com/science/article/pii/S0960077919304424, doi:https://doi.org/10.1016/j.chaos.2019.109496.

[2] Elvio Accinelli, Filipe Martins, and Alberto A. Pinto. The basins of attraction in the generalized Baliga–Maskin public good model. *Journal of Evolutionary Economics*, 2021. URL: https://link.springer.com/article/10.1007/s00191-021-00758-z, doi:10.1007/s00191-021-00758-z.

[3] Sandeep Baliga and Eric Maskin. Mechanism design for the environment. In K.G. Mäler and J. Vincent, editors, *Handbook of environmental economics*, pages 306–324. Elsevier Science/North Holland, 2003. URL: https://www.sciencedirect.com/science/article/abs/pii/S157400990301012X, doi:10.1016/S1574-0099(03)01012-X.

[4] Scott Barrett. Self-Enforcing International Environmental Agreements. *Oxford Economic Papers*, 46(Supplement 1):878–894, 10 1994. URL: https://

academic.oup.com/oep/article-abstract/46/Supplement_1/878/2568546, doi:10.1093/oep/46.Supplement_1.878.

[5] Scott Barrett. Climate treaties and approaching catastrophes. *Journal of Environmental Economics and Management*, 66(2):235–250, 2013. URL: https://www.sciencedirect.com/science/article/abs/pii/S0095069612001222, doi:10.1016/j.jeem.2012.12.004.

[6] Basak Bayramoglu, Michael Finus, and Jean-François Jacques. Climate agreements in a mitigation-adaptation game. *Journal of Public Economics*, 165:101–113, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0047272718301300, doi:https://doi.org/10.1016/j.jpubeco.2018.07.005.

[7] Pierre Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Springer-Verlag, 1991. doi:https://doi.org/10.1007/978-3-030-45982-6.

[8] R. H. Coase. The problem of social cost. *The Journal of Law and Economics*, 3:1–44, 1960. arXiv:https://doi.org/10.1086/466560, doi:10.1086/466560.

[9] Astrid Dannenberg, Andreas Lange, and Bodo Sturm. On the Formation of Coalitions to Provide Public Goods - Experimental Evidence from the Lab. NBER Working Papers 15967, National Bureau of Economic Research, Inc, May 2010. URL: https://ideas.repec.org/p/nbr/nberwo/15967.html.

[10] Claude d'Aspremont, Alexis Jacquemin, Jean Gabszewicz, and John Weymark. On the stability of collusive price leadership. *Canadian Journal of Economics*, 16(1):17–25, 1983. URL: https://www.jstor.org/stable/134972, doi:https://doi.org/10.2307/134972.

[11] Robyn M Dawes. Social dilemmas. *Annual review of psychology*, 31(1):169–193, 1980. URL: https://www.annualreviews.org/doi/abs/10.1146/annurev.ps.31.020180.001125?journalCode=psych, doi:https://doi.org/10.1146/annurev.ps.31.020180.001125.

[12] Effrosyni Diamantoudi and Eftichios S Sartzetakis. Stable international environmental agreements: An analytical approach. *Journal of Public Economic Theory*, 8(2):247–263, 2006. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9779.2006.00262.x, doi:10.1111/j.1467-9779.2006.00262.x.

[13] Klaus Eisenack and Leonhard Kähler. Adaptation to climate change can support unilateral emission reductions. *Oxford Economic Papers*, 68(1):258–278, 2016. URL: https://academic.oup.com/oep/article-abstract/68/1/258/2362411, doi:10.1093/oep/gpv057.

[14] Michael Finus, Francesco Furini, and Anna Viktoria Rohrer. The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs Stackelberg leadership. *Journal of Environmental Economics and Management*, 109:102461, 2021. URL: https://www.sciencedirect.com/science/article/pii/S0095069621000449, doi:https://doi.org/10.1016/j.jeem.2021.102461.

[15] Michael Finus and Matthew McGinty. The anti-paradox of cooperation: Diversity may pay! *Journal of Economic Behavior & Organization*, 157:541 – 559, 2019. URL: http://www.sciencedirect.com/science/article/pii/S0167268118302890, doi:https://doi.org/10.1016/j.jebo.2018.10.015.

[16] Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, 71(3):397–404, 2001. doi:10.1016/S0165-1765(01)00394-9.

[17] Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr. Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24(3):153 – 172, 2003. URL: http://www.sciencedirect.com/science/article/pii/S1090513802001575, doi:https://doi.org/10.1016/S1090-5138(02)00157-5.

[18] Eban S. Goodstein and Stephen Polasky. *Economics and the Environment*. Wiley, 8th edition, 2017. URL: https://www.wiley.com/en-us/EconomicsandtheEnvironment,8thEdition-p-9781119369868.

[19] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. URL: https://science.sciencemag.org/content/162/3859/1243, arXiv:https://science.sciencemag.org/content/162/3859/1243.full.pdf, doi:10.1126/science.162.3859.1243.

[20] Russell Hardin. *Collective Action*. Resources for the Future, 1982.

[21] Shirli Kopelman, John Mark Weber, and David M. Messick. *Factors influencing cooperation in commons dilemmas: A review of experimental psychological research*, chapter 4, pages 113–156. The National Academies Press., 2002. doi:10.17226/10287.

[22] A. Mas-Colell. *Cooperative Equilibrium*, pages 95–102. Palgrave Macmillan UK, London, 1989. doi:10.1007/978-1-349-20181-5_7.

[23] Eric Maskin. Mechanism design: How to implement social goals. *The American Economic Review*, 98(3):567–576, 2008. URL: http://www.jstor.org/stable/29730086.

[24] Mancur Olson. *The logic of collective action*. Harvard University Press, 2009 (originally published 1965). URL: https://www.hup.harvard.edu/catalog.php?isbn=9780674537514.

[25] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Canto Classics. Cambridge University Press, 2015 (originaly published 1990). doi:10.1017/CBO9781316423936.

[26] Plato. *Republic*. Project Gutenberg, 2008. Translated by Benjamin Jowett. URL: http://www.gutenberg.org/ebooks/1497.

[27] John E. Roemer. Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, 127:45–57, 2015. doi:https://doi.org/10.1016/j.jpubeco.2014.03.011.

[28] John E. Roemer. *How We Cooperate: A Theory of Kantian Optimization*. Yale University Press, 2019. URL: http://www.jstor.org/stable/j.ctvfc52jk.

[29] Santiago J. Rubio and Alistair Ulph. Self-enforcing international environmental agreements revisited. *Oxford economic papers*, 58(2):233–263, 2006. URL: https://academic.oup.com/oep/article-abstract/58/2/233/2361959, doi:10.1093/oep/gpl002.