Essays in Applied Microeconomic Theory

Inauguraldissertation
zur Erlangung des Grades eines Doktors
der Wirtschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität
Bonn

vorgelegt von
Melina Maria Cosentino
aus München

2024

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Dezsö Szalay
Zweitreferent:	Prof. Dr. Sven Rady
Tag der mündlichen Prüfung:	15.02.2024

Acknowledgements

This thesis would not have been possible without the generous support of many people:

First of all, I would like to thank Dezsö Szalay, my first supervisor, who gave me the freedom to pursue the projects I thought interesting while still providing invaluable guidance. Dezsö always knew the correct questions to ask and hinted at potential gaps in my analyses without ever giving me the impression of having failed or gone in the wrong direction. He immediately understood my priorities, intentions and needs and adjusted his advise accordingly.

Sven Rady went above and beyond what would be expected from a second supervisor. In spite of his busy schedule, he took the time to provide extraordinarily detailed feedback on my papers without being pedantic. He always found time for my questions, even before becoming my second supervisor. Besides giving me valuable content-related advise, he also provided extensive support in various other respects: together with Daniel Krähmer, he made it possible for me to join the Collaborative Research Center TR 224 (CRC), appointed me as a member of the CRC's board and, as did Dezsö, supported my transition to a life outside academia.

Furthermore, I want to express my gratitude to Daniel Krähmer who, even though not my supervisor, invested a lot of time and effort into advising me. Not only did he make it possible for me to join the CRC, he also provided highly valuable feedback on all of my projects, welcoming me in his office for multiple hour-long meetings. His advise was crucial for the progress of my papers.

Moreover, I benefited significantly from joining the CRC, which introduced me to a network of experienced and motivated researchers and enabled me to participate in fruitful discussions in internal workshops. In particular, I would like to mention the monthly meetings of the project group I was assigned to. Daniel Krähmer, Volker Nocke and Nicolas Schutz have created a room for interesting and insightful discussions without any pressure to perform, which I really enjoyed.

Lastly, I am infinitely grateful for my fiancé and co-author's support: Philipp helped me push through the downs during my time as a PhD student, he always lent a patient ear to my thoughts, offered detailed feedback and gave me confidence

when I found myself lacking in it. He had my back—always and unconditionally. Without him, I would not be where I am. Thank you, Philipp, thank you for being who you are, for letting me live this life by your side—forever.

Contents

List of figures

List of tables

Introduction

This dissertation is composed of three self-contained chapters whose main objective lies in providing model-based explanations of and policy recommendations for "real-world" phenomena that attracted my attention throughout the doctoral studies. While the first chapter aims to assess the effect of negative preferences on political parties' behaviour, thereby providing an explanation for the rise of political extremism, the second and third chapters focus on policy recommendations: in Chapter 2, Philipp Hamelmann and I analyse a communication protocol that enhances information transmission between experts and a decision maker in the presence of a conflict of interest; Chapter 3 sheds light on (potential) adverse effects of climate policies and suggests procedures that enable regulators to avoid them.

Antipartisanship—an explanation for extremism?, Chapter 1: In this chapter, I adapt the Hotelling-Downs model with three parties and add an "antipartisan" component: antipartisans vote for the party located furthest away from their most disliked party. While the standard game without antipartisanship and uniformly distributed voters has no pure-strategy equilibrium, the present model allows for equilibria with, depending on the share of antipartisan voters, either distinct moderate or extreme party-positions. This provides a theoretical explanation for phenomena such as those observed in Brazil in 2018: an exogenous increase in antipartisanship, followed by polarisation. I characterise the conditions under which a change in antipartisanship by itself can explain such comparative statics.

Eliciting information from multiple experts via grouping, Chapter 2: This chapter is joint work with Philipp Hamelmann. We analyse a set-up in which a decision maker (DM) seeks to determine whether to adopt a new policy or maintain the status quo. To do so, she consults (finitely many) experts whose common interests differ significantly from those of the DM. As suggested by Wolinsky (2002)[1], partial communication ("grouping mechanisms") among experts can—requiring neither transfers nor commitment—result in revelation of more information than full communication: by allowing for communication within groups of experts only and,

[1] **Wolinsky, Asher.** 2002. "Eliciting information from multiple experts." *Games and Economic Behavior*, 41(1): 141 – 160.

hence, changing the events in which votes are pivotal, the DM may be able to manipulate experts' strategies to her advantage. We elaborate on this, inter alia, characterising optimal grouping mechanisms and conditions under which grouping can improve upon full communication.

Climate clubs: adverse effects and how to avoid them, Chapter 3: This chapter revolves around the "G7 Climate Club" whose main objective lies in promoting the implementation of the Paris Agreement. One of the proposed measures is a reduction in the production of emission-intensive goods. Analysing imperfect competition in a market for such a good, I highlight risks of said intervention: a reduced production by club members may increase the total level of emissions. In the new equilibrium, non-members raise production—potentially offsetting reduced emissions in member countries. For some parametrisations, the club needs to increase, not decrease its production and emissions to minimise the aggregate emission level. I discuss (1) conditions under which there is a risk of such adverse effects and (2) analyse the optimal club production levels; both (1) and (2) are highly dependent on the exact market structure and may, hence, be unknown to the club. As a remedy, I propose interventions that are, for virtually all parametrisations, guaranteed to reduce emissions, not harm consumers and can, other than the optimal production levels, be implemented without detailed knowledge of the market structure.

Chapter 1

Antipartisanship—an explanation for extremism?*

1.1 Introduction

In most voting models considered in economic research, preferences are formulated positively: voters are assumed to vote for their most preferred option. By assumption, such models are unable to describe situations in which agents cannot pinpoint a favourite outcome but only one they like least. Political scientists, however, have come to the conclusion that "negative partisanship"[1] has been relevant in parliamentary and presidential elections all over the world (e.g. cf. Medeiros and Noël, 2014; Caruana, McGregor and Stephenson, 2015; Mayer, 2017; Samuels and Zucco, 2018). To address this blind spot in the literature, I propose a Hotelling-Downs-type model in which some voters have negative rather than positive preferences. At first glance, one might think it irrelevant for an election's outcome whether agents vote for their favourite or try to impede their least favourite party's success: the two could be mistaken as primal and dual of the same underlying problem. As will be shown below, this is not the case; the presence of negative partisanship can change an election's outcome substantially.

As a case-study, consider Brazil's general election in 2018. The country's workers' party (left-wing), Partido dos Trabalhadores (PT), had maintained both the plurality of seats in the chamber of deputies and won all presidential elections since

*University of Bonn, Bonn Graduate School of Economics (BGSE), *melina.cosentino@outlook.de*.
Funding by the Bonn Graduate School of Economics (BGSE), the German Research Foundation (DFG) through CRC TR 224 (project B03) and Studienstiftung des Deutschen Volkes e.V. is gratefully acknowledged.
I am grateful for valuable comments by Philipp Hamelmann, Dezsö Szalay, Sven Rady, Daniel Krähmer, Florian Brandl and participants of the BGSE-workshop.

[1] Caruana, McGregor and Stephenson (2015) define: "Holding a negative partisanship toward a party is an affective repulsion from that party".

2002. In 2018, it was beaten by a (previously) minor party, the Partido Social Liberal (PSL). Lead by Jair Bolsonaro, the PSL had evolved from a median-right-wing party, hardly ever obtaining a single seat in the Chamber of Deputies, to a radical right party, winning the presidential election by a margin of ten percentage points. What can explain this change in political tides?

In 2015, after a long period of high levels of government spending, an economic downturn and high inflation levels during the reign of the PT, the so called "anti-PT-ismo" (also referred to as "antipetismo") movement gained momentum.[2] The PSL's leaders recognised this as a unique chance to increase their vote share. Expressing explicit aversion against the PT and moving even further away from it on the political spectrum, the PSL had soon become the archetypal "anti-PT" party. Among political scientists, there seems to be a broad agreement with respect to the relevance of antipetismo for the PSL's sudden success (e.g. Samuels and Zucco, 2018; do Amaral, 2020; Fuks, Ribeiro and Borba, 2021). This pattern—the rise of a movement *against* a specific agenda or party, followed by polarisation—has been observed in other countries, including Canada, the United States and Europe (cf. Mayer, 2017; Abramowitz and Webster, 2016; Caruana, McGregor and Stephenson, 2015; Casalecchi, Borges and Renno, 2020; Arzheimer and Berning, 2019; Medeiros and Noël, 2014).

This paper provides a deeper understanding of such occurrences. To this end, I adapt the Hotelling-Downs framework for political decision-making with three parties (Downs, 1957) and add the "antipartisan" component: antipartisans vote for the party located furthest away from their most disliked party. While the standard model without antipartisanship and uniformly distributed voters has no equilibrium in pure strategies, the present one is able to explain both the spreading of parties over the political spectrum and extreme positions. Moreover, it provides a theoretical explanation for the aforementioned phenomena: an exogenously triggered increase in the share of antipartisans changing the best response of parties, followed by relocation on the political spectrum and the forming of a new equilibrium. Specifically, I characterise all possible outcomes in the present adaptation of the Hotelling-Downs model under convex relocation costs (Theorem 1): depending on the share of antipartisans, there is either no equilibrium, parties locate at the extrema, or they choose a specific combination of moderate positions (not all equal). Thereafter, I derive under which circumstances a change in the share of antipartisans can, by itself, explain transitions between them (Theorems 2 and 4).

[2]Antipetismo had been present before 2015, the year in which, fuelled by scandals and general discontent with Brazil's current leadership, it reached an unprecedented size and impact, ending with the call for impeachment of president Dilma Rousseff (PT). In 2016, Dilma Rousseff was officially suspended (Davis and Straubhaar, 2020).

In essence, these results provide a theoretical characterisation of political systems whose structure may be affected by an increase or decrease in negative partisanship.

The next sections are structured as follows: In Section 1.2, I describe the model and explain its novel features. Section 1.3, the core of the analysis, is divided into three parts: In sections 1.3.1 and 1.3.2, I present and discuss the most important findings. Readers interested in an in-depth and more technical analysis thereof (along with some additional results) are referred to Section 1.3.3. Section 1.4 relates my work to similar adaptations of the Hotelling-Downs model; Section 1.5 concludes.

1.2 Model

The election consists of three stages: First, parties choose their positions simultaneously. Then, every voter submits one vote, after which the election's winner is announced (plurality voting). The following paragraphs contain detailed descriptions of parties' and voters' behaviour, respectively.

Parties

There are three parties, A, B and C, which simultaneously choose positions p_A, p_B and p_C on the political spectrum $[0,1]$ to maximise their expected payoff. Each party has a predetermined ideology, α, β and γ in $[0,1]$. All ideologies are common knowledge. Labelling is always chosen such that (w.l.o.g.) $\alpha \leq \beta \leq \gamma$.

A party's payoff depends on its winning or losing the election and its position. Let $w_i = 1$ $i \in \{A,B,C\}$ if party i receives the plurality of votes, $w_i = 0$ otherwise; then:

$$U(w_A, p_A) = w_A - c(|p_A - \alpha|)$$

(*mutatis mutandi* for B and C), where $c:\ [0,1] \to \mathbb{R}^+$ is a non-decreasing, continuous function satisfying $c(0) = 0$ and $c(x) > 0\ \forall x \neq 0$.
Hereinafter, the term "expected" in "expected vote share" and "expected payoff" is dropped when there is no risk of confusion.

The cost function has many interpretations. It could be the cost a party faces upon changing its political agenda: posters, speeches, websites, video-clips and any other advertising would have to be adjusted. Similarly, the process of persuading fellow party members of the change to be undergone may also cost resources. Another way to justify the introduction of c could be career concerns; it may be that a member of the party's list of candidates himself cares only about his success in an election. To get his party's nomination, however, he first has to gain support among

his fellow party members. Hence, an intrinsically office-only-motivated candidate has to take into account (his party members') ideological concerns as well. I find that, even if the degree of ideological concerns is small, the introduction of ideologies and respective costs is able to explain movements across the spectrum that seem, from an empirical point of view, very familiar and cannot be captured by standard Hotelling-Downs models.

Voters

There are two types of voters: *partisans* and *antipartisans*. Partisans and antipartisans are uniformly distributed on the unit interval. There is a share of q partisan- and $1-q$ antipartisan-voters, where $q \in (0,1)$.

A partisan votes (as in standard models) for the party closest to his position on the political spectrum. Below, I consider a representative partisan, referred to as voter j with position x_j.

In contrast, an antipartisan votes according to an "anti-preference". There is one party on the political spectrum—the party closest to the voter's anti-preference—whose probability of winning the election he wants to *minimise*. He votes for the party located furthest away from this most disliked party. Below, I consider a representative antipartisan, referred to as k, with position x_k. To formalise and contrast, the next paragraph describes the representative voters' problems.[3]

Partisan j solves

$$\max_{i\in\{A,B,C\}} -(p_i - x_j)^2.$$

By contrast, antipartisan k solves

$$\max_{i\in\{A,B,C\}} \left(p_i - p^-(x_k)\right)^2,$$

where $p^-(x_k)$ is defined to be the location of the party closest to voter k's position:

$$p^-(x_k) := \operatorname{argmax}_{p\in\{p_A,p_B,p_C\}} -(p - x_k)^2.$$

An antipartisan's preference is not to be confused with a simple inversion of a partisan's preference: in some cases, an antipartisan might not vote for the party furthest away from x_k. To illustrate the distinction between an inversion of partisanship and antipartisanship, consider Figure 1.1.

For any realization of x_k in the interval $(0, 0.55)$, antipartisan k votes for party C. In this case, B or A are the closest parties to x_k (i.e. the antipartisan's most disliked parties). C is the party located furthest away from p_A and p_B. If x_k lies in

[3] Note that the distance need not be quadratic but could be any symmetric distance function.

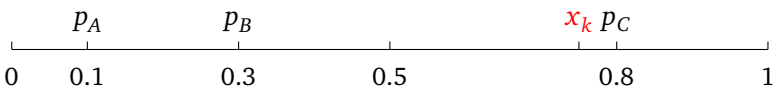

Figure 1.1: Antipartisan on the political spectrum

$(0.55, 1)$, party A receives antipartisan k's vote: then, C is k's most disliked party (p_A lies furthest away from p_C). Mind that, as the median party can never be the one furthest away from any other party, it never receives antipartisans' votes.

Now, consider a pure inversion of partisanship wherein k simply votes for the party furthest away from x_k. Such an "inverted partisan" would solve the following problem:

$$\max_{i \in \{A,B,C\}} (p_i - x_k)^2$$

which is not necessarily equivalent to an antipartisan's maximisation problem. To illustrate, consider Figure 1.2: The closest party to the antipartisan's antipreference point $x_k = 0.5$ is party B. He votes for party C, as C is located furthest away from B. An inverted partisan however, would vote for party A, being the party located furthest away from $x_k = 0.5$.

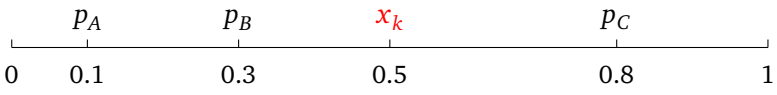

Figure 1.2: Antipartisanship $\neq$ inverted partisanship

Put differently, antipartisans follow a "minimax" strategy and inverted partisans are minimisers in the traditional sense. More precisely, an antipartisan minimises his maximum loss—he maximises his utility conditional on his most disliked party winning the election. For illustration, suppose the antipartisan in Figure 1.2 votes for party A instead of C and B receives enough votes to win the election—the worst-case scenario for the antipartisan. Suppose now that B aims to pass a bill. The antipartisan has an interest in impairing B's ability to do so. Having supported the opposition and the party (C) that is least likely to vote for any proposal made by B, he is best off in such scenario. Had he voted for A, he would have risked A to cooperate with B in the process of passing the bill. Hence, had he acted as an inverted partisan, the antipartisan would have, indirectly, supported the success of party B. Considering the motivation of the model, it seems more reasonable to study antipartisans rather than inverted partisans.

Throughout the analysis, I consider only strict Nash equilibria. A Nash equilibrium is said to be strict if best responses are unique. By definition, this rules out

mixed strategies, which I consider a reasonable restriction due to the context my model is built around. Adopting a mixed strategy in this environment would imply parties not making clear statements as to which policies they would implement when elected.[4] A central aspect of political campaigning is the clear statement of plans and aims a party has for its potential legislature. In discussions with competitors, candidates need to have clear and transparent opinions and ideas. These fundamental aspects of campaigning for elections hardly allow for mixed strategies. The ruling out of weak best responses, on the other hand, precludes equilibria that occur only for very particular parameter constellations and break down as soon as one of them changes infinitesimally.[5] Hence, unless stated otherwise, the term "equilibrium" always refers to "strict Nash equilibrium".

1.3 Main analysis and results

This section will be structured as follows: I start by providing a quick discussion of the expected vote-share function in Section 1.3.1. While technical in nature, a fundamental understanding of the expected vote share is essential for all later analyses as it determines strategies and outcomes. Thereafter, I state the paper's main results and insights (Section 1.3.2): Theorem 1 characterises the set of equilibria that can arise in games with convex cost functions. Theorem 2 states under which conditions a change in the share of antipartisanship by itself can explain transitions between moderate and extreme positions (i.e. comparative statics similar to those observed in Brazil in 2018). In Section 1.3.3, I seek to provide a deeper understanding of the main findings' origins and discuss further results; as its content is supplementary in nature, it may be skipped by readers who prefer not to delve into technical details and are satisfied by the explanations provided in Section 1.3.2.

1.3.1 Expected vote share

In this section, I discuss determinants of parties' expected vote-share function, which is at the core of all below analyses and results. Figure 1.3 depicts a combination of positions of parties A, B and C. In this case, none of the parties share a position and the distances between them are not equal ($p_B - p_A \neq p_C - p_B$). As equal

[4] During the election, not even the party members themselves would know the policy they would implement when elected.

[5] For instance, if, under the assumption of linear costs, the share of antipartisans happens to be equal to exactly half of the constant marginal cost of movement, one party might be indifferent between moving towards one of its opponents and remaining on its point of ideology. As soon as q or the cost parameter change infinitesimally, the indifference breaks down and one action is strictly superior.

distances and positions ($p_A = p_B = p_C$) virtually never constitute an equilibrium (Lemma 1) and neither do equal distances (Lemma 2), this discussion covers almost all important combinations for later analyses. The case in which two parties share a position will not be considered, as, resembling a game with two players, it is relatively simple to analyse.

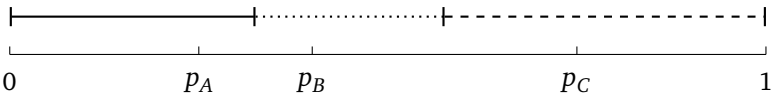

Figure 1.3: Positions on the spectrum

The solid, dotted and dashed lines in Figure 1.3 facilitate the reading of Table 1.1: The solid line, for instance, is of length $p_A + \frac{p_B - p_A}{2}$. As Table 1.1 suggests, all partisans whose position lies within this section of the spectrum vote for party A. Similarly, all partisans located on the dashed line vote for party C.

Party	Partisan-votes	Antipartisan-votes
A	$\left(p_A + \frac{p_B - p_A}{2}\right)q$ solid line	$\left(1 - p_C + \frac{p_C - p_B}{2}\right)(1-q)$ dashed line
B	$\frac{p_C - p_A}{2}q$ dotted line	— none
C	$\left(1 - p_C + \frac{p_C - p_B}{2}\right)q$ dashed line	$\left(p_A + \frac{p_B - p_A}{2} + \frac{p_C - p_A}{2}\right)(1-q)$ solid and dotted line

Table 1.1: Expected vote share for positions in Figure 1.3

In the equilibrium depicted in Figure 1.3, party A has chosen a position closer to party B than C—an important detail, as it determines whether party B's antipartisans (antipartisans that are located closest to party B; dotted line) vote for party A or party C. In this case, party A receives no votes by B's antipartisans and only those of antipartisans located closest to party C (dashed line). Party C, on the other hand, receives votes of both A's and B's antipartisans, resulting in a relatively strong position. Party B is worst off when it comes to antipartisans' votes: as the median, it does not receive any votes by antipartisans. B's vote share equals $\frac{p_C - p_A}{2}q$ (dotted line) and cannot be influenced by movements of B within the bounds of p_A and p_C. As parties A and C also receive votes by partisans (solid and dashed line, respectively), their positions are stronger than B's even for very small shares of antipartisanship (cf. Table 1.1).

This short analysis delivers a number of key insights: First, it illustrates that, in most cases, the weakest party is the median one, as it cannot influence its vote

share by movements within the bounds of its opponents' positions (here (p_A, p_C)). Second, the strongest party is likely to be the party located further away from the median, as it receives votes by both opponents' antipartisans. Accordingly, the median is the party most tempted to deviate from its position and, hence, important when determining whether a combination of positions constitutes an equilibrium. This will be apparent in two of the paper's main results (Theorems 2 and 3). Lastly, Figure 1.3 illustrates how antipartisanship determines parties' positions: for high shares of antipartisanship, the two extreme parties are likely to move closer to the extrema (thereby, increasing the length of the dotted line) to receive more votes by antipartisans; for low shares of antipartisanship, they are tempted to move closer to the median (increasing the length of the solid and dashed lines, respectively).

1.3.2 Main results

In this section, I discuss the paper's main results, which characterise the set of equilibria and state under which conditions antipartisanship by itself can explain transitions between equilibria with moderate and extreme party-positions. To do so, define the following *classes* of equilibria:

Definition 1. *(Three classes of equilibria)*

1. ***Ideology-faithful** equilibrium: parties choose distinct positions $p_A < p_B < p_C$.*
2. ***Knife-edge alliance** equilibrium: exactly two parties share a position that is neither 0 nor 1; either $p_A = p_B < p_C$ or $p_A < p_B = p_C$.*
3. ***Extreme** equilibrium: parties locate at the extrema.*

*Equilibria that are not extreme are referred to as **moderate**.*

The definition of ideology-faithful equilibria is self-explanatory: the ordering of parties' ideologies determines that of their positions. The definition of knife-edge alliance equilibria, on the other hand, might seem odd at first glance; it is, however, not arbitrary: As will be discussed in Section 1.3.3 (Lemma 4), alliances between two parties are only stable if they constitute a knife-edge alliance equilibrium. Furthermore, such equilibria are, as the name suggests, unstable: the shared position must be exactly equal to a certain value and even small changes in parameter values or other parties' positions disrupt the equilibrium. The degree of coordination between parties required in such equilibrium seems rather questionable given their rivalry. Ideology-faithful and extreme equilibria, on the other hand, are more stable and require less coordination: in the vast majority of cases, small deviations of other parties do not impact the best response of the party under consideration,

while knife-edge alliance equilibria are not robust to such "trembles". The same holds for small changes in parameter values.

The technical language at hand, the main results can be stated: Theorem 1 characterises the set of possible outcomes of games with convex cost functions.

Theorem 1. *Suppose the cost function is convex and continuously differentiable. If an equilibrium exists, it is unique and either an ideology-faithful, a knife-edge alliance, or an extreme equilibrium.*
If the cost function is linear, parties locate at their ideologies in ideology-faithful equilibria.

The insights in Theorem 1 are manifold, as it characterises the combinations of positions that may constitute an equilibrium.[6] To illustrate, consider the case in which costs are linear: While the standard game without antipartisans does not have an equilibrium in pure strategies (cf. Osborne, 1993), antipartisans allow for the existence of three classes of equilibria. Parties may locate at the extrema, or they choose their points of ideology. For relatively small shares of antipartisanship and under a number of requirements on parties' ideologies (Lemma 4), it can be a best response to share a position on the spectrum.

The existence of equilibria in the presence of antipartisans is, actually, rather intuitive: the most important properties of the standard Hotelling-Downs game with three parties that preclude the existence of equilibria are (1) the continuous action space and (2) the fact that all parties are drawn towards the median of the spectrum. While this model also satisfies the continuity assumption (1), which reduces the set of equilibria substantially, the presence of antipartisans counteracts the convergence result. In contrast to the standard model, parties are not only drawn towards the median (trying to gain partisans' votes) but also to the extrema (for antipartisans' votes). Antipartisanship encourages maximum, partisanship minimum differentiation. This explains the existence and properties of equilibria: For high shares of antipartisanship, parties are drawn towards the extrema (maximum average distance) and only extreme equilibria exist. For low shares of antipartisanship, the game is analogous to the standard game; that is, no equilibrium exists as all parties seek to locate at the median. For moderate shares of antipartisanship, the two forces may balance one another out and moderate equilibria arise.[7]

[6] Theorem 1 characterises all equilibria that may exist when costs are convex. It does not state that such equilibria do, indeed, exist for some parameter constellations: this will follow from Theorem 2.

[7] Note that while the term "balance" might suggest that moderate equilibiria only exist when $q = 0.5$ (as many antipartisans as partisans), moderate equilibria can arise for a continuum of values of q.

Moreover, Theorem 1 states that when the cost function is linear, parties locate at their points of ideology in any equilibrium with distinct positions $p_A < p_B < p_C$: as the marginal cost of movement is independent of a party's position, deviating from one's point of ideology and maintaining $p_A < p_B < p_C$ either strictly dominates (very high or very low shares of antipartisanship) or is strictly dominated (moderate shares of antipartisanship) by not moving. When marginal costs of movement are not constant, this might not be the case. As will be discussed closer in Section 1.3.3, parties choose "ideology-faithful" ($p_A < p_B < p_C$ if $\alpha < \beta < \gamma$) positions in moderate equilibria under convex costs as well, but they move to positions at which the marginal cost equals the marginal benefit of moving—not exactly to their points of ideology.

To be able to state the second main theorem—a result on comparative statics—one last technical aspect remains to be addressed. Theorem 2 considers games with linear and bounded cost functions:

Assumption linearity. *Costs are of the form* $c(x) = \rho x$ *for all* $x \in [0,1]$*, where* $0 < \rho < 0.5$.

The upper bound on ρ ensures the existence of equilibria in which parties locate at extrema (for small values of q) and games that do not have equilibria (for high values of q). In other words, $\rho < 0.5$ precludes the existence of parameter values (α, β, γ) and ρ for which moving is strictly dominated, no matter the value of q.

As alluded to above, the share of antipartisans determines whether parties seek to locate as close to or as far away from the median as possible and, hence, the class of equilibrium that arises. Consequently, an increase in antipartisanship could result in the disruption of a moderate and the transition to an extreme equilibrium—as was observed in Brazil in 2018. Theorem 2 formally addresses this conjecture: It states that there is a non-empty set of combinations of parameter values for which antipartisanship can explain transitions between moderate and extreme equilibria. There are, on the other hand, also games in which antipartisanship cannot do so (cf. Theorem 3). In essence, antipartisanship can explain transition between moderate and extreme equilibria only if parties' ideologies are such that a moderate equilibrium exists for some value of q: each party must receive sufficiently many votes in such equilibrium; that is, their ideologies must be sufficiently spread over the spectrum. To simplify the statement of Theorem 2, consider the following definition.

Definition 2. *A combination of ideologies is referred to as* ***sufficiently spread*** *if*

$$\alpha < 1-\beta,\ \gamma < 3\beta \text{ and } \gamma > 1-\beta,$$

where $\beta-\alpha < \gamma-\beta$ *and* $\alpha \neq \beta$.

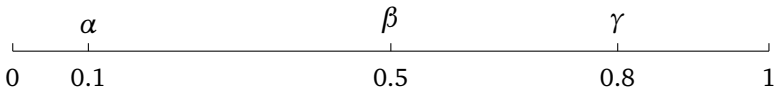

Figure 1.4: Sufficiently spread ideologies

Note that while Theorem 2 only considers sufficiently spread ideologies, the appendix contains a version of Theorem 2 that holds for all possible combinations of ideologies. As the main insight remains the same, Theorem 2 serves as a good and short representation of the more general Theorem 4.

Theorem 2. *Consider a combination of sufficiently spread ideologies* (α,β,γ)*. There are cost functions that satisfy linearity for which a change in the share of antipartisans* $1-q$ *by itself can explain transitions from no equilibrium to a moderate and a moderate to an extreme equilibrium if and only if*

$$\gamma-4\alpha+\beta>0 \text{ and } 4\gamma-2-\beta-\alpha>0.$$

Note that as there is no equilibrium-multiplicity (Theorem 1), a change between extreme and moderate positions can never be a mere "switching" between coexisting equilibria. Parties may relocate in such way only as a response to a change in parameter values, such as the share of antipartisanship $1-q$. For a closer discussion of this result, see Section 1.3.3.

To determine when and how antipartisanship can explain transitions between moderate and extreme equilibria, one needs to characterise the conditions on parameter values under which each of these classes of equilibria exist for some value of q. It turns out that there are values of q for which there is either no or an extreme equilibrium for *any* combination of ideologies and a cost function satisfying Assumption linearity: For very high values of q, the game is subject to the same forces as the "classic" Hotelling-Downs version and has no equilibrium—no matter the ideologies or the cost parameter. For small values of q, the presence of antipartisans makes parties move as far away from the median of the spectrum as possible—again, for any combination of parameter values. Hence, to establish the theorem, it suffices to characterise the set of parameter values (α, β, γ) that allow for a value

of q for which a moderate, namely the ideology-faithful equilibrium exists for some cost function satisfying linearity.[8] Doing so, one obtains the two constraints on ideology values stated in Theorem 2 ($\gamma - 4\alpha + \beta > 0$ and $4\gamma - 2 - \beta - \alpha > 0$) that, if satisfied, rule out deviations by the median party B to positions $\lim_{\varepsilon \to 0^+} \alpha - \varepsilon$ and $\lim_{\varepsilon \to 0^+} \gamma + \varepsilon$ in an ideology-faithful equilibrium.

Interestingly, the holding of party B's constraints is both necessary and sufficient for the existence of an ideology-faithful equilibrium when ideologies are sufficiently spread: As the median, B is the party that needs to overcome the smallest distance to reach other parties. Additionally, it is the only party that never receives any votes by antipartisans, which can be changed by deviating beyond one of its opponents' positions (i.e. $\alpha - \varepsilon$ and $\gamma + \varepsilon$). Such movement changes party B's antipartisan vote share discontinuously—a potentially very tempting deviation. Party C, on the other hand, has a relatively strong position as it is the party voted for by B's antipartisans (as $\beta - \alpha < \gamma - \beta$); hence, it is not tempted to deviate. Deviations by party A are accounted for by the lower bounds on β and γ: the distance to be overcome is too long and A remains on its point of ideology. To sum up, if ideologies are sufficiently spread, antipartisanship can explain transitions from an extreme to a moderate equilibrium if and only if party B's best response to positions $p_A = \alpha$ and $p_C = \gamma$ is $p_B = \beta$.

To illustrate, consider Figure 1.5, which depicts the combinations of parameter values that satisfy constraints stated in Theorem 2. Note that the actual space of possible values is twice as large as the theorem assumes $\beta - \alpha < \gamma - \beta$.[9] The main economic insight in Figure 1.5 is the following: The distance between parties' ideologies has to be of sufficient size such that each of them receives an adequate vote share in the ideology-faithful equilibrium. This prevents deviations, in particular by party B, towards the respective opponents' positions. Not only is it important for party B to receive a sufficiently large share of votes staying in between the other two parties, but also does B have to be sufficiently far away from both A and C to make it costlier to move beyond their positions. As depicted in Figure 1.5, this results in a space that allows for adequate distance between all parties.

[8] Knife-edge alliance equilibria are generally not possible in games with sufficiently spread ideologies, as the distance between parties is too large for a shared position to be optimal.

[9] Assuming distances between ideologies to not be equal is without loss of generality ($\alpha < \beta < \gamma$ and $\beta - \alpha < \gamma - \beta$): Lemmas 1, 2 and the results obtained in the proof of Lemma 4 rule out the existence of ideology-faithful equilibria in games that do not satisfy these requirements, making them of no interest for the derivation of Theorem 2.

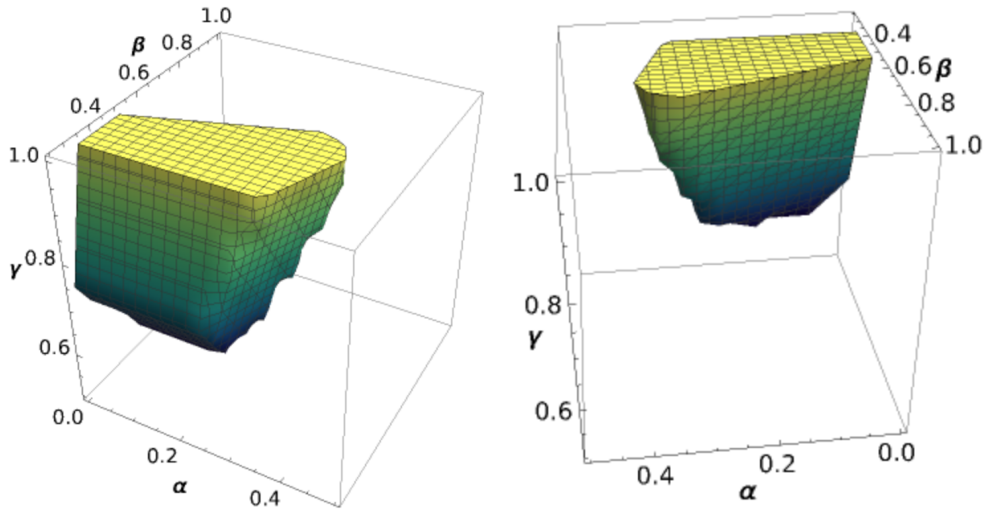

Figure 1.5: Parameter combinations described in Theorem 2

Referring back to the motivation of this paper, Theorem 2 states: were one to find that Brazil's parties incur costs proportional to changes they make to their political agenda (Assumption linearity) and their ideologies are sufficiently spread over the political spectrum, the rise and success of Bolsonaro's party could indeed be explained by antipartisanship.

So far, this section only provided a very rough understanding of underlying forces, intuition and the origin of the paper's main results. The following section will be dedicated to answering the "how" and "why" behind each of them: Why does the game only allow for such narrow set of equilibria? Why are equal positions or positions on the spectrum that are not equal to the points of ideology not an equilibrium? Why are knife-edge alliance equilibria the only equilibria in which two parties share a position on the spectrum and what makes them unstable? Under which conditions can antipartisanship *not* explain transitions between equilibria and why?

1.3.3 Driving forces and deeper analysis

Characterisation of equilibria

Theorem 1 states that equilibria can only be ideology-faithful, knife-edge alliance or extreme equilibria. In what follows, I discuss why any other combination of positions is not stable.

A natural candidate for an equilibrium would be one in which all parties locate at

the median of the spectrum. Here, the average distance to voters is minimised and high vote shares could be more probable. Despite being a, seemingly, reasonable candidate for an equilibrium, such combination of positions is never optimal. In fact, Lemma 1 states that in equilibrium, three parties can never all locate at the same (and not extreme) position. As both the spectrum and the cost function are continuous, any infinitesimal deviation comes at virtually no cost and makes the deviating party more attractive to both (some) partisans and antipartisans. This constitutes a profitable deviation and precludes the existence of equilibria in which all parties share a position.

Lemma 1. *It is never optimal for all parties to locate at the same position that is neither* 0 *nor* 1.

Similarly, equal distances between parties' position are virtually never optimal.[10] In the presence of antipartisans, it is always better to be a little more extreme than at least one of the opponents, as this increases votes by antipartisans discontinuously. This also rules out equal distances between parties. When both parties are located equally far away from the median party, they share the median party's antipartisans' votes. An infinitesimal move towards the extremum by one of the extreme parties makes it the best candidate for the median's antipartisans and is always profitable:

Lemma 2. *Suppose* $(\alpha, \beta, \gamma) \neq (0, 0.5, 1)$ *and an equilibrium exists in which parties locate at distinct positions. Let (w.l.o.g.)* $p_A < p_B < p_C$. *Then,* $p_B - p_A \neq p_C - p_B$.

The next result, Lemma 3, may be surprising: it states that the central party always locates at its point of ideology. This stems from the simple fact that moving within the bounds of its two opponents' positions (below (p_A, p_C)) does not change the central party's payoff as it never receives votes by antipartisans. Consequently, a central party would either (1) seek to move as close as possible to one of its opponents (if its ideology does not lie within their positions) or (2) remain on its point of ideology (if its ideology lies within their positions). As the action space is continuous, (1) cannot be an equilibrium; hence, central parties locate on their points of ideology:

Lemma 3. *The central party's position is equal to its point of ideology in all equilibria in which parties locate at distinct positions.*

Beyond its relevance for characterisations of equilibria, Lemma 3 provides a very interesting general insight: in the presence of antipartisans, the central party

[10] Only if $(\alpha, \beta, \gamma) = (0, 0.5, 1)$, they might constitute an equilibrium, cf. Theorem 4.

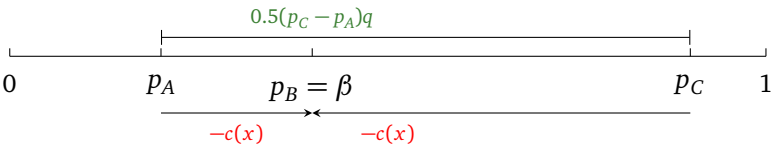

Figure 1.6: Lemma 3; the central party chooses its point of ideology.

never changes its political agenda to gain more votes. On pages 17 to 18, I discuss that and why non-central parties, on the other hand, do engage in such strategic movements along the political spectrum.

The next result discusses the already mentioned unstable knife-edge alliance equilibria. Recall, a knife-edge alliance equilibrium is an equilibrium in which either A and B or B and C share a position on the spectrum (see Definition 1).

Lemma 4. *Suppose the cost function is convex and continuously differentiable. Knife-edge alliance equilibria are the only equilibria in which two parties share a position that is neither* 0 *nor* 1. *They break down as soon as any parameter value changes infinitesimally.*

Lemma 4 states that parties may only share positions on the spectrum in a knife-edge alliance equilibrium. Thus, parties A and C can never choose the same non-extreme position: Alliances are generally only possible if the share of antipartisans is relatively small ($1-q < 0.5$), as high shares of antipartisans encourage differentiation, not convergence. This implies that if A and C share position $p_A \neq p_B$, (1) B must choose a position p_B closer to p_A than β ($\mid p_B - p_A \mid \leq \mid \beta - p_A \mid$). Moreover, (2) A and C must both have moved in the same direction ($\alpha \leq p_A \iff \gamma \leq p_A$), as p_A can only be optimal if it perfectly balances the deviation pay-offs of moving infinitesimally towards 0 or 1 for both parties. This is only possible if their respective ideologies are either both smaller or higher than p_A. As $\alpha \leq \beta \leq \gamma$, (1) and (2) are not compatible and an alliance between A and C is never optimal. Furthermore, alliances between three parties (at a position that is neither 0 nor 1) are precluded by Lemma 1. Hence, only in knife-edge alliance equilibria, two parties may share a non-extreme position.

Besides that, Lemma 4 makes a statement about the stability of such equilibria: A knife-edge alliance between two parties can only be stable if a number of requirements are fulfilled. Their ideologies cannot be too moderate, as otherwise a deviation towards an extremum would be profitable—the deviating party would receive all votes by partisans that are extremer than the initially shared position. The single party, on the other hand, must be located at a position that allows for

sufficient space between itself and its opponents to make sure a comparatively large share of partisans vote for the alliance. Furthermore, the share of antipartisans has to be high enough to encourage a shared position but small enough for the single party to not be tempted to move towards its opponents. These requirements can only be met if the alliance is located at one exact position—determined by parameter values and the single party's position—at which all above forces are balanced as if on a knife edge. Consequently, the equilibrium breaks down as soon as either one of the parameter values or a position of one player changes. To summarise, alliances can only be formed between two parties with similar ideologies and are relatively unstable.[11]

A plethora of combinations of positions ruled out as candidates for equilibria, Theorem 1 follows almost immediately. Recall, Theorem 1 characterises all equilibria of games with convex costs and shows that, given a value of q, equilibria in which parties locate at the extrema and ones in which they choose positions on the spectrum cannot exist simultaneously. Given the above results, the proof is rather simple: it shows that there cannot be an equilibrium (hereinafter *other equilibrium*) that is neither extreme nor a knife-edge alliance or ideology-faithful equilibrium. To illustrate, consider a game with linear costs: Lemmas 1, 2 and 4 imply that the other equilibrium would have to be (1) one in which at least one party does not locate at its point of ideology, (2) all parties choose different positions, and (3) the distance between the parties is not the same. This is not possible: Roughly speaking, if one party deviates from its ideology, all others do so as well, as the marginal cost of movement is equal for all of them. This, however, either requires very small shares of antipartisans, which preclude the existence of such other equilibrium (parties seek minimum differentiation, Lemma 6) or rather high shares of antipartisans, which rule out all equilibria but extreme ones (parties seek maximum differentiation, Lemma 7). Accordingly, the forces pushing parties towards the extrema and the median, respectively, can only be balanced if, for all parties, moving is relatively costly compared to any possible gain. This, in turn, results in the ideology-faithful equilibrium in which parties locate at their points of ideology.[12] Consequently, there

[11] As the proof of Theorem 2 suggests, extreme and ideology-faithful equilibria, on the other hand, are robust to small changes in parameter values.

[12] When costs are strictly convex, parties A and C move to the position at which the marginal cost of moving is equal to the marginal increase in votes, which, in turn, depends on q. Hence, many combinations of positions can constitute an ideology-faithful equilibrium under convex costs; when costs of moving towards the extrema (the center) are small compared to the increase in antipartisan (partisan) votes, parties may choose to move further away from their points of ideology than when they are relatively high. Put differently, games with strictly convex costs allow for "smoother" transitions between extreme, moderate and no equilibrium as a response to changes in the share of antipartisanship.

is no such other equilibrium.

Classes of equilibria

An important question that remains to be answered is whether equilibria of different classes can exist simultaneously. Were I to find that an extreme and a moderate equilibrium may coexist for some q, changes between them could be mere transitions between coexisting equilibria and not necessarily attributable to fundamental changes in the environment. Theorem 1 states that this is not the case if costs are convex. This finding is rather intuitive: Extreme equilibria can only occur if the share of antipartisanship is relatively high (antipartisans encourage maximum differentiation). Ideology-faithful equilibria, on the other hand, arise when the shares of antipartisans and partisans are moderate and, hence, neither minimum nor maximum differentiation are dominant strategies. Lastly, knife-edge alliance equilibria occur when the share of antipartisans is lower than 0.5 but not "too low": in such equilibrium, there are enough antipartisans for the single party to not want to join the alliance but not enough to prevent the alliance partners from choosing the same position. Put differently, one could think of the class of equilibrium as a function of the share of antipartisans (cf. Figure 1.7). Note that other than Figure 1.7 might suggest at first glance, there are parameter combinations (α, β, γ) and ρ that do not allow for a knife-edge alliance or an ideology-faithful equilibrium (Theorem 3), no matter the share of antipartisans. For a closer discussion of this finding, see Section 1.3.3 and Theorem 3.

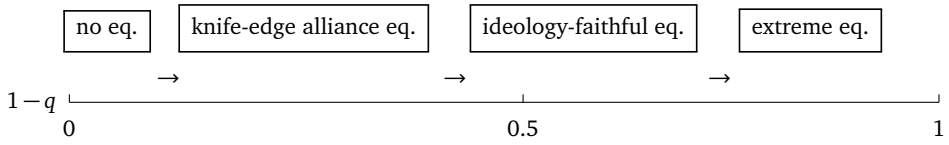

Figure 1.7: Classes of equilibria and the share of antipartisans $1-q$

Hence, comparative statics as those observed in Brazil in 2018 *have* to be the result of a change in the conditions the parties operate under—for instance, an increase in antipartisanship.

Comparative statics

Having discussed Theorem 1 and its origin, I pivot to a closer analysis of the results on comparative statics. Recall, Theorem 2 states under which conditions transitions between a moderate and an extreme equilibrium can be explained by changes in antipartisanship and, hence, answers this paper's main question. In some sense,

Theorem 2 characterises conditions under which the phenomena observed in Brazil in 2018 could be explained by the rise of antipartisans. Being positive, it only provides limited insights about the opposite case: when is antipartisanship *not* able to explain transitions—and if so, why?

As discussed in Section 1.3.2, Theorem 2 describes a parameter space for which moderate, extreme and no equilibrium are possible outcomes for some value of q and a cost function satisfying linearity. Due to party B's weak position in ideology-faithful equilibria, the space is implied by two constraints that ensure there being enough distance between parties A and C (enough votes for B in an ideology-faithful equilibrium) and B and A/C (deviations by B beyond A's or C's position too costly), respectively. Consequently, the existence of parameter values for which moderate equilibria do not exist, no matter the value of $1-q$, is not surprising. Clearly, for those parameter values, changes in antipartisanship cannot explain transitions between moderate and extreme equilibria:

Theorem 3. *Suppose linearity and ideologies are such that*

$$\gamma \leq 0.5 \text{ and } \gamma \leq 3\beta \text{ or}$$
$$\alpha \geq 0.5 \text{ and } \alpha \geq 3\beta - 2.$$

Transitions between moderate and extreme equilibria cannot be explained by a change in the share of antipartisanship $1-q$.

Figure 1.8: Ideologies that do not allow for moderate equilibria

As stated above, B does not receive any votes by antipartisans; thus, moderate equilibria can only exist if there is sufficient space between the extreme parties. Accordingly, if $\gamma \leq 0.5$[13] and $(p_A, p_B, p_C) = (\alpha, \beta, \gamma)$, party B does not receive enough votes by partisans located between α and γ[14] and party C's vote share is so high (all partisans to its right vote for party C), B is even more tempted to deviate towards $\lim_{\varepsilon \to 0^+} \gamma + \varepsilon$.

However, one may think that there are knife-edge alliance equilibria, which make up for the absence of the ideology-faithful equilibrium. Those do not exist for relatively small values of γ (i.e. $\gamma \leq 3\beta$) either, as an alliance between party A and

[13] The arguments for the case in which $\alpha \geq 0.5$ and $\alpha \geq 3\beta - 2$ are analogous.

[14] In fact, these are the only votes B receives in the ideology-faithful equilibrium.

B would be too weak (not enough votes to share) and one between C and B too strong (any alliance partner would be tempted to move towards 1 to receive all votes by partisans to its right). Hence, for the parameter space described in Theorem 3, non-extreme equilibria do not exist and changes in antipartisanship alone cannot explain comparative statics as those observed in Brazil in 2018.

1.4 Related literature

The Hotelling-Downs model, an adaptation of Hotelling's "model of spatial competition" (Hotelling, 1929), provides a theoretical framework for political agenda-making in the presence of uncertainty about voters' ideologies. As shown by Osborne (1993), the classic version with more than two parties has no Nash equilibrium in pure strategies when the voters' distribution on the spectrum is unimodal. Many extensions alleviating this negative result have since been proposed, of which I mention the most relevant to this paper below.

Ronayne (2018) introduces non-strategic idealists who do not move for the purpose of gaining votes. He finds two of those idealist candidates along with an unlimited number of strategic players to be enough to allow for the existence of pure strategy equilibria. The model analysed above considers idealism as well—though a, so to say, moderate version of it. The parties in this paper have ideological concerns, they incur costs when deviating from their predetermined ideology. They are, however, strategic: that is, if the gain from deviations is sufficiently high, they may deviate—Ronayne (2018)'s idealist never move. Furthermore, in this model, there are no players that are *not* subject to such idealist concerns.

Models similar to that proposed in Calvert (1985) constitute a further related adaptation: the authors analyse the outcome of games in which candidates are (also) motivated by the policy-outcome of the election, no matter whether they win or not. Their model distinguishes between policy- and office-motivated candidates: policy-motivated candidates care about the policy implemented after the election, no matter who implements it, while office-motivated candidates only care about being elected, no matter the political agenda they might have to adapt to do so. Note that the parties analysed in the present paper resemble a mixture of the above: In my model, parties care about (1) being elected and (2) the policy they, themselves, claim to implement. Importantly, a losing party does not care about the policies implemented by the winner of the election. Hence, their policy motivation could be understood as more selfish than that of parties analysed in Calvert (1985). The

convergence result obtained in the standard model is robust towards the introduction of policy motivation à la Calvert (1985) (cf. Duggan and Fey, 2005), while the present model allows for other types of equilibria.

Cahan and Slinko (2018) analyse the outcome of an election subject to a "best-worst voting rule": voters on a Hotelling line are able to submit both a positive and a negative vote; they can vote *for* their favourite and *against* their least favourite candidate. The votes are weighted, such that they do not necessarily cancel one another out. The authors find that best-worst voting allows for non-convergent equilibria. Note that I consider a different voting mechanism than Cahan and Slinko (2018) do (plurality voting): in my model, voters can submit only one ("positive") vote.

To conclude, I would like to mention related analyses that do not pertain to economic research:
In political science, negative partisanship has been the subject of analysis for some time now (one of the earliest works being Maggiotto and Piereson, 1977). Arzheimer and Berning (2019), for instance, analyse the rise of the "Alternative for Germany" (AfD)—also referred to as "party of protest"—in Germany in 2015. As a response to an increase in negative attitudes towards immigration, likely triggered by the European refugee crisis, the previously median-right AfD adopted an extreme-right platform opposing the government's left-leaning and welcoming immigration policy—a, as the authors argue, successful strategic move. Using survey data, Arzheimer and Berning (2019) provide evidence for negative attitudes towards immigration being one of the main drivers of the AfD's sudden success. Empirical in nature, the methods used in this and other work on the matter in political science are, however, rather distinct from classic game theoretical analyses—creating a gap in the literature I seek to fill: to the best of my knowledge, this is the first paper introducing explicit negative preferences to a plurality voting model in economic research.

1.5 Conclusion and outlook

In this paper, I introduce antipartisans to the Hotelling-Downs model with three parties: antipartisans vote for the candidate located furthest away from their most disliked candidate on the political spectrum. The presence of antipartisans gives rise to three classes of equilibria: (1) for comparatively high shares of antipartisanship, parties locate at the extrema of the spectrum, (2) for moderate shares, they locate at moderate positions (two parties share a position or none do) and (3) for low shares, there is no equilibrium.

Besides allowing for the existence of pure strategy Nash-Equilibria, which the standard Hotelling-Downs model does not, the present adaptation is able to explain movements across the political spectrum by changes in antipartisanship. Phenomena such as the sudden success of Jair Bolsonaro's originally median-right wing party after a relocation towards the extreme-right could be rooted in an increase in antipartisanship. In particular, this paper answers the question under which conditions such transitions between equilibria can, ceteris paribus, be explained by antipartisan voting behaviour.

While my analysis suggests *that* and *how* changes in the share of antipartisanship can explain movements along the spectrum, it remains agnostic about the *why* of such change: why should the the number of voters with negative feelings towards political stances change in the first place? One possible source of such change could be a shift in the relative importance of different political matters. One could imagine there being questions of political nature relatively more voters do not have positive but rather negative preferences about: a voter may not know which social security system he prefers over others but only which foreign policy regulations he generally *dislikes*. In Germany in 2015, the refugee crisis could have increased the relevance of foreign policy regulations that comparatively many voters had negative rather than positive preferences about (Arzheimer and Berning, 2019); these voters only knew how they did *not* want the high number of refugees to be dealt with. In Brazil, it might have been the economic downturn and inflation combined with increased public spending that created general unrest and the want for certain policies explicitly *not* to be maintained or adopted (Davis and Straubhaar, 2020). A satisfying answer backed with theoretical or empirical arguments as to why the share in antipartisanship should change goes beyond the scope of this project and is left for future research.

Appendix 1.A

Proofs and omitted results

Remark: Despite my considering only strict Nash equilibria, in some cases, I need to allow for weak inequalities when ruling out deviations after applying limits. For instance, when the deviation payoff contains a term that is strictly smaller than but converges to 0, the deviation payoff can (dropping the limit term) be equal to the non-deviation payoff. This results in weak inequalities instead of strict ones.
Remark: To enhance readability, I often omit intermediate steps in which I apply limits on the cost function. Consider the following equivalences that hold due to c's continuity: $\lim_{\varepsilon\to 0^+} c(p_A+\varepsilon-\alpha) = c(p_A+\lim_{\varepsilon\to 0^+}\varepsilon-\alpha) = c(p_A-\alpha)$. Instead of stating intermediate steps, I often just write $c(p_A-\alpha)$.
Proof of Lemma 1. Towards contradiction, suppose there is an equilibrium in which all parties locate at some position $p \in (0,1)$. For such position, their expected vote share is equal to $1/3$.[15] Moving infinitesimally apart from its competitors, one of them can change its (expected) vote share to be equal to

$$\lim_{\varepsilon\to 0^+} 0.5(2p-\varepsilon)q + \big(1-0.5(2p-\varepsilon)\big)(1-q) = pq + (1-p)(1-q) \qquad \text{or}$$
$$\lim_{\varepsilon\to 0^+} 0.5(2p+\varepsilon)(1-q) + \big(1-0.5(2p+\varepsilon)\big)q = p(1-q) + (1-p)q,$$

depending on which end of the spectrum it decided to move towards. Those deviations cannot both yield an expected vote share of less than $1/3$—hence, at least one of them constitutes a profitable deviation. As the cost function is assumed to be continuous, the cost of such infinitesimal deviation can be disregarded when comparing the payoffs and the deviation remains profitable. Accordingly, there exists no such equilibrium. □
Proof of Lemma 2. If $p_B - p_A = p_C - p_B$, one party is tempted to move slightly towards the closer extremum, making sure to gain all antipartisan-votes against the other non-central party:
Consider party A. As I assume costs to be continuous, the following needs to hold for $p_B - p_A = p_C - p_B$ to be an equilibrium:

$$0.5(p_A+p_B)q + \big(1-0.5(p_C+p_B)\big)(1-q) + 0.25(p_B-p_A)(1-q) - c(p_A-\alpha) \geq$$
$$\lim_{\varepsilon\to 0^+} 0.5(p_A-\varepsilon+p_B)q + \Big(1-0.5(p_A-\varepsilon+p_B)\Big)(1-q) - c(p_A-\varepsilon-\alpha) \Rightarrow$$
$$p_B - p_C > 0, \text{ contradicting } p_C > p_B > p_A.$$

[15] This holds for all such equal positions p.

Hence, in equilibrium, distances between parties cannot be equal. □

Proof of Lemma 3. Say (w.l.o.g.), the central party is B. Its ideology can neither be smaller nor greater than its position, as any movement within the bounds of p_A and p_C towards the ideology does not change its vote share but decrease the cost it incurs. Note that this holds even for the case in which the party's ideology is more extreme than p_A or p_C: Any movement towards the other parties' positions within the bounds of p_A and p_C would be an improvement. As the action space is continuous, B would never be able to settle on one such very close position.

Were party B to move to a position that does not lie within $[p_A, p_C]$, there would be a new central party subject to the very same forces. Hence, when parties choose distinct positions, the central party has to choose a position equal to its point of ideology. □

Remark: Below, I make use of the partial derivative of the cost function with respect to a party's position, for instance $\frac{\partial c(|p_A-\alpha|)}{\partial p_A}$. I, hereafter, refer to it as $c'(|\, p_A-\alpha\,|)$.

Proof of Lemma 4. Below, I derive necessary conditions for equilibria in which two parties locate at the same (interior) position and show that they imply any such equilibrium to be a knife-edge alliance equilibrium.

For it to be optimal for two parties to take the same position, infinitesimal but profitable deviations towards or away from the third party must be impossible. Suppose $p_A = p_B < p_C$ and consider party A. As costs are continuous, the following needs to hold:

$$
\begin{aligned}
&0.5\Big(0.5(p_A+p_C)q+\big(1-0.5(p_A+p_C)\big)(1-q)\Big)-c(|p_A-\alpha|)> \\
&\lim_{\varepsilon\to 0^+} 0.5(2p_A-\varepsilon)q+\big(1-0.5(p_A+p_C)\big)(1-q)-c(|p_A-\varepsilon-\alpha|) \iff \\
&(1-2p_A)q > 1-0.5(p_A+p_C)-\lim_{\varepsilon\to 0^+}\varepsilon q+2\Big(c(|p_A-\alpha|)-c(|p_A-\varepsilon-\alpha|)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&0.5\Big(0.5(p_A+p_C)q+\big(1-0.5(p_A+p_C)\big)(1-q)\Big)-c(|p_A-\alpha|)> \\
&0.5(p_C-p_A)q-\lim_{\varepsilon\to 0^+} c(|p_A+\varepsilon-\alpha|) \iff \\
&(1-2p_A)q < 1-0.5(p_A+p_C)-\lim_{\varepsilon\to 0^+} 2\Big(c(|p_A-\alpha|)-c(|p_A+\varepsilon-\alpha|)\Big)
\end{aligned}
$$

The above system of inequalities can only have a solution if (1) $c(|p_A-\alpha|) < c(|p_A+\varepsilon-\alpha|)$ and (2) $0.5q > c(|p_A-\alpha|)-c(|p_A-\varepsilon-\alpha|)$ for any $\varepsilon > 0$. Hence, the parties sharing a position need to have ideology values that are weakly smaller than p_A *and* $c'(|p_A-\alpha|) < 0.5q \;\wedge c'(|p_A-\beta|) < 0.5q$. Besides that, the position they share has to be equal to $p_A = \frac{p_C+2q-2}{4q-1}$.

As party C must not be tempted to move to a position closer to p_A or closer to

the extremum, its marginal cost of movement has to satisfy $c'(|p_C - \gamma|) \geq q - 0.5$ and $c'(|p_C - \gamma|) \geq 0.5 - q$. As $q > 0.5$ (this follows from $p_A = \frac{p_C+2q-2}{4q-1}$), for any combination of positions that are not extreme, $\gamma \geq p_C$ has to hold. Hence, the following requirements need to be met in any such equilibrium:

$$\begin{aligned}
&\alpha \leq p_A \text{ and } \beta \leq p_A,\\
&c'(|p_A - \alpha|) < 0.5q \text{ and } c'(|p_A - \beta|) < 0.5q,\\
&p_A = \frac{p_C + 2q - 2}{4q - 1} \text{ and } p_A < 1/3 p_C < 1/3,\\
&q > 0.5.
\end{aligned}$$

As p_A has to be exactly equal to $\frac{p_C+2q-2}{4q-1}$ and p_C is determined by parameter values, the equilibrium breaks down as a response to any change in any parameter value. Hence, the equilibrium is a knife-edge alliance equilibrium. The case in which two parties share a position greater than 2/3 (here, B and C) is analogous. □

Lemma 5. *$q \leq 0.5$ in any equilibrium in which parties locate at the extrema.*

Proof. In general, there could be two types of extreme equilibria. Either all extrema are or only one extremum is occupied. The first type of extreme equilibrium can only appear if all parties have an ideology of value greater than 0.5 (analogously for values smaller than 0.5).[16] If such equilibrium exists, parties must not be tempted to stay at their point of ideology. Considering party A and assuming $\alpha \neq 1$, this results in the following inequality:

$$\begin{aligned}
&1/3 - c(1-\alpha) > 0.5(1+\alpha)q + \big(1 - 0.5(1+\alpha)\big)(1-q) \iff\\
&1/3 - c(1-\alpha) > 1 - 0.5(1+\alpha) + \alpha q = 0.5 + \alpha(q - 0.5),
\end{aligned}$$

implying $q < 0.5$, as $c(1-\alpha) > 0$. If all parties have the same ideology, the following needs to hold for any $\varepsilon > 0$:

$$\begin{aligned}
&1/3 > 0.5\big(1 + (1-\varepsilon)\big)q + \big(1 - 0.5\big(1 + (1-\varepsilon)\big)\big)(1-q) - c(\varepsilon) \iff\\
&\varepsilon(q - 0.5) > 2/3 - c(\varepsilon),
\end{aligned}$$

which is not possible as the cost function is continuous and $c(0) = 0$. Hence, in such case, the extreme equilibrium in which all parties locate at the same extremum is not possible. In the case in which only two parties occupy the same extremum and at least one of them has an ideology that is not equal to the extremum, $q < 0.5$ as

[16] If this were not the case, at least one of the parties would be tempted to move to the unoccupied extremum, as this would save costs and yield a higher vote share.

well, as $0.25-c(1-\beta)>0.5q \Rightarrow q<0.5$. If the ideology is equal to the extremum, $q \leq 0.5$, as the following needs to hold for any $\varepsilon > 0$: $0.25 > 0.5q - c(\varepsilon) \iff 0.5(0.5-q) > -c(\varepsilon) \Rightarrow q \leq 0.5$. □

Lemma 6. *Suppose linearity,* $(\alpha,\beta,\gamma) \neq (0,0.5,1)$*,* $0.5q > \rho$ *and* $q > 0.5$*. Parties do not choose distinct positions. Extreme positions are not possible either.*

Proof. Lemmas 1 and 2 imply: any equilibrium with distinct positions has to be such that there is one extreme party that receives no votes by the central party's antipartisans. This party can increase its vote share by $0.5xq$ by moving x closer towards the central party (not beyond). The party incurs cost $x\rho$ for doing so. It does not lose any antipartisan-votes when doing so, as it does not change the distance between its two competitors.[17] If now $0.5q > \rho$ and the party's ideology value is smaller than the initial and new positions, this is a profitable deviation. If the movement reduces the costs (ideology closer to new position than old one), it is an even more profitable deviation. Thereby, there is no equilibrium with distinct positions.

An extreme equilibrium is not possible either, as the two parties sharing a position always have incentives to move towards the middle: Those two parties both receive a payoff of 0.25 less their respective costs of moving. Whenever one of them moves slightly towards the middle, thereby almost not changing its cost-term, it can make sure to receive a payoff of $0.5q$ less the slightly, almost not measurably, different cost.[18] This is a profitable deviation, as $0.5q > 0.25$ ($q > 0.5$ by assumption). An extreme equilibrium in which all parties share a position is not possible either, as an infinitesimal move towards the center would come at almost no cost and increase the vote share to be equal to $q > 0.5 > 1/3$. □

Lemma 7. *Suppose the cost function is convex and continuously differentiable,* $c'(x) < 0.5-q \ \forall x \in (0,1]$ *and an equilibrium exists. Parties locate at the extrema.*

Proof. Towards contradiction, suppose $c'(x) < 0.5-q \ \forall x \in [0,1]$ and there is an equilibrium that is not extreme. By Lemmas 1, 2 and 4, this has to be an equilibrium in which all parties choose different positions and distances between them are different.

Let $p_A < p_B < p_C \ \wedge \ p_B - p_A < p_C - p_B$. Consider party C. By moving x units towards 1, it looses $0.5xq$ votes by partisans and gains $0.5x(1-q)$ votes by antipartisans. Hence, the marginal benefit of moving towards the extremum is equal to $0.5-q$. As the cost function is convex, for party C's position to be stable, the

[17] It moves closer to the central party whose antipartisans never vote for it anyway.
[18] The cost could be higher or lower, depending on the ideology.

marginal cost of moving towards p_C has to be equal to $0.5-q$ if $p_C < 1$. This is not possible by assumption. Thereby, party C always has an incentive to move towards the extremum if $p_A \neq p_B \neq p_C$. If $p_C = 1$, party B has an incentive to move to $\lim_{\varepsilon \to 0^+} p_A - \varepsilon$: To show this, I use that the median party always locates at its point of ideology (by Lemma 3); hence:

$$
\begin{aligned}
&p_A q + (1-0.5(1+p_A))(1-q) - c(\beta - p_A) \geq 0.5(1-p_A)q \iff \\
&p_A q + (1-p_A)(0.5-q) \geq c(\beta - p_A) \text{ which is true, as:} \\
&p_A q + (1-p_A)(0.5-q) \geq (1-p_A)(0.5-q) \geq (\beta - p_A)(0.5-q) \\
&> (\beta - p_A)c'(\beta - p_A) \geq c(\beta - p_A)
\end{aligned}
$$

An equilibrium in which $p_B < p_A = \alpha < p_C$ is not possible either as now party A would be tempted to move beyond p_B.

Furthermore, no knife-edge alliance equilibrium exists: $q < 0.5$ by assumption, which violates the necessary condition obtained in the proof of Lemma 4; hence, only extreme equilibria are possible. □

Proof of Theorem 1. The proof is divided into three steps. First, I show that there are no other equilibria. Thereafter, I show that, given linear costs, parties locate at their points of ideology in any ideology-faithful equilibrium. In the last step, I show uniqueness.

No other equilibria:

Suppose there is an equilibrium that is neither a knife-edge alliance nor an extreme equilibrium. Below, I refer to this as the *other equilibrium*. Lemmas 1, 2 and 4 imply: this equilibrium has to be one in which all parties choose different positions and the distance between the parties is not the same. Let (w.l.o.g.) $p_A < p_B < p_C \wedge p_B - p_A < p_C - p_B$.

By Lemma 3, party B has to be located at its point of ideology.

Consider now party A. Its point of ideology cannot be greater than p_A, as any movement of x units towards α, staying closer to 0 than p_B, yields an improvement: it is less costly and increases the vote share by $0.5xq$. A's ideology can only be smaller than p_A if $c'(p_A - \alpha) = 0.5q$: the marginal gains of moving towards party B (keeping $p_A < \beta$) are equal to $0.5q$; a local maximum can only be achieved if marginal costs are equal to gains and, hence, $c'(p_A - \alpha) = 0.5q$. Note that $c'(p_A - \alpha) < 0.5q$ is not possible if $p_A < p_B$: such position cannot be stable if $c'(p_A - \alpha) < 0.5q$, as A would move as close to p_B as possible (marginal benefit > marginal cost); as there is no such position that also satisfies $p_A < p_B$, I get that either $\alpha = p_A$ or $\alpha < p_A$ and $c'(p_A - \alpha) = 0.5q$.

Lastly, consider C. A similar line of reasoning applies here: By moving x units

towards 1, C gains $x(0.5-q)$ votes; moving towards the middle, it gains $x(q-0.5)$ votes. Marginal costs must be equal to gains if C chooses a position $p_B \neq p_C \neq 1$ $p_C \neq \gamma$; hence, in any other equilibrium, either $1 \neq p_C \neq \gamma$ and $c'(\mid p_C - \gamma \mid) = |q-0.5|$, $p_C = 1$ and $c'(\mid 1-\gamma \mid) < 0.5-q$ or $p_C = \gamma$ and $c'(x) > 0.5-q$ for all $x \in [0,1]$.

It remains to be shown that the other equilibrium is ideology-faithful; that is, $p_A < p_B < p_C$ if $\alpha < \beta < \gamma$ and relative political attitudes are sustained.

First note that the party closest to 0 must have an ideology weakly smaller than its position (as $p_A \geq \alpha$, which was shown above).[19] Thereby, as the median party must be located at its point of ideology and $\alpha \leq \beta \leq \gamma$, $p_B < p_A < p_C$ and $p_B < p_C < p_A$ are not possible; hence, A must be located closest to 0. Lastly, $p_A < p_C < p_B$ is not possible either: in this case, $\gamma = p_C$ by Lemma 3. Then $p_B > \beta$, $c'(p_B - \beta) \leq 0.5-q$ and $q < 0.5$. By assumption, $\beta \leq \gamma = p_C < p_B$ and the cost is increasing and convex, thus, it has to hold that

$$c(p_B-\gamma) \leq c(p_B-\beta) \leq c'(p_B-\beta)(p_B-\beta) \leq (0.5-q)(p_B-\beta). \tag{1.1}$$

Comparing party C's payoff upon choosing $p_C = \gamma$ to a deviation to $\lim_{\varepsilon \to 0} p_B + \varepsilon$, I find that such deviation is always profitable given the assumptions on parameters:

$$\begin{aligned}
&0.5(p_B-p_A)q \lesseqgtr (1-p_B)q + 0.5(p_B+p_A)(1-q) - (0.5-q)(p_B-\beta) \iff \\
&(p_B-1+\beta)q \lesseqgtr 0.5(p_A+\beta), \text{ now as } q<0.5 \text{ and } 0 \leq p_A < p_B \leq 1, \\
&(p_B-1+\beta)q < 0.5(p_B-1+\beta) \leq 0.5(p_A+\beta) \Rightarrow \\
&0.5(p_B-p_A)q < (1-p_B)q + 0.5(p_B+p_A)(1-q) - c(p_B-\gamma) \text{ by 1.1.}
\end{aligned}$$

Note that the sharing of a position on the spectrum by two parties is only possible in a knife-edge alliance equilibrium, as was shown in Lemma 4. Thereby, there are, if any, only knife-edge alliance, ideology-faithful and extreme equilibria when the cost function is convex and continuously differentiable.[20]

$p_A = \alpha$, $p_B = \beta$ and $p_C = \gamma$ in an ideology-faithful equilibrium if costs are linear: For this part of the proof, assume the cost function to be of the form $c(x) = \rho x$, $\rho > 0$. As argued above, $\alpha < p_A$ only if $c'(p_A - \alpha) = \rho = 0.5q$. As $\rho = 0.5q$ implies p_A to be a weak best response, this does not constitute an equilibrium (I consider only strict Nash equilibria). Hence, $\alpha = p_A$. The argument for $p_C = \gamma$ is analogous. Note that $p_C = 1 > \gamma$ is not possible in an ideology-faithful equilibrium by Lemma 7. $p_B = \beta$ follows from Lemma 3.

[19] The case in which $p_B - p_A \geq p_C - p_B$ is analogous.
[20] Cf. footnote 19.

Uniqueness:
This part of the proof is divided into four steps. First, I show that ideology-faithful and extreme equilibria do not coexist. Thereafter I argue that the same holds for knife-edge alliance and extreme equilibria. In the third step, I rule out coexistence of ideology-faithful and knife-edge alliance equilibria. To conclude, I show that two equilibria of the same class cannot coexist either.
Step 1: A necessary condition for the existence of an ideology-faithful equilibrium is $\alpha < \beta < \gamma$, as was argued in the first part of this proof.
There are two types of extreme equilibria: all parties share a position or only two do so. Consider the case in which all parties share a position. For such equilibrium to be possible, the following constraints need to hold for some ideology of value $x \in (0,1)$:[21]

$$\frac{1}{3} - c(x) > \frac{x}{2}(1-q) + (1-\frac{x}{2})q \iff \frac{1}{3} - q - x(\frac{1}{2} - q) > c(x) \tag{1.2}$$

$$\frac{1}{3} > q \tag{1.3}$$

$$\frac{1}{3} - c(x) > \frac{1}{2} - c(1-x) \Rightarrow x < \frac{1}{2}, \tag{1.4}$$

where the first inequality rules out a deviation of some party with ideology x to its point of ideology in such extreme equilibrium. The second inequality accounts for an infinitesimal deviation towards the center of the spectrum. The third inequality considers deviations to the other extremum (here 1). Consider the case in which $\alpha \neq 0$. For the ideology-faithful equilibrium to be possible, the median party cannot be tempted to move beyond $p_A = \alpha$. At such position, it would receive at least all votes by antipartisans of party C, which must have an ideology smaller that $1/2$; hence, the deviation payoff is bounded below by $1/2(1-q) - c(x)$. The payoff B receives at $p_B = \beta$ is smaller that $1/4q$. Thus, it must hold that $c(x) > 1/2 - 3/4q$, which is not compatible with Line 1.2, as $1/2 - 3/4q > 1/3 - q > 1/3 - q - x(0.5 - q)$. Consider the case in which $\alpha = 0$. If $\alpha = 0$, a deviation to α by party B in the ideology-faithful equilibrium needs to be ruled out. As was established above, party B's payoff in the ideology faithful equilibrium is bounded above by $1/4q$. A deviation to α yields a payoff of at least $0.5\big(1/4q + 3/4(1-q)\big) - c(x)$. The former being greater than the latter is not compatible with Line 1.2 either. Thus, ideology-faithful and extreme equilibria in which only one extremum is occupied are not compatible.
The same holds for ideology-faithful and extreme equilibria in which all extrema are occupied: Suppose A and B occupy 0. A necessary condition for the existence

[21] The case in which all parties locate at 1 is analogous.

of such equilibrium is $\beta < 0.5$, as B would be better off sharing position with C if $\beta \geq 0.5$. As $\alpha \leq \beta \leq \gamma$, party B never occupies an extremum alone in such equilibrium. Furthermore, party B must not be tempted to remain on its point of ideology. This implies $0.25 - \min\{c(\beta), c(1-\beta)\} > 0.5q$. Using this finding, the following inequalities show that, given $0 < \alpha < 0.5$ and B does not want to deviate to $\lim_{\varepsilon \to 0^+} \alpha - \varepsilon$ in the ideology-faithful equilibrium, the extreme equilibrium is not possible:

$$
\begin{aligned}
&0.5(\gamma-\alpha)q \geq \alpha q + \big(1-0.5(\gamma+\alpha)\big)(1-q) - c(\beta-\alpha) \Rightarrow \\
&0.5(\gamma-\alpha)q \geq \alpha q + \big(1-0.5(\gamma+\alpha)\big)(1-q) - c(\beta) \iff \\
&q(1-2\alpha) - 1 + 0.5(\gamma+\alpha) \geq -c(\beta) > 0.5q - 0.25 \iff \\
&q(0.5-2\alpha) > 0.75 - 0.5(\gamma+\alpha) \Rightarrow 0.5(0.5-2\alpha) > 0.75 - 0.5(\gamma+\alpha) \iff \\
&0.5\gamma > 0.5 + 0.5\alpha \iff \gamma > 1 + \alpha,
\end{aligned}
$$

contradicting $\gamma \leq 1$. Note that $q \leq 0.5$ follows from Lemma 5. If $\alpha = 0$, the following must hold:

$$
\begin{aligned}
&0.5\gamma q > 0.5\big(0.5\gamma q + (1-0.5\gamma)(1-q)\big) - c(\beta) \iff \\
&0.25\gamma q > 0.5(1-q)(1-0.5\gamma) - c(\beta) > 0.5(1-q)(1-0.5\gamma) + 0.5q - 0.25 = \\
&0.5 - 0.25\gamma - 0.5q + 0.25\gamma q + 0.5q - 0.25 \iff 0 > 0.25 - 0.25\gamma,
\end{aligned}
$$

contradicting $\gamma \leq 1$. The case in which $\beta > 0.5$ is analogous (B wants to deviate to γ in this case). Hence, extreme and ideology-faithful equilibria cannot coexist, as was to be shown.

Step 2: The same holds for knife-edge alliance and extreme equilibria: The proof of Lemma 4 states that $q > 0.5$ in knife-edge alliance equilibria. In extreme equilibria $q \leq 0.5$ by Lemma 5, making them incompatible.

Step 3: Knife-edge alliance and ideology-faithful equilibria cannot coexist either, as, by the proof of Lemma 4, in any knife-edge alliance equilibrium, $c'(\mid p_A - \alpha \mid) < 0.5q$. As $\alpha \leq \beta \leq p_A$, this would imply party A being tempted to move towards party B's point of ideology[22] in any ideology-faithful equilibrium: any such movement increases votes marginally by $0.5q$ and comes at marginal costs that are weakly smaller than $c'(\mid p_A - \alpha \mid)$ (as $p_A \geq \beta$) and is, hence, profitable. The case in which party B and C share a position in a knife-edge alliance equilibrium is analogous.

Step 4: For a given parameter combination, there cannot be two ideology-faithful equilibria as, by the proof of Theorem 1, extreme parties locate at positions at which

[22] By Lemma 3, the median party chooses its point of ideology in any equilibrium with distinct positions.

the marginal cost of movement equals the marginal increase in vote shares. Additionally, $p_A \leq p_B = \beta \leq p_C$. Given a set of parameter values there is only one such combination of positions and, hence, only one ideology-faithful equilibrium. The same holds for knife-edge alliance equilibria: the proof of Lemma 4 shows that, given a combination of parameter values, there is only one stable combination of positions that allows for such equilibrium. Lastly, extreme equilibria do not coexist either: Consider an extreme equilibrium in which all extrema are occupied. The two parties sharing an extreme position cannot be indifferent between either of the extrema as only strict Nash equilibria are considered, hence two extreme equilibria in which all extrema are occupied cannot coexist. An extreme equilibrium in which all parties locate at the same extremum (1) cannot coexist with one in which all extrema are occupied (2), as one of the parties (the party occupying one extremum alone) in (2) would be indifferent between two positions (again, no strict Nash equilibrium). Two different extreme equilibria in which all parties locate at the same extremum cannot coexist either, as a necessary condition for an extreme equilibrium in which $p_A = p_B = p_C = 0$ is their ideologies being smaller than $1/2$. This is due to the fact that a deviation to 1 would be profitable otherwise. An analogous argument holds for $p_A = p_B = p_C = 1$. □

Lemma 8. *Suppose linearity. There exists a share of antipartisans* $0 < q < 1$ *for which the equilibrium is extreme.*

Proof. In what follows, I show the existence of an extreme equilibrium in which two parties share a position: In such equilibrium, a deviation of one of the two parties sharing a position (say, w.l.o.g., A and B) towards the median needs to be ruled out. Consider the case in which both parties have an ideology smaller than (or equal to 0.5). In this case, both parties locate at 0.[23] Thus, for an extreme equilibrium to exist, the following needs to hold: $0.25-c(\alpha) \geq 0.25-c(\beta) > 0.5q \Rightarrow 0.5(0.5-q) > c(\beta) = \beta\rho$. As $\beta \leq 0.5$, $\rho < 0.5-q$ is sufficient for the above to hold.[24] In this case party, C is not tempted to deviate to another position $p \neq 1$, as $\rho < 0.5-q$: any movement away from a position $p \neq 1$ further away from the other parties would be profitable. Consequently, the best party C can do is choose $p_C = 1$. The case in which both have an ideology of value greater than 0.5 is analogous. As ρ is bounded above by 0.5, a value of $q > 0$ can always be found to satisfy $\rho < 0.5-q$ and an extreme equilibrium exists. □

[23] A deviation to 0 given a shared position at 1 yields the same vote share and come at less cost. Hence, a shared position at 1 is not possible if the parties sharing a position have ideologies smaller than 0.5.

[24] If both A and B have an ideology of 0, deviating from 0 is unprofitable as well if $\rho < 0.5-q$.

Lemma 9. *Suppose linearity. There is a share of antipartisans* $0 < q < 1$ *for which no equilibrium exists.*

Proof. Assume $(\alpha, \beta, \gamma) \neq (0, 0.5, 1)$. By Lemma 6 and the results obtained in the proof of Lemma 4, for values $0.5q \geq q - 0.5 > \rho$ and $q > 0.5$, there is no equilibrium. A q to satisfy this inequality can be found as long as $\rho < 0.5$, which is true by assumption. The same holds for the case in which $(\alpha, \beta, \gamma) = (0, 0.5, 1)$: By Theorem 1, distinct positions are only possible if parties locate at their points of ideology when costs satisfy linearity. Such combination of positions does not constitute an equilibrium either if $0.5q \geq q - 0.5 > \rho$ and $q > 0.5$, as party A, for instance, would be better off at $\lim_{\varepsilon \to 0^+} \beta - \varepsilon$: $0.25q + 0.5(1-q) < 0.5q + 0.25(1-q) - 0.5\rho \iff \rho < q - 0.5$. □

Lemma 10. *Suppose linearity.* $\alpha \leq 0.5$ *and* $\gamma \geq 0.5$ *if an ideology-faithful equilibrium exists.*

Proof. For the ideology-faithful equilibrium to exist, the central party (B) must not be tempted to move to $\lim_{\varepsilon \to 0^+} p_A - \varepsilon$; I refer to this condition as (B beyond α). A necessary condition for such movement to not be profitable is $\alpha < 0.5$:

$$\begin{aligned} &0.5(\gamma - \alpha)q \geq \alpha q + \big(1 - 0.5(\gamma + \alpha)\big)(1-q) - (\beta - \alpha)\rho \Rightarrow \\ &2(1-2\alpha)q \geq 2 - \gamma - \alpha - 2(\beta - \alpha)\rho \end{aligned}$$

Towards contradiction, suppose $\alpha \geq 1/2$ and (B beyond α) holds. As $\rho > 0$:

$$\begin{aligned} 0 \geq 2(1-2\alpha)q \geq 2 - \gamma - \alpha - 2(\beta - \alpha)\rho > 2 - \gamma - \alpha - (\beta - \alpha) \\ = 1 - \beta + 1 - \gamma > 0, \end{aligned}$$

as $\beta < 1$ and $\gamma \leq 1$, which leads to a contradiction. An analogous derivation can be made for γ implying $\gamma > 0.5$ being a necessary condition for the existence of an ideology-faithful equilibrium.[25] □

Proof of Theorem 2. Below, I derive the conditions under which an ideology-faithful equilibrium exists for some value of q and a cost function that satisfies linearity. The existence of values of q and ρ that allow for no and extreme equilibria is implied by Lemmas 9 and 8. Below, I refer to constraints implied by ideologies being sufficiently spread as Assumption (Spread). Note that, as was shown in the proof of Lemma 10, in any ideology-faithful equilibrium $\alpha < 0.5$ and $\gamma > 0.5$.

[25] If $\gamma \leq 0.5$, a deviation of B towards $\lim_{\varepsilon \to 0^+} \gamma + \varepsilon$ is always profitable.

To show under which conditions ideology-faithful equilibria are possible, I derive inequalities that rule out deviations from the parties' ideologies in such equilibrium. To do so, I make use of the below characterisation of expected vote shares:

Expected vote share

The expected vote share of party A, given fixed positions p_B and p_C, where $p_A < p_C$ takes the following form. Note that combinations of positions such as $p_A = p_B = p_C$ or $p_B - p_A = p_C - p_B$ are not depicted as they were ruled out by Lemmas 1 and 2. Cases in which $p_A \geq p_C$ are analogous.

$\mathbb{E}[U(p_A)] + c(\mid p_A - \alpha \mid) =$

$$\begin{cases} \frac{p_A+p_B}{2}q + \left(1 - \frac{p_A+p_B}{2}\right)(1-q) & \text{if } p_A < p_B \wedge p_B - p_A > p_C - p_B, \quad \text{(i)} \\ \frac{p_A+p_B}{2}q + \left(1 - \frac{p_C+p_B}{2}\right)(1-q) & \text{if } p_A < p_B \wedge p_B - p_A < p_C - p_B, \quad \text{(ii)} \\ \frac{p_C-p_B}{2}q & \text{if } p_B < p_A < p_C, \quad \text{(iii)} \\ 0.5\left(\frac{p_A+p_C}{2}q + \left(1 - \frac{p_A+p_C}{2}\right)(1-q)\right) & \text{if } p_A = p_B, \quad \text{(iv)} \\ 0.5\left(\left(1 - \frac{p_A+p_B}{2}\right)q + \frac{p_A+p_B}{2}(1-q)\right) & \text{if } p_A = p_C, \quad \text{(v)} \end{cases}$$

Suppose, A's position lies within (ii) and A locates at its ideology. If for instance, party A decides to locate further away from B than C, one needs to only compare the payoff of the closest position to α within the bounds of (i). Any deviation towards 0 within the bounds of (i) is not profitable if $q > 0.5 - \rho$. Moving x units towards the extremum costs A $(\rho + 0.5q)x$, as moving is costly and such movement decreases partisans' votes. The share of antipartisans' votes gained is equal to $0.5(1-q)x$. This is unprofitable if

$$0.5(1-q) - 0.5q - \rho < 0 \iff \rho > 0.5 - q. \qquad \text{(towards extremum)}$$

A similar reasoning can be applied to deviations within bounds but towards the center. They are ruled out by keeping $\rho > 0.5q$: Again, say, party A's position lies within (ii) and it locates at its ideology. Moving x units towards the center but staying within (ii) costs $x\rho$ units and yields $0.5xq$ units more votes by partisans. This is unprofitable if

$$0.5q - \rho < 0 \iff \rho > 0.5q. \qquad \text{(towards center)}$$

Note that $0.5q > q - 0.5$; hence, (towards center) and (towards extremum) are

sufficient for party C to not be tempted to move away from γ to some position $p > \beta$ such that $p - \beta > \beta - \alpha$. Having ruled out improvements within bounds, it remains to be secured that parties cannot improve by switching to "another part" of the function's domain; that is, level differences need to be analysed.[26]

(A beyond β):

$$0.5(\alpha+\beta)q + \big(1-0.5(\gamma+\beta)\big)(1-q) \geq 0.5(\gamma-\beta)q - (\beta-\alpha)\rho$$

(A beyond γ):

$$0.5(\alpha+\beta)q + \big(1-0.5(\gamma+\beta)\big)(1-q) \geq (1\text{-}\gamma)q + 0.5(\gamma+\beta)(1-q) - (\gamma-\alpha)\rho$$

(A towards 0; $2\beta-\gamma \geq 0$):

$$0.5(\alpha+\beta)q + \big(1-0.5(\gamma+\beta)\big)(1-q) \geq 0.5(3\beta-\gamma)q + \big(1-0.5(3\beta-\gamma)\big)(1-q) - (\alpha-2\beta+\gamma)\rho$$

(A towards 1; $2\gamma-\beta \leq 1$):

$$0.5(\alpha+\beta)q + \big(1-0.5(\gamma+\beta)\big)(1-q) \geq \big(1\text{-}0.5(3\gamma-\beta)\big)q + 0.5(3\gamma-\beta)(1-q) - (2\gamma-\beta-\alpha)\rho$$

(B beyond α):

$$0.5(\gamma-\alpha)q \geq \alpha q + \big(1-0.5(\gamma+\alpha)\big)(1-q) - (\beta-\alpha)\rho$$

(B beyond γ):

$$0.5(\gamma-\alpha)q \geq (1-\gamma)q + 0.5(\gamma+\alpha)(1-q) - (\gamma-\beta)\rho$$

(B towards 0; $2\alpha-\gamma \geq 0$):

$$0.5(\gamma-\alpha)q \geq 0.5(3\alpha-\gamma)q + \big(1-0.5(3\alpha-\gamma)\big)(1-q) - (\beta-2\alpha+\gamma)\rho$$

(B towards 1; $2\gamma-\alpha \leq 1$):

$$0.5(\gamma-\alpha)q \geq \big(1-0.5(3\gamma-\alpha)\big)q + 0.5(3\gamma-\alpha)(1-q) - (2\gamma-\alpha-\beta)\rho$$

[26] Note that deviations to another party's position (to its exact position) are ruled out by Lemma 4 when ideologies are sufficiently spread. In its proof, I show that there is generally one position (either slightly more extreme or slightly more moderate) that gives a weakly higher payoff than a shared position. Hence, a deviation to a shared position does not have to be considered and is ruled out by other constraints.

(C beyond α):

$$\big(1-0.5(\gamma+\beta)\big)q+0.5(\gamma+\beta)(1-q)\geq \alpha q+\big(1-0.5(\alpha+\beta)\big)(1-q)-(\gamma-\alpha)\rho$$

(C beyond β):

$$\big(1-0.5(\gamma+\beta)\big)q+0.5(\gamma+\beta)(1-q)\geq 0.5(\beta-\alpha)q-(\gamma-\beta)\rho$$

(C towards 0; $2\alpha-\beta\geq 0$):

$$\big(1-0.5(\gamma+\beta)\big)q+0.5(\gamma+\beta)(1-q)\geq 0.5(3\alpha-\beta)q+\big(1-0.5(3\alpha-\beta)\big)(1-q)-(\gamma-2\alpha+\beta)\rho$$

(towards center): $\rho>0.5q$

(towards extremum): $\rho>0.5-q$

The above system of inequalities represents the relevant deviations for each party. Each inequality makes sure that the party referred to in parentheses (e.g. party A for constraint (A beyond β)) is not tempted to deviate to another position in an ideology-faithful equilibrium. If those inequalities hold, there is a value of ρ for which ideology-faithful equilibria are possible. Note that constraint (B beyond α), for instance, does not compare a deviation of B to α but to $\lim_{\varepsilon\to 0^+}\alpha-\varepsilon$. Isolating expressions that may be negative- or zero-valued, I obtain the following system of inequalities, referred to as (q-bounds).

$$(3\beta+\alpha-2)q\geq\gamma+\beta-2-2(\beta-\alpha)\rho \qquad (A \text{ beyond } \beta)$$

$$(\alpha-4+3\beta+4\gamma)q\geq 2\big(\gamma+\beta-1-(\gamma-\alpha)\rho\big) \qquad (A \text{ beyond } \gamma)$$

$$q\geq\frac{2\gamma-2\beta-2(\alpha-2\beta+\gamma)\rho}{\alpha+3\gamma-4\beta} \qquad (A \text{ towards } 0;\ 2\beta-\gamma\geq 0)$$

$$(\alpha-4+7\gamma)q\geq 4\gamma-2-2(2\gamma-\beta-\alpha)\rho \qquad (A \text{ towards } 1;\ 2\gamma-\beta\leq 1)$$

$$q\geq\frac{2-\gamma-\alpha-2(\beta-\alpha)\rho}{2(1-2\alpha)} \qquad (B \text{ beyond } \alpha)$$

$$q\geq\frac{\gamma+\alpha-2(\gamma-\beta)\rho}{2(2\gamma-1)} \qquad (B \text{ beyond } \gamma)$$

$$q\geq\frac{2-3\alpha+\gamma-2(\beta-2\alpha+\gamma)\rho}{2-7\alpha+3\gamma} \qquad (B \text{ towards } 0;\ 2\alpha-\gamma\geq 0)$$

$$q\geq\frac{3\gamma-\alpha-2(2\gamma-\alpha-\beta)\rho}{7\gamma-3\alpha-2} \qquad (B \text{ towards } 1;\ 2\gamma-\alpha\leq 1)$$

$$(4-2\gamma-3\beta-3\alpha)q \geq 2-\alpha-\gamma-2\beta-2(\gamma-\alpha)\rho \qquad (C \text{ beyond } \alpha)$$

$$(2+\alpha-3\beta-2\gamma)q \geq -\beta-\gamma-2(\gamma-\beta)\rho \qquad (C \text{ beyond } \beta)$$

$$(4-2\gamma-6\alpha)q \geq 2-3\alpha-\gamma-2(\gamma-2\alpha+\beta)\rho \qquad (C \text{ towards } 0;\ 2\alpha-\beta \geq 0)$$

$$\rho > 0.5q \qquad \text{(towards center)}$$

$$\rho > 0.5-q \qquad \text{(towards extremum)}$$

Luckily, the number of constraints can be reduced drastically. An important step therein lies in showing that 2ρ is the smallest upper bound for q. The next paragraphs rule out any other bounds being smaller than 2ρ in sufficiently spread games.

Consider, for instance, constraint (A beyond β). As $(3\beta+\alpha-2)$ may be negative-valued, this constraint could represent an upper bound for q. In the following derivations, I show that, in games with sufficiently spread ideologies, $3\beta+\alpha-2<0$ implies 2ρ is a lower upper bound than (A beyond β).

$$\frac{\gamma+\beta-2-2(\beta-\alpha)\rho}{3\beta+\alpha-2} \geq 2\rho \iff 2\rho(3\beta+\alpha-2) \geq \gamma+\beta-2-2(\beta-\alpha)\rho$$
$$\iff 2\rho(4\beta-2) \geq \gamma+\beta-2,$$

which is true, as by Assumption (Spread), $3\beta>\gamma$ and ρ can reach its upper bound (which will be shown to be the case below). Hence, when $3\beta+\alpha-2<0$, (A beyond β) can be accounted for as long as there is no other upper bound on ρ except 0.5. If $3\beta+\alpha-2\geq 0$, the constraint is satisfied, as $\gamma+\beta-2-2(\beta-\alpha)\rho<0$. Hence, ($A$ beyond β) can be disregarded.

(A towards 1) imposes no upper bound on q, as $\alpha-4+7\gamma>0$ by Assumption (Spread):

$$\alpha-4+7\gamma > \alpha-4(\gamma+\beta)+7\gamma = \alpha+3\gamma-4\beta = 3(\gamma-\beta)-(\beta-\alpha) > 0.$$

When (A beyond γ) imposes an upper bound on q, it does not have to be considered, as it can be met as long as 0.5 is the tightest upper bound on ρ:

$$2\rho \leq \frac{2\big(\gamma+\beta-1-(\gamma-\alpha)\rho\big)}{\alpha-4+3\beta+4\gamma} \iff$$
$$2\rho(\alpha-4+3\beta+4\gamma+\gamma-\alpha) \geq 2\gamma+2\beta-2 \iff$$
$$2\rho(-4+3\beta+5\gamma) \geq 2\gamma+2\beta-2$$

As $-2+\beta+3\gamma > -2+(1-\gamma)+3\gamma = -1+2\gamma > 0$.
Consider now (C beyond α) and suppose it imposes an upper bound on q as $4-2\gamma-3\beta-3\alpha \leq 0$. The below derivations show that 2ρ imposes a tighter bound; hence, (C beyond α) can be omitted when $4-2\gamma-3\beta-3\alpha \leq 0$.

$$\begin{aligned}&(4-2\gamma-3\beta-3\alpha)2\rho \geq 2-\alpha-\gamma-2\beta-2(\gamma-\alpha)\rho \iff\\ &2\rho(4-\gamma-3\beta-4\alpha) \geq 2-\alpha-\gamma-2\beta,\end{aligned}$$

as by Assumption (Spread) $1-2\alpha+1-\beta-\alpha > 0$.
Now, consider (C beyond β). Note that when $2+\alpha-3\beta-2\gamma \geq 0$, the constraint is met (RHS negative). The following derivations show that (C beyond β) is accounted for by $2\rho > q$ when $2+\alpha-3\beta-2\gamma < 0$ and can, hence, be ignored.

$$\begin{aligned}&(2+\alpha-3\beta-2\gamma)2\rho \geq -\beta-\gamma-2(\gamma-\beta)\rho \iff\\ &2\rho(2+\alpha-4\beta-\gamma) \geq -\beta-\gamma,\end{aligned}$$

which is true as long as there is no other upper bound on ρ than 0.5 (as $2-3\beta+\alpha > 0$ as $2(1-\beta) > \gamma-\beta > \beta-\alpha$). Hence, the constraint is redundant.
Consider now (C towards 0). The below derivations show that if $4-2\gamma-6\alpha < 0$, (C towards 0) is accounted for by making sure $2\rho > q$:

$$\begin{aligned}&(4-2\gamma-6\alpha)2\rho \geq 2-3\alpha-\gamma-2(\gamma-2\alpha+\beta)\rho \iff\\ &(4-\gamma-8\alpha+\beta)2\rho \geq 2-3\alpha-\gamma,\end{aligned}$$

which holds as long as ρ is only bounded above by 0.5, as $2-5\alpha+\beta > 0$ (as $\alpha < 1/2$ Lemma 10).
To determine bounds for ρ in terms of α, β and γ, one can insert 2ρ in all inequalities imposing a lower bound on q. To find out whether an admissible value of ρ can be found such that all constraints hold at once, the following inequalities must be consistent: whenever a lower bound is higher than an upper bound, there is no such value of ρ.

$$\begin{aligned}\rho &> \frac{\gamma+\beta-1}{3\beta+5\gamma-4} && (A \text{ beyond } \gamma)\\ \rho &> \frac{\gamma-\beta}{2\alpha+4\gamma-6\beta} && (A \text{ towards } 0;\ 2\beta-\gamma \geq 0)\\ \rho &> \frac{2\gamma-1}{9\gamma-\beta-4} && (A \text{ towards } 1;\ 2\gamma-\beta \leq 1)\end{aligned}$$

$$\rho > \frac{2-\gamma-\alpha}{2(2-5\alpha+\beta)} \qquad (B \text{ beyond } \alpha)$$
$$\rho > \frac{\gamma+\alpha}{2(5\gamma-2-\beta)} \qquad (B \text{ beyond } \gamma)$$
$$\rho \geq \frac{2-3\alpha+\gamma}{2(2-9\alpha+4\gamma+\beta)} \qquad (B \text{ towards } 0;\ 2\alpha-\gamma \geq 0)$$
$$\rho \geq \frac{3\gamma-\alpha}{9\gamma-4\alpha-2-\beta} \qquad (B \text{ towards } 1;\ 2\gamma-\alpha \leq 1)$$

$$\rho > \frac{2-\alpha-\gamma-2\beta}{2(4-\gamma-3\beta-4\alpha)} \qquad (C \text{ beyond } \alpha)$$
$$\rho > \frac{2-3\alpha-\gamma}{2(4-\gamma-8\alpha+\beta)} \qquad (C \text{ towards } 0;\ 2\alpha-\beta \geq 0)$$

$$\rho > 0.5q \qquad \text{(towards center)}$$
$$\rho > \frac{1}{6} \qquad \text{(towards extremum)}$$

As was the case above, some expressions arise that may be negative- or zero-valued, namely $4-\gamma-3\beta-4\alpha$ and $3\beta+5\gamma-4$. These expressions correspond to constraints that have to only be considered if their version in system (q-bounds) imposes a lower bound on q (as was shown above). As both expressions are positive-valued if their counterparts in system (q-bounds) impose lower bounds, negative values do not need to be considered. The other expressions (in denominators) are always positive-valued, as I assume ($\alpha < \beta < \gamma$ and $\beta - \alpha < \gamma - \beta$). Theorem 2 states that it is sufficient to consider the constraints imposed on ρ by (B beyond α) and (B beyond γ) under these assumptions. To verify this, I show that each of the other lower bounds on ρ are smaller than its lowest upper bound 0.5 and do thereby not impose restrictions that cannot be met, provided that (B beyond α) and (B beyond γ) hold.
(A beyond γ) is redundant:

$$\frac{\gamma+\beta-1}{3\beta+5\gamma-4} < \frac{1}{2} \iff 2\gamma+2\beta-2 < 3\beta+5\gamma-4 \iff$$
$$0 < \beta+3\gamma-2 \iff 0 < 2\beta+6\gamma-4,$$

which is true if (A beyond γ) imposes a lower bound on q (implying $0 \leq \alpha-4+3\beta+4\gamma < 2\beta+6\gamma-4$).

(A towards 0) is redundant:

$$\frac{\gamma-\beta}{2\alpha+4\gamma-6\beta}<\frac{1}{2} \iff \frac{\gamma-\beta}{\alpha+2\gamma-3\beta}<1 \iff$$
$$\gamma-\beta<2(\gamma-\beta)-(\beta-\alpha) \iff \beta-\alpha<\gamma-\beta,$$

which is true by assumption.
(A towards 1) is redundant:

$$\frac{2\gamma-1}{9\gamma-\beta-4}<\frac{1}{2} \iff 4\gamma-2<4(2\gamma-1)+\gamma-\beta \iff 0<2(2\gamma-1)+\gamma-\beta,$$

which is true in any ideology-faithful equilibrium as $2\gamma>1$ is necessary (Lemma 10).
(B towards 0) and (B towards 1) can be ruled out right away, as will be shown below (steps (1) and (2)):
(1) (B beyond α) violated if $2\alpha-\gamma\geq 0$: Towards contradiction, suppose $2\alpha-\gamma\geq 0$ and (B beyond α) holds, then

$$\frac{2-\gamma-\alpha}{2(2-5\alpha+\beta)}<\frac{1}{2} \iff 2-\gamma-\alpha<2-5\alpha-\beta \iff$$
$$0<\gamma-\beta-4\alpha=(\gamma-2\alpha)-2\alpha-\beta<0,$$

contradicting $2\alpha-\gamma\geq 0$ and $\beta>0$.
(2) (B beyond γ) violated if $2\gamma-\alpha\leq 1$: $2\gamma-\alpha\leq 1 \iff -1+2\gamma\leq\alpha \iff -1+3\gamma\leq\alpha+\gamma$. Furthermore note that as $1\geq 2\gamma-\alpha>2\gamma-\beta$:

$$\alpha+\gamma\geq 3\gamma-1>5\gamma-2-\beta \Rightarrow \frac{\gamma+\alpha}{2(5\gamma-2-\beta)}>\frac{1}{2} \iff (B \text{ beyond } \gamma) \text{ violated.}$$

(C beyond α) is redundant if (B beyond α) holds: Note that $2-\gamma+\alpha-2\beta\geq 2\gamma-\gamma+\alpha-2\beta=\gamma-\beta-(\beta-\alpha)>0$ by assumption.

$$\frac{2-\gamma-\alpha}{2(2-5\alpha+\beta)}<\frac{1}{2} \iff 2-\gamma-\alpha<2-5\alpha+\beta \Rightarrow$$
$$2-\alpha-\gamma<2-5\alpha+\beta+(2-\gamma+\alpha-2\beta) \iff$$
$$2-\alpha-\gamma-2\beta<2-5\alpha+\beta+(2-\gamma+\alpha-4\beta).$$

Hence, (C beyond α) holds.

(C towards 0) is redundant:

$$\frac{2-3\alpha-\gamma}{2(4-\gamma-8\alpha+\beta)} < \frac{1}{2} \iff 2-3\alpha-\gamma < 4-\gamma-8\alpha+\beta \iff$$
$$0 < 2(1-2\alpha)+\beta-\alpha, \text{ which is true by assumption and } \alpha < 0.5.$$

(towards extremum) is accounted for by (B to α):

$$\frac{2-\gamma-\alpha-2(\beta-\alpha)\rho}{2(1-2\alpha)} > 0.5-\rho \iff$$
$$2-\gamma-\alpha-1+2\alpha > 2\rho(\beta-\alpha-1+2\alpha) \iff 1-\gamma+\alpha > 2\rho(\beta+\alpha-1),$$

which is true for all values of ρ as $2-\gamma-\beta > 0 \iff 1-\gamma+\alpha > \beta+\alpha-1$. The only constraints of interest are now (B beyond α) and (B beyond γ). One can thereby deduce: when (B beyond α) and (B beyond γ) impose lower bounds on ρ that are smaller than 0.5, admissible values of ρ and q can be found such that the ideology-faithful equilibrium exists. This is the case if

$$\frac{1}{2} > \max\left\{\frac{2-\gamma-\alpha}{2(2-5\alpha+\beta)}, \frac{\gamma+\alpha}{2(5\gamma-2-\beta)}\right\}.$$

This space is, as can be seen in Figure 1.5, non-empty. Two examples of parameter combinations satisfying the above are
$(\alpha,\beta,\gamma) \in \{(0.2,0.35,0.64),(0.15,0.4,0.85)\}$.
The above statement is equivalent to $\gamma-4\alpha+\beta > 0 \wedge 4\gamma-2-\beta-\alpha > 0$ (simple reordering of the inequalities implied by the maximum operator). For all models satisfying Assumption (Spread), this condition is both necessary and sufficient. Sufficiency was shown above (e.g. by showing other constraints' redundancy). Necessity is implied by the fact that if either $\gamma-4\alpha+\beta < 0$ or $4\gamma-2-\beta-\alpha < 0$, ($B$ beyond α) and (B beyond γ) impose lower bounds on ρ that are higher than the smallest upper bound and can, hence, not be met. Furthermore, knife-edge alliance equilibria in which party A and B share a position are precluded by the assumption $\beta > 1/3\gamma$ (see necessary conditions for the existence of a knife-edge alliance equilibrium in the proof of Lemma 4). An alliance between party C and B is not possible either, as for such alliance to be possible, $\beta-\alpha > \gamma-\beta$ is a necessary condition. To see why, note that the alliance has to locate at $p_B = \frac{\alpha+2q}{4q-1}$ (derivations analogous to those in Lemma 4). Towards contradiction, suppose $\beta-\alpha < \gamma-\beta$ and $p_B = \frac{\alpha+2q}{4q-1}$. As $p_B \leq \beta$ (by the proof of Lemma 4), $p_B-\alpha \leq \beta-\alpha < \gamma-\beta \leq 1-p_B \iff 2p_B < 1+\alpha$;

hence, as $1 > q > 0.5$[27],

$$2\frac{\alpha+2q}{4q-1} < 1+\alpha \iff 2\alpha+4q < (1+\alpha)(4q-1) \iff$$
$$2\alpha+1+\alpha < 4q(1+\alpha-1) \iff 3+1/\alpha < 4q, \text{ contradicting } \alpha < 1,\ q < 1.$$

Lastly, by Theorem 1, there are no other classes of equilibria than knife-edge alliance, ideology-faithful and extreme equilibria when linearity is satisfied. Hence, the conditions stated in Theorem 2 are both necessary and sufficient for the existence of values of q that allow for moderate, extreme equilibria and no equilibrium, respectively, for some ρ when ideologies are sufficiently spread. Thus, if the conditions are met, a change in antipartisanship by itself can explain transitions between them. □

Proof of Theorem 3. The Theorem follows from the necessary conditions for the existence of an ideology-faithful equilibrium in the proof Lemma 10 and those for the existence of a knife-edge alliance equilibrium in the proof of Lemma 4. □

Appendix 1.B

Supplementary results

Theorem 4. *Consider a combination of ideology values* (α,β,γ)*. There are cost functions that satisfy linearity for which a change in the share of antipartisans* $1-q$ *by itself can explain transitions from no equilibrium to moderate equilibria and from moderate to extreme equilibria if and only if* (α,β,γ) *lie in the parameter space described by the consolidated constraints.*

Proof. I consider the case in which $\beta-\alpha \leq \gamma-\beta$. The case in which $\beta-\alpha \geq \gamma-\beta$ follows by symmetry.

Theorem 2 covers all combinations of sufficiently spread ideologies. Theorem 3 covers the cases in which moderate equilibria are generally not possible. Hence, combinations of ideologies that are neither sufficiently spread nor satisfy the conditions in Theorem 3 remain to be analysed. The analysis is divided into five steps: First, I show that no moderate equilibria are possible in games in with $\beta-\alpha=\gamma-\beta$ and $(\alpha,\beta,\gamma) \neq (0,0.5,1)$ and that for $(\alpha,\beta,\gamma) \neq (0,0.5,1)$, an ideology-faithful equilibrium exists for some q and ρ. Thereafter, I derive the conditions under which knife-edge alliance equilibria are possible. Then, I derive under which conditions,

[27] This holds by the results obtained in the proof of Lemma 4, as $0.5 > \rho > q-0.5$ and $q > 0.5$ are necessary conditions for the existence of a knife-edge alliance equilibrium.

the ideology-faithful equilibrium is possible if $\gamma < 9/14$ and $\beta \leq 1/3\gamma$ (then, knife-edge alliance equilibria might not exist). Thereafter, I consider combinations of ideology values that are not sufficiently spread but satisfy $\beta \geq 1/3\gamma$. The last step is the analysis of the parameter space considered in Theorem 2 including the case in which $\beta = 1/3\gamma$.
Step 1: "no moderate equilibrium if $\beta - \alpha = \gamma - \beta$ and $(\alpha, \beta, \gamma) \neq (0, 0.5, 1)$ and only if $(\alpha, \beta, \gamma) = (0, 0.5, 1)$, there may be an ideology-faithful equilibrium"

Suppose $(\alpha, \beta, \gamma) \neq (0, 0.5, 1)$. Ideology-faithful equilibria are ruled out by Lemma 2. Knife-edge alliance equilibria are not possible either: Equal distances would imply $\gamma - p_A \leq \gamma - \beta = \beta - \alpha \leq \beta \leq p_A \iff \gamma \leq 2p_A$. This, however, contradicts the necessary condition for the existence of a knife-edge alliance equilibrium $\gamma > 3p_A$. The case in which C and B share a position is analogous. Hence, when $\beta - \alpha = \gamma - \beta$, there is no moderate equilibrium.
However, when $(\alpha, \beta, \gamma) = (0, 0.5, 1)$, there is an ideology faithful equilibrium. Relevant deviations from $(p_A, p_B, p_C) = (0, 0.5, 1)$ are

(A to β):
$$0.25q + 0.5\big(1-q\big) > 0.5\big(0.75q + 0.25(1-q)\big) - 0.5\rho \iff \rho > q - 0.75$$
(A to $\lim_{\varepsilon \to 0^+} \beta - \varepsilon$):
$$0.25q + 0.5\big(1-q\big) \geq 0.5q + 0.25(1-q) - 0.5\rho \iff \rho \geq q - 0.5$$
(B to α):
$$0.5q > 0.25q + 0.25(1-q) - 0.5\rho \iff \rho > 0.5 - q.$$

A combination of ρ and q to satisfy the above can be found. Note that the above constraints are necessary and sufficient as deviations by C and to γ are analogous.
Step 2: "knife-edge alliance equilibria."

To determine the conditions under which a knife-edge alliance equilibrium exists, consider, besides the necessary conditions stated in the proof of Lemma 4, the

below inequalities that rule out deviations:

(A to α):

$$0.5q > \rho$$

(A to γ):

$$0.5\big(0.5(p_A+\gamma)q+\big(1-0.5(p_A+\gamma)\big)(1-q)\big)-(p_A-\alpha)\rho > 0.5\big((1\text{-}0.5(p_A+\gamma))q+0.5(p_A+\gamma)(1-q)\big)-(\gamma-\alpha)\rho$$

(A beyond γ):

$$0.5\big(0.5(p_A+\gamma)q+\big(1-0.5(p_A+\gamma)\big)(1-q)\big)-(p_A-\alpha)\rho \geq (1\text{-}\gamma)q+0.5(p_A+\gamma)(1-q)-(\gamma-\alpha)\rho$$

(A towards 1; $2\gamma-p_A\leq 1$):

$$0.5\big(0.5(p_A+\gamma)q+\big(1-0.5(p_A+\gamma)\big)(1-q)\big)-(p_A-\alpha)\rho \geq \big(1\text{-}0.5(3\gamma-p_A)\big)q+(3\gamma-p_A)(1-q)-(2\gamma-p_A-\alpha)\rho$$

(C to p_A):

$$\big(1-0.5(\gamma+p_A)\big)q+0.5(\gamma+p_A)(1-q) > 1/3-(\gamma-p_A)\rho$$

(C beyond p_A):

$$\big(1-0.5(\gamma+p_A)\big)q+0.5(\gamma+p_A)(1-q) \geq \mathrm{p}_A q+(1-p_A)(1-q)-(\gamma-p_A)\rho$$

(towards center): $\rho > q-0.5$

(A to α) follows from the proof of Lemma 4. The above does not consider any deviations of party B, as those are accounted for by ruling out deviations of party A. All that matters are level differences, which are the same for both parties as they choose the same position. Note that as $\gamma > 3p_A$ in any knife-edge equilibrium with $p_B = p_A$, it is not possible for any party to deviate to a position $\lim_{\varepsilon\to 0^+} p_A-(\gamma-$

$p_A) - \varepsilon$. The system can be simplified and expressed as follows:

$$
\begin{aligned}
&q > 2\rho && (A \text{ to } \alpha)\\
&(2q-1)(\gamma + p_A - 1)2 > -4(\gamma - p_A)\rho && (A \text{ to } \gamma)\\
&(2q-1)(4\gamma + 2p_A - 3) - (1 - \gamma + p_A) \geq -4(\gamma - p_A)\rho && (A \text{ beyond } \gamma)\\
&(2q-1)(7\gamma - p_A - 3) \geq -8(\gamma - p_A)\rho && (A \text{ towards 1; } 2\gamma - p_A \leq 1)\\
&(2q-1)(1 - \gamma - p_A) + 1/3 > -2(\gamma - p_A)\rho && (C \text{ to } p_A)\\
&(2q-1)(2 - \gamma - 3p_A) \geq -2(\gamma - p_A)\rho && (C \text{ beyond } p_A)\\
&\rho > q - 0.5 && (\text{towards center})
\end{aligned}
$$

Most of the constraints are redundant. To see why, consider the below derivations: (A to γ) accounted for by (A beyond γ): Note that (A to γ) is always satisfied if $\gamma + p_A - 1 \geq 0$. Hence, only the case in which $\gamma + p_A - 1 < 0 \iff p_A < 1 - \gamma \Rightarrow p_A^2 < (1-\gamma)^2$ has to be considered. Below, I show that $\gamma + p_A - 1 < 0$ implies (A beyond γ) being tighter than (A to γ). To do so, is use that by the proof of Lemma 4, $q = \big(1 - 0.5(\gamma + p_A)\big)/(1 - 2p_A)$. The following expression is the result of the subtraction of (A beyond γ) from (A to γ).

$$
\begin{aligned}
&(2q-1)(1-2\gamma) + 1 - \gamma + p_A > 0 \iff\\
&\frac{(1-2\gamma)(2-\gamma-p_A)}{1-2p_A} + \frac{(1-2p_A)(\gamma+p_A)}{1-2p_A} = \frac{2(1-2\gamma) + (\gamma+p_A)(1-2p_A+2\gamma-1)}{1-2p_A}\\
&= 2\frac{(1-2\gamma) + (\gamma+p_A)(\gamma-p_A)}{1-2p_A} > 0 \iff (1-2\gamma) + (\gamma+p_A)(\gamma-p_A) =\\
&1 - 2\gamma + \gamma^2 - \gamma^2 + \gamma^2 - p_A^2 = (1-\gamma)^2 - p_A^2 > 0, \text{ which is true as } \gamma + p_A - 1 < 0.
\end{aligned}
$$

Consider now (A towards 1). Interestingly, (A towards 1) can never be of relevance as otherwise (A beyond γ) would be violated. To see why, note that ρ is bounded above by 0.5. As the constraints are more likely to be met if ρ is of high values, one can use this bound to show under which conditions the constraints can be met for some value of ρ. Were I to find that they cannot hold for the highest admissible value of ρ, I would need to conclude their holding to be impossible for any such value of ρ. A necessary requirement for (A towards 1) to be relevant is $2\gamma - p_A \leq 1$. Using the upper bound on ρ imposed by Assumption linearity (0.5) and the expression for q derived in the proof of Lemma 4, I can show that (A towards 1) is never relevant: Towards contradiction, suppose $2\gamma - p_A \leq 1$ and (A beyond γ)

holds. Then,

$$2q(4\gamma+2p_A-3)=2(4\gamma+2p_A-3)\frac{1-0.5(\gamma+p_A)}{1-2p_A}\geq 3\gamma+3p_A-2-4(\gamma-p_A)\rho$$
$$>3\gamma+3p_A-2-(\gamma-p_A)\iff\frac{2-4p_A^2-5\gamma+2\gamma^2+p_A(1+2\gamma)}{p_A-0.5}>0\iff$$
$$2-4p_A^2-5\gamma+2\gamma^2+p_A(1+2\gamma)<0\iff 2-4p_A^2-3\gamma+2\gamma^2+2\gamma p_A<$$
$$<2\gamma-p_A\leq 1\iff 1+\gamma(2\gamma-3)+2p_A(\gamma-2p_A)<0.$$

Taking into account that $p_A<1/3p_C<1/3$ (necessary condition for a knife-edge equilibrium derived in proof of Lemma 4) must hold, the expression $1+\gamma(2\gamma-3)+2p_A(\gamma-2p_A)$ is negative-valued in the following cases

$$\begin{aligned}
&0\leq p_A<1/6 \text{ and } 0.5(1+2p_A)<\gamma<1-2p_A && \Rightarrow 2\gamma-p_A>1+p_A\geq 1\\
&1/6<p_A\leq 0.2 \text{ and } 1-2p_A<\gamma<0.5(1+2p_A) && \Rightarrow 2\gamma-p_A>2-5p_A\geq 1\\
&0.2<p_A<0.25 \text{ and } 3p_A<\gamma<0.5(1+2p_A) && \Rightarrow 2\gamma-p_A>5p_A\geq 1.
\end{aligned}$$

Hence, for values $2\gamma-p_A\leq 1$, the expression cannot be negative-valued and (A beyond γ) is violated, as was to be shown. Hence, (A towards 1) is not relevant. (C to p_A) is always satisfied:

$$(2q-1)(1-\gamma-p_A)+1/3\geq -1/3(2q-1)+1/3=$$
$$1/3(2-2q)\geq 0>-2(\gamma-p_A)\rho.$$

(C beyond p_A) is always satisfied: This holds true as $0<2-\gamma-3p_A=(1-\gamma)+(1-3p_A)$ and $q>0.5$ by the proof of Lemma 4. The only constraints to be considered are, hence, (A to α), (A beyond γ) and (towards center).

$$\begin{aligned}
&q>2\rho && (A \text{ to } \alpha)\\
&2q(4\gamma+2p_A-3)\geq 3\gamma+3p_A-2-4(\gamma-p_A)\rho && (A \text{ beyond } \gamma)\\
&\rho>q-0.5 && (\text{towards center})
\end{aligned}$$

(A beyond γ) is always satisfied if $4\gamma+2p_A-3\geq 0$. To see why this statement is true, consider the following derivations, in which I use that in any knife-edge alliance equilibrium $q=\big(1-0.5(\gamma+p_A)\big)/(1-2p_A)$ and $q-0.5<\rho$ (by constraint

(towards center)).

$$2q(4\gamma+2p_A-3)\geq 3\gamma+3p_A-2-4(\gamma-p_A)(q-0.5)>$$
$$>3\gamma+3p_A-2-4(\gamma-p_A)\rho \iff \frac{2-p_A^2+p_A(1-2\gamma)-5\gamma+3\gamma^2}{p_A-0.5}\geq 0 \iff$$
$$2-p_A^2+p_A(1-2\gamma)-5\gamma+3\gamma^2\leq 0.$$

By the proof of Lemma 4, $p_A<1/3\gamma$ and, as $4\gamma+2p_A-3\leq 4\gamma+2/3\gamma-3$, $\gamma\geq 9/14$. Furthermore, $p_A\geq 0.5(3-4\gamma)$; hence,

$$2-p_A^2+p_A(1-2\gamma)-5\gamma+3\gamma^2<2-p_A^2-5\gamma+3\gamma^2\leq$$
$$2-5\gamma+3\gamma^2-(1.5-2\gamma)^2=-(0.5-\gamma)^2<0.$$

Hence, as long as $\gamma\geq 9/14$ and the ideologies are neither sufficiently spread nor lie in the space described by Theorem 3, constraint (A to γ) can be satisfied. Constraint (A to α) and (towards center) can be satisfied as well, as the above holds even for $\rho=q-0.5<0.5q$. As the constraint is more likely to be met for higher values of ρ and $q-0.5$ is strictly smaller than $0.5q$, there are values $q-0.5<\rho<0.5q$ for which all constraints are satisfied. Hence, a moderate equilibrium exists for some value of ρ and q. If $\gamma<9/14$, constraint (A beyond γ) can be satisfied if the below quadratic expression implied by constraint (A beyond γ) is greater than 0:

$$\frac{2\big(1-0.5(p_A+p_C)\big)(4\gamma+2p_A-3)}{1-2p_A}>3\gamma+3p_A-2-2(\gamma-p_A)0.5q \tag{1.5}$$
$$\iff \frac{0.5-4.5p_A^2+p_A(5-4\gamma)-3\gamma+2.5\gamma^2}{p_A-0.5}>0 \iff \tag{1.6}$$
$$0>0.5-4.5p_A^2+p_A(5-4\gamma)-3\gamma+2.5\gamma^2 \tag{1.7}$$

Note that I am using the smallest upper bound ($0.5q$) on ρ. If the above holds with strict inequality, values $q-0.5<\rho<0.5q$ can be found that satisfy all constraints simultaneously. Line 1.7 can be shown to hold (implying all constraints to hold) for $\gamma<9/14$ in the following cases:

$0\leq p_A\leq 1/6$ and $\quad 0.2(3+4p_A)-0.2\sqrt{4-26p_A+61p_A^2}<\gamma<9/14$ or

$1/6<p_A<3/14$ and $\quad 3p_A<\gamma<9/14.$

Note that in any knife-edge alliance equilibrium, $\beta\leq p_A$. Furthermore, the bound in the first line is increasing in p_A. Hence, to make sure a knife-edge alliance equilibrium exists, one would like to decrease p_A as much as possible. This is only

admissible as long as $p_A \geq \beta$; hence, the final constraints on γ are

$$0 \leq \beta \leq 1/6 \text{ and } \quad 0.2(3+4\beta) - 0.2\sqrt{4-26\beta+61\beta^2} < \gamma < 9/14 \text{ or}$$
$$1/6 < \beta < 3/14 \text{ and } \quad 3\beta < \gamma < 9/14.$$

It remains to be determined under which conditions an ideology-faithful equilibrium can make up for the missing of the knife-edge alliance equilibrium when $\beta < 1/3\gamma$. As a knife-edge alliance equilibrium always exists for values of $\gamma \geq 9/14$, it suffices to consider cases in which $\gamma < 9/14$ and $\beta < 1/3\gamma$:
Step 3: "ideology-faithful equilibrium when $\gamma < 9/14$ and $\beta < 1/3\gamma$"

As derived in the proof of Theorem 2, in any ideology-faithful equilibrium, the constraints on q stated in (q-bounds) need to hold. As in this case $\beta < 1/3\gamma$, deviations of party A to position β need to be considered as well. Hence, one additional constraint needs to be added to the system stated in (q-bounds). Note that, as was argued in Theorem 2, deviations of B or C to their opponents' positions or A to γ do not need to be considered.

$$\begin{aligned}(A \text{ to } \beta)\colon\ & 0.5(\alpha+\beta)q + \big(1-0.5(\gamma+\beta)\big)(1-q) > \\ & 0.5\big(0.5(\beta+\gamma)q + \big(1-0.5(\beta+\gamma)\big)(1-q)\big) - (\beta-\alpha)\rho \iff \\ & q < \frac{0.5(\gamma+\beta)-1-(\beta-\alpha)2\rho}{\alpha+\beta-1}\end{aligned}$$

For the parameter constellation considered here, this is an upper bound, as $\alpha+\beta-1<0$. The system can, again, be reduced drastically:
First, note that (A towards 0) is never relevant, as $2\beta-\gamma<0$ by assumption.
Consider now the case in which (A towards 1) imposes an upper bound on q. The constraint can be disregarded, as the implied bound is higher than 2ρ for $\rho = 0.5$ (see Lemma 11):

$$1 < \frac{4\gamma-2-2(2\gamma-\beta-\alpha)0.5}{\alpha-4+7\gamma} \iff -4+9\gamma-\beta \geq 4\gamma-2 \iff 5\gamma-\beta-2>0.$$

This is true as $5\gamma-\beta-2 > 4\gamma-2 > 0$. The constraint does, however, not matter even if (A towards 1) imposes a lower bound on q. This is the case if the LHS of the equation is weakly positive. For sufficiently high ρ, the RHS can be assured to be negative:

$$4\gamma-2-2\gamma+\beta+\alpha = 2\gamma+\beta+\alpha-2 < 8/3\gamma-2 < 12/7-2 < 0.$$

Hence, (A towards 1) does not need to be considered. Further, note that $3\beta+\alpha-2 <$

$4\beta - 2 < 4/3\gamma - 2 < 6/7 - 2 < 0$ and $\alpha - 4 + 3\beta + 4\gamma < 4/3\gamma + 4\gamma - 4 < 0$ by assumption; hence, both (A beyond β) and (A beyond γ) impose upper bounds on q. Note that for all parameter constellations considered in this step, (A to β) is accounted for by (A beyond β). To show this, I use Lemma 11:

$$\frac{0.5(\gamma-\beta)-1+\alpha}{\alpha+\beta-1} > \frac{\gamma-2+\alpha}{3\beta+\alpha-2} \iff$$
$$(0.5(\gamma-\beta)-1+\alpha)(3\beta+\alpha-2) > (\gamma-2+\alpha)(\alpha+\beta-1) \iff$$
$$(0.5(\gamma-\beta)-1+\alpha)(2\beta-1) > (\gamma-2+\alpha-0.5(\gamma-\beta)+1-\alpha)(\alpha+\beta-1) =$$
$$(0.5(\gamma+\beta)-1)(\alpha+\beta-1) \iff$$
$$(0.5(\gamma+\beta)-\beta-1+\alpha)(2\beta-1) > (0.5(\gamma+\beta)-1)(\alpha+\beta-1) \iff$$
$$(0.5(\gamma+\beta)-1)(2\beta-1-\alpha-\beta+1) = (0.5(\gamma+\beta)-1)(\beta-\alpha) >$$
$$> (\beta-\alpha)(2\beta-1) \iff 0.5(\gamma+\beta)-1 > 2\beta-1 \iff \gamma > 3\beta,$$

which is true by assumption. Later statements follow very similar arguments. As the derivations implying them are purely algebraic and not really insightful, I refrain from expressing them in detail and only state that "it can be shown that" the constraint under consideration is accounted for by some other constraint.

As was shown in Theorem 2, (B towards 0) and (B towards 1) can never be relevant.[28]
Further, note that (C beyond α) and (C towards 0) impose lower bounds on q, as $1/3\gamma > \beta > \alpha$ by assumption.
(C beyond α) is accounted for by (B beyond α):

$$2-\gamma-\alpha-2(\beta-\alpha)\rho > 2-\gamma-\alpha-2(\gamma-\alpha)\rho > 2-\gamma-\alpha-2\beta-2(\gamma-\alpha)\rho$$
$$4-2\gamma-3\beta-3\alpha = 2(1-2\alpha)+2+\alpha-3\beta-2\gamma >$$
$$> 2(1-2\alpha)+2-3\gamma+\alpha > 2(1-2\alpha)+1/14+\alpha,$$

hence, the bound imposed by (B beyond α) is always tighter than the one imposed by (C beyond α).
(C beyond β) is always satisfied under the assumptions made, as $2+\alpha-3\beta-2\gamma > 2+\alpha-\gamma-2\gamma > 0$.

[28] In the proof of Theorem 2, this statement was proven using an upper bound on q that might not be the tightest one for all combinations of ideology values. The statement holds even if the lowest upper bound is not the one invoked in the proof: if the constraint cannot be satisfied using a potentially higher upper bound on q than the actual one, it cannot hold for the smallest upper bound on q either.

Lastly, (C towards 0) is accounted for by (B beyond α):

$$\begin{aligned}
&2-\gamma-\alpha-2(\beta-\alpha)\rho > 2-\gamma-3\alpha-2(\beta-\alpha+\gamma-\alpha)\rho\\
&4-2\gamma-6\alpha = 2(1-2\alpha)+2-2\alpha-2\gamma > 2(1-2\alpha)+2-3\gamma >\\
&> 2(1-2\alpha)+1/14.
\end{aligned}$$

In the proof of Theorem 2, (towards extremum) was shown to be accounted for by (B to α) for all admissible parameter values. The system can, hence, be reduced to upper bounds on q imposed by (A beyond β), (A beyond γ) and (towards center) and lower bounds imposed by (B beyond α) and (B beyond γ).

Another case in which knife-edge alliance equilibria are generally not possible remains to be analysed: $\gamma \leq 3\beta$, but the ideology values are not sufficiently spread:
Step 4: "$\gamma \leq 3\beta$ and ($\alpha \geq 1-\beta$ or $\gamma \leq 1-\beta$)"

The analysis follows mostly from Theorem 2 and the above. The main difference lies in the fact that there are parameter constellations for which (A beyond γ) imposes a lower bound and some for which (C beyond α) imposes an upper bound on q. Note that only one of them can ever impose an upper bound on q, never both:

$$\begin{aligned}
&4-2\gamma-3\beta-3\alpha<0 \iff -2\alpha<\alpha-4+3\beta+2\gamma \iff\\
&0<2(\gamma-\alpha)<\alpha-4+3\beta+4\gamma.
\end{aligned}$$

The same is true in case one of them imposes a lower bound on q: as $\alpha \geq 1-\beta$ or $\gamma \geq 1-\beta$, the other constraint has to be an upper bound. To see why, suppose $\alpha \geq 1-\beta$. Then, $4-2\gamma-3\beta-3\alpha<0$, as $4-2\gamma-3\beta-3\alpha \leq 4-2\gamma-3\beta-3(1-\beta)=1-2\gamma<0$. If $\gamma \leq 1-\beta$, on the other hand, $\alpha-4+3\beta+2\gamma<0$, as $\alpha-4+3\beta+4\gamma \leq \alpha-4+3\beta+4(1-\beta)=\alpha-\beta<0$.
As $\alpha<1-\beta$ and $\gamma>1-\beta$ cannot hold at the same time under the assumptions made in this step, only one of the bounds can be a lower bound. Note that as $\gamma \leq 3\beta$, deviations to positions that are equal to an opponent's ideology value do not need to be considered (cf. Theorem 2).
(A beyond β) was shown to be irrelevant in the proof of Theorem 2 under the assumption $1/3\gamma<\beta$. When $\gamma=3\beta$, (A beyond β) and 2ρ impose the same bound on q for $\rho=0.5$. As $\rho<0.5$ by Assumption linearity and 2ρ is tighter than (A beyond β) for any $\rho \in (0,0.5)$, the upper bound 2ρ accounts for (A beyond β) for all values $\gamma \leq 3\beta$. To see why, consider the following derivations: Suppose $3\beta=\gamma$ and $3\beta+\alpha-2<0$ and note that this is the only instant in which the analysis of constraint (A beyond β) in the proof of Theorem 2 differs from the analysis in

Step 4. For any value of $\varepsilon \in (0, 0.5)$ (and, hence, for $\rho = 0.5 - \varepsilon$), the following holds:

$$\frac{\gamma + \beta - 2 - 2(\beta - \alpha)(0.5 - \varepsilon)}{3\beta + \alpha - 2} - 2(0.5 - \varepsilon) > 0 \iff 0 > -2\varepsilon\left(1 + \frac{\beta - \alpha}{3\beta + \alpha - 2}\right)$$
$$\iff 1 + \frac{\beta - \alpha}{3\beta + \alpha - 2} > 0 \iff \beta - \alpha < 2 - \alpha - 3\beta \iff \beta < 0.5,$$

which is true as $\beta = 1/3\gamma \leq 1/3$.

Now consider (A beyond γ). It can be shown that this constraint is either not a higher lower bound than (B beyond γ) or trivially met. This holds for all parameter values satisfying $\alpha - 4 + 3\beta + 4\gamma > 0$ and $0 \leq \alpha < \beta < \gamma \leq 1$. Note that this is the only case in which (A beyond γ) imposes a lower bound.[29]

(A towards 0) is satisfied when 2ρ imposes the tightest upper bound on q (see proof of Theorem 2) and has to only be considered if there is a tighter upper bound than 2ρ.

(A towards 1) is not relevant: If it imposes an upper bound on q, it is accounted for by upper bound 2ρ (this was shown above). If (A towards 1) is a lower bound, it is either trivially met or bound (B beyond γ) is tighter. This can be shown to hold for all values of α, β and γ that satisfy $\alpha - 4 + 7\gamma \geq 0$ and $0 \leq \alpha < \beta < \gamma \leq 1$. Note that this is the only case in which (A towards 1) can impose a lower bound on q.

(C beyond α) is accounted for by (B beyond γ) if it imposes a lower bound or trivially met. This, again, can be shown for all parameter values satisfying $4 - 2\gamma - 3\beta - 3\alpha \geq 0$ and $0 \leq \alpha < \beta < \gamma \leq 1$ and all values of $\rho \in (0, 0.5)$. Note that this is the only case in which (C to α) imposes a lower bound.

(C beyond β) was shown to not be relevant in the proof of Theorem 2 under the assumption $1/3\gamma < \beta$, which is satisfied in this case.

(C towards 0) is irrelevant if it imposes an upper bound on q, as was shown in the proof of Theorem 2. In case it imposes a lower bound, it is either accounted for by constraint (B beyond α) or trivially met. This can be shown to hold for all parameter values satisfying $4 - 2\gamma - 6\alpha \geq 0$ and $0 \leq \alpha < \beta < \gamma \leq 1$, which is the only case in which (C towards 0) imposes a lower bound. In the proof of Theorem 2, (towards extremum) was shown to be accounted for by (B to α) for all admissible parameter values. (B towards 1) and (B towards 0) were shown to be irrelevant in the proof of Theorem 2.[30] Hence, the only constraints of relevance are the three lower bounds (A towards 0), (B beyond α) and (B beyond γ) and the upper bounds (A beyond γ), (C beyond α) and (towards center).

[29] It is easy to see that the constraint is met when $\alpha - 4 + 3\beta + 4\gamma = 0$ (RHS negative).

[30] Cf. footnote 28.

Step 5: "$\gamma \leq 3\beta$ and $\alpha < 1-\beta$ and $\gamma > 1-\beta$."

This represents a slightly more general version of the analysis in Theorem 2; that is, all cases have been considered in the respective proof and only $\gamma = 3\beta$ needs to be accounted for below. As the fact that $\gamma < 3\beta$ was only used in one instance (when (A beyond β) imposes an upper bound), it suffices to adjust only this step and keep the remaining analysis equivalent. As was argued in Step 4, the upper bound 2ρ accounts for (A beyond β) for all values $\gamma \leq 3\beta$. Necessary and sufficient conditions for the existence of a moderate equilibrium for $\gamma \leq 3\beta$ and $\alpha < 1-\beta$ and $\gamma > 1-\beta$ are, hence, equivalent to the conditions stated in Theorem 2.

As was argued in Lemma 9, no equilibrium is always possible. The same holds for extreme equilibria (Lemma 8).

Having derived both the conditions under which a knife-edge alliance equilibrium exists and those under which an ideology-faithful equilibrium can make up for the lack of such equilibrium, I have considered all combinations that may allow for a moderate equilibrium.[31] Finally, it needs to be assured that no lower bound on q is higher than any upper bound. This has to hold for a non-empty set of values of ρ. By Lemma 11, I can use the upper bound for ρ to express the final and consolidated constraints for the existence of moderate equilibria:
Let $\mathbb{1}$ denote the indicator function, then the consolidated constraints can be expressed as follows:
Consolidated constraints:
$(\alpha, \beta, \gamma) = (0, 0.5, 1)$ or
$\beta - \alpha < \gamma - \beta$ and $\alpha \neq \beta$:

Case 1:
1. $\gamma \leq 3\beta$
2. $\alpha < 1-\beta$ and $\gamma > 1-\beta$
3. $\gamma - 4\alpha + \beta > 0$ and $4\gamma - 2 - \beta - \alpha > 0$

Case 2:
1. $\gamma \leq 3\beta$
2. $\alpha \geq 1-\beta$ or $\gamma \leq 1-\beta$
3. $\alpha < 0.5$ and $\gamma > 0.5$
4.

[31] By Theorem 1, there are no other classes of equilibria than knife-edge alliance, ideology-faithful and extreme equilibria when linearity is satisfied.

$4-2\gamma-3\beta-3\alpha<0$ and

$$\min\left[\frac{2-2\gamma-2\beta}{4-2\gamma-3\beta-3\alpha},1\right]>$$
$$\max\left[\frac{2-\gamma-\beta}{2(1-2\alpha)},\frac{\alpha+\beta}{2(2\gamma-1)},\mathbb{1}(2\beta-\gamma\geq 0)\frac{\gamma-\beta-\alpha}{\alpha+3\gamma-4\beta}\right], \text{ or}$$

$\alpha-4+3\beta+4\gamma<0$ and

$$\min\left[\frac{\gamma+2\beta-2+\alpha}{\alpha-4+3\beta+4\gamma},1\right]>$$
$$\max\left[\frac{2-\gamma-\beta}{2(1-2\alpha)},\frac{\alpha+\beta}{2(2\gamma-1)},\mathbb{1}(2\beta-\gamma\geq 0)\frac{\gamma-\beta-\alpha}{\alpha+3\gamma-4\beta}\right]$$

Case 3:

1. $\gamma>3\beta$
2. $\gamma\geq 9/14$

Case 4:

1. $\gamma>3\beta$
2. $\gamma<9/14$
3. $0\leq\beta\leq 1/6$ and $0.2(3+4\beta)-0.2\sqrt{4-26\beta+61\beta^2}<\gamma$ or $1/6<\beta<3/14$

Case 5:

1. $\gamma\geq 3\beta$
2. $\gamma<9/14$
3. $\alpha<0.5$ and $\gamma>0.5$
4. $$\min\left[\frac{\gamma-2+\alpha}{3\beta+\alpha-2},\frac{\gamma+2\beta-2+\alpha}{\alpha-4+3\beta+4\gamma},1\right]>\max\left[\frac{2-\gamma-\beta}{2(1-2\alpha)},\frac{\alpha+\beta}{2(2\gamma-1)}\right].$$

Games in which $\alpha=\beta=\gamma$ do not have moderate equilibria (by Lemma 1). Games with $\alpha=\beta$ support them only if they satisfy the constraints stated in **Cases 3 or 4**. The case in which $\beta-\alpha>\gamma-\beta$ follows by symmetry. □

The below figure depicts all parameter combinations that satisfy the consolidated constraints.

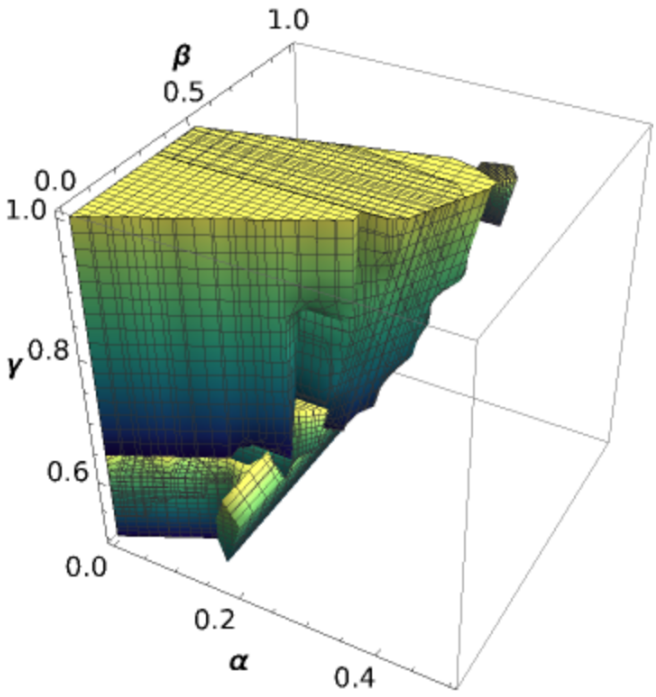

Figure 1.9: Parameter constellations satisfying the constraints in Theorem 4

Lemma 11. *Consider a system of lower and upper bounds on q derived from comparisons of payoffs. It suffices to compare the ordering of bounds for* $\rho = 0.5$ *(maintaining strict inequalities) to determine whether a combination of parameter values can satisfy the the constraints for some* $\rho \in (0, 0.5)$*. Any parameter constellation that satisfies the constraints for some value of* $\rho \in (0, 0.5)$ *satisfies them for* $\rho = 0.5$*.*

Proof. First, note that upper bounds are either of the form (1) $\frac{a-b\rho}{c}$, where $a, b \geq 0$ and $c < 0$ or (2) 2ρ. Lower bounds are of the form $\frac{d-e\rho}{f}$, where $d, e \geq 0$ and $f > 0$. Consider case (1). Comparing $\frac{a-b\cdot 0.5}{c}$ and $\frac{d-e\cdot 0.5}{f}$, I can determine whether there is a continuum of values of ρ for which the constraints can be met simultaneously:

$$\frac{a-b\cdot 0.5}{c} > \frac{d-e\cdot 0.5}{f} \iff \frac{a-b\cdot 0.5}{c} - \frac{d-e\cdot 0.5}{f} > \varepsilon(e/f - b/c) \iff$$
$$\frac{a-b(0.5-\varepsilon)}{c} > \frac{d-e(0.5-\varepsilon)}{f},$$

for some $0.5 > \varepsilon > 0$ and $\rho = 0.5 - \varepsilon$, as $e/f - b/c > 0$ by assumption. There are no combinations of parameter values satisfying the constraints simultaneously for

some value of $\rho \in (0, 0.5)$ that do not satisfy them for $\rho = 0.5$, as the derivation is valid in both directions. Consider now case (2):

$$2 \cdot 0.5 > \frac{d - e \cdot 0.5}{f} \iff 2 \cdot 0.5 - \frac{d - e \cdot 0.5}{f} > \varepsilon(e/f + 2) \iff$$
$$2(0.5 - \varepsilon) > \frac{d - e(0.5 - \varepsilon)}{f},$$

for some $0.5 > \varepsilon > 0$ and $\rho = 0.5 - \varepsilon$, as $e/f + 2 > 0$ by assumption. Again, there are no combinations of parameter values satisfying the constraints simultaneously for some value of $\rho \in (0, 0.5)$ that do not satisfy them for $\rho = 0.5$, as the derivation is valid in both directions. Note that as $e/f + 2 > 0$ and $e/f - b/c > 0$, equality is not admissible here. ρ is strictly bounded above by 0.5, hence equality would imply there not being any value of $\rho \in (0, 0.5)$ satisfying all constraints.

Consider now two lower bounds on q of the forms $\frac{d_1 - e_1 \rho}{f_1}$ and $\frac{d_2 - e_2 \rho}{f_2}$, where $d_1, d_2, e_1, e_2 \geq 0$ and $f_1, f_2 > 0$. If the first bound is greater than the second for $\rho = 0.5$, it suffices to verify that first bound is smaller than the smallest upper bound for $\rho = 0.5$ to make sure the second bound is satisfied as well for some value of $\rho \in (0, 0.5)$

$$\frac{d_1 - e_1 \cdot 0.5}{f_1} - \frac{d_2 - e_2 \cdot 0.5}{f_2} > 0 \iff \frac{d_1 - e_1 \cdot 0.5}{f_1} - \frac{d_2 - e_2 \cdot 0.5}{f_2} >$$
$$> \varepsilon(e_2/f_2 - e_1/f_1) \iff \frac{d_1 - e_1(0.5 - \varepsilon)}{f_1} > \frac{d_2 - e_2(0.5 - \varepsilon)}{f_2}$$

for some $0.5 > \varepsilon > 0$ and $\rho = 0.5 - \varepsilon$. This holds trivially if $e_2/f_2 - e_1/f_1 < 0$. It also holds for some value of ε if $e_2/f_2 - e_1/f_1 \geq 0$, as the first inequality was strict. Therefore, there is some value of $\rho \in (0, 0.5)$ for which both bounds are smaller than the smallest upper bound if the first lower bound is smaller than the smallest upper bound for $\rho = 0.5$ when $\frac{d_1 - e_1 \cdot 0.5}{f_1} - \frac{d_2 - e_2 \cdot 0.5}{f_2} > 0$. Hence, it suffices to determine whether the first bound is smaller than the smallest upper bound for $\rho = 0.5$ to ensure the existence of some $\rho \in (0, 0.5)$ for which all constraints are satisfied at the same time. As the first lower bound needs to be smaller than the smallest upper bound for $\rho = 0.5$ (see reasoning above), there are no parameter constellations for which all requirements are met for $\rho \in (0, 0.5)$ but not for $\rho = 0.5$ if $\frac{d_1 - e_1 \cdot 0.5}{f_1} - \frac{d_2 - e_2 \cdot 0.5}{f_2} > 0$ and the first bound is greater than the smallest upper bound for $\rho = 0.5$. An analogous argument can be made for upper bounds.

To conclude, it is both necessary and sufficient to consider the relative ordering of constraints for $\rho = 0.5$ to determine whether there is a continuum of values of $\rho \in (0, 0.5)$ for which all constraints are met simultaneously. □

References

Abramowitz, Alan I., and Steven Webster. 2016. "The rise of negative partisanship and the nationalization of U.S. elections in the 21st century." *Electoral Studies*, 41(1): 12–22.

Arzheimer, Kai, and Carl C. Berning. 2019. "How the Alternative for Germany (AfD) and their voters veered to the radical right, 2013–2017." *Electoral Studies*, 60(1): 102040.

Cahan, Dodge, and Arkadii Slinko. 2018. "Electoral competition under best-worst voting rules." *Social Choice and Welfare*, 51(2): 259–279.

Calvert, R. 1985. "Robustness of the Multidimensional Voting Model: Candidate Motivations, Uncertainty, and Convergence." *American Journal of Political Science*, 29(1): 69–95.

Caruana, Nicholas J., R. Michael McGregor, and Laura B. Stephenson. 2015. "The Power of the Dark Side: Negative Partisanship and Political Behaviour in Canada." *Canadian Journal of Political Science*, 48(4): 771–789.

Casalecchi, Gabriel Avila, Andre Borges, and Lucio Renno. 2020. "Generalized anti-partisans, conservative and moderate antipetistas: unpacking Bolsonaro's vote in Brazil's 2018 elections." *Working Paper*.

Davis, Stuart, and Joe Straubhaar. 2020. "Producing Antipetismo: Media activism and the rise of the radical, nationalist right in contemporary Brazil." *The International Communication Gazette*, 82(1): 82–100.

do Amaral, Oswaldo. 2020. "The victory of Jair Bolsonaro according to the Brazilian Electoral Study of 2018." *Brazilian Political Science Review*, 14(1): 1–13.

Downs, Anthony. 1957. "An Economic Theory of Political Action in a Democracy." *Journal of Political Economy*, 65(2): 135–150.

Duggan, John, and Mark Fey. 2005. "Electoral competition with policy-motivated candidates." *Games and Economic Behavior*, 51(2): 490–522.

Fuks, Mario, Ednaldo Ribeiro, and Julian Borba. 2021. "From Antipetismo to Generalized Antipartisanship: The Impact of Rejection of Political Parties on the 2018 Vote for Bolsonaro." *Brazilian Political Science Review*, 15(1): 1–28.

Hotelling, Harold. 1929. "Stability in Competition." *The Economic Journal*, 39(153): 41–57.

Maggiotto, Michael A, and James E Piereson. 1977. "Partisan Identification and Electoral Choice: The Hostility Hypothesis." *American Journal of Political Science*, 21(4): 745–767.

Mayer, Sabrina Jasmin. 2017. "How negative partisanship affects voting behavior in Europe: Evidence from an analysis of 17 European multi-party systems with proportional voting." *Research and Politics*, 4(1): 1–7.

Medeiros, Mike, and Alain Noël. 2014. "The Forgotten Side of Partisanship: Negative Party Identification in Four Anglo-American Democracies." *Comparative Political Studies*, 47(7): 1022–1046.

Osborne, Martin J. 1993. "Candidate Positioning and Entry in a Political Competition." *Games and Economic Behavior*, 5(1): 133–151.

Ronayne, David. 2018. "Extreme idealism and equilibrium in the Hotelling-Downs model of political competition." *Public Choice*, 176(1): 389–403.

Samuels, David J., and Cesar Zucco. 2018. "Partisans, Antipartisans, and Nonpartisans: Voting Behavior in Brazil." *Cambridge University Press*.

Chapter 2

Eliciting information from multiple experts via grouping*

Joint with Philipp Hamelmann

2.1 Introduction

Conflicts of interest and information asymmetries are at the core of many problems that economic research seeks to address; including principal-agent problems, voting schemes, markets for lemons and a plethora of other contributions. In most applications, resulting inefficiencies can be ameliorated by either commitment to a certain mechanism (e.g. contracts) or monetary transfers aligning incentives. However, these tools are not always available: in many environments, monetary transfers are either not feasible (budget) or not admissible (regulation, corruption-prevention, ethical reasons). Mechanisms that do not require the above are therefore very attractive. This paper elaborates on a mechanism suggested by Wolinsky (2002) that can alleviate the conflict of interest between a decision maker (DM) and a set of experts without the use of monetary transfers or commitment.

The underlying model considers a typical principal-(multiple-)agent problem with misaligned incentives and information asymmetry: Each of several experts possesses some noisy, private information pertaining to the unknown state of the world. Jointly, their information determine whether or not some proposal should be accepted. Note that, in the context of this model, "determine" need not be understood as causative, but merely correlative; that is, we remain agnostic as to whether the experts' information affect the value of the policy, or simply stand in one-to-one

*Funding by the Bonn Graduate School of Economics (BGSE) and the German Research Foundation (DFG) through CRC TR 224 (project B03) is gratefully acknowledged. This project profited from valuable comments by Sven Rady, Dezsö Szalay, Stephan Lauermann, Daniel Krähmer, Julius Kappenberg and the participants of the CRC retreat in October 2023.

correspondence with the factors that do so. The experts prefer the policy over the status quo if the sum of their signals exceeds some fixed threshold; that is, the experts share a common preference.

By contrast, the DM who ultimately decides on the proposal does not receive any private information—she relies purely on experts' recommendations. Moreover, her threshold for accepting the proposal differs from that of the experts, creating a conflict. As abstract this set-up might seem, it is pertinent to many real-world applications. To elaborate, consider the following example: The CEO of a company considers switching to a new software. To learn whether this would lead to an increase in efficiency, she consults the employees, who—being the users of said software—are better informed about the effect of the switch. However, while the CEO only considers the change in output, the employees prefer not having to invest time into learning to use the new software unless the gains in efficiency are substantial. When the conflict of interest is severe and employees are not able to share their information among one another (no communication), they never recommend the switch: their own information is not sufficient for them to be sure the new software is worth the costs of adapting to it; accordingly, they *always* choose the "conservative" action and discourage the CEO from changing the status quo. Hence, the CEO does not gain from consulting them. In case the CEO gives the employees a platform to privately discuss their information (full communication), they recommend the switch if and only if the software is sure to benefit themselves—though not necessarily the CEO. Accordingly, neither no nor full communication are able to resolve the conflict of interest.

Surprisingly maybe, for many parametrisations, the "intermediate" case (partial communication) can improve upon the two extrema: by allowing her employees to communicate within smaller groups (but not between them), she may be able to influence what they can infer in the event of being pivotal and, hence, elicit more information. This mechanism (proposed by Wolinsky, 2002) makes use of the fact that voters, if rational, condition their actions on being pivotal; that is, they are only concerned with situations wherein "their vote matters". In particular, they deduce the information (and corresponding actions) that other players must have received in order for such situation to arise. Grouping, if chosen optimally, changes the "pivotal information set"[1] and reveals just enough information such that the employees do not always vote against the proposal, yet not enough for them to enforce only outcomes that are optimal from their point of view.

Such a, as we call it, "grouping mechanism" (GM) requires neither commitment nor monetary transfers and is therefore broadly applicable. While proposing partial

[1] That is, the information set at which a vote is pivotal.

communication as a potential remedy to such conflict of interest and discussing a number of examples thereof, Wolinsky (2002) does not go into much detail. We elaborate on his findings and provide a closer analysis by, inter alia, characterising equilibria induced by grouping mechanisms and providing conditions under which grouping can improve upon full communication.

This paper's contributions are threefold: First, we characterise outcomes of games with more general signal distributions than those considered in Wolinsky (2002): he shows that no communication never results in a change of the status quo if signals are Bernoulli trials and the conflict of interest sufficiently high; full communication leads to adoption if and only if the policy is optimal for the experts. We provide similar results for a broader set of signal distributions; however, note that we consider less general utility functions than Wolinsky (2002) does in the first part of his paper.

Second, we further characterise grouping mechanisms similar to those suggested by Wolinsky (2002) and discuss the relationship between group sizes, the conflict of interest and the degree to which information can be elicited (Proposition 1). Thereafter, we show that grouping can only improve upon full communication if it implies a higher probability of adoption; otherwise, the "safe option"[2] of full communication yields a higher expected utility than the GM.

Third, we elaborate on how and when specific grouping mechanisms may benefit the DM: In "expected value grouping mechanisms" (μ-GMs), experts are divided into groups and asked to submit a positive vote if the sum of signals within the group is higher than expected, while the DM announces the number of positive votes needed for her to choose the policy. We characterise the optimal group size for such μ-GMs given normally distributed signals (Proposition 3). Thereafter, we discuss the conditions under which such grouping can improve upon the outcome of full communication (Proposition 4) and show that, qualitatively, the insights generalise to GMs with arbitrary group-thresholds (Proposition 5). This implies our focusing on μ-GMs to serve a reasonable and not overly limited simplification. Thereafter, analysing the case in which the DM partitions experts into two groups, we provide conditions under which even such simple grouping can improve upon full communication (Lemma 13) and assess how changes in the conflict of interest and the gains from the policy affect the benefits thereof (Proposition 6). Lastly, we show that the fundamental idea behind grouping mechanisms does not hinge on the groups being equally sized; that is, grouping can be beneficial for the DM even if there is a *prime* number of experts (Proposition 7).

[2]Full communication can be understood as the "safe option" as it never leads to adoption of unprofitable policies (adoption if and only if the policy is optimal for the experts).

The remainder of this paper is structured as follows: In Section 2.2, we describe the model; Section 2.3 constitutes the main part of the analysis wherein we discuss the basics of the game and provide first hints as to why grouping may be beneficial (Section 2.3.1). Thereafter, we characterise grouping mechanisms (Section 2.3.2) and determine optimal expected value grouping mechanisms given normally distributed signals along with the conditions under which such grouping improves upon full communication (Section 2.3.3). Section 2.3.4 shows that our restricting attention to μ-GMs does not alter results qualitatively. The section may be skipped by readers who are satisfied with the analyses obtained in the previous sections and prefer not to delve into the technical details of more general grouping mechanisms. In Section 2.3.5, we provide conditions under which even a partition of experts into two groups is preferable for the DM and analyse comparative statics. Lastly, we show that the requirement of groups being equally sized is not the driver of our results (Section 2.3.6). Section 2.4 relates our work to the literature; Section 2.5 concludes.

2.2 Model

A decision maker (DM; she) has to decide whether or not to adopt a new policy. The desirability of said policy depends on an unknown multi-dimensional state of the world, $\boldsymbol{s} = (s_1, ..., s_N) \in S \subseteq \mathbb{R}^N$. To inform her decision, she consults N experts (he/they), each of whom possesses knowledge about one dimension of the state-space; that is, expert $i \in \{1, ..., N\}$ receives "signal" s_i. The state is realised according to some joint distribution $f_S : \mathcal{S} \to \mathbb{R}$, where $\mathcal{S}$ is a sigma-algebra on S, with identical and independent marginal distributions.[3] Therefore, the experts' signals, s_i, are distributed iid according to f_{S_i}.

The utilities of both the DM and experts, depend on the sum over elements of $\boldsymbol{s}$, denoted $\boldsymbol{s}^\Sigma$, such that the experts' signals[4] may be thought of as cumulative evidence in favour of the positive impact of the policy. While both types of players are thus more inclined towards adoption if $\boldsymbol{s}^\Sigma$ is large, they differ with respect to the thresh-

[3] Wolinsky (2002) assumes signals to be binary ($s_i \in \{0, 1\}$).

[4] As the desirability of the policy is determined by the sum of signals $\boldsymbol{s}^\Sigma$, one could argue the term "signal" to be misleading. Alternatively, the signals could be thought of as "substates" that jointly determine the decision-relevant variable. For instance (cf. example in Section 2.1), if the decision relevant quantity is the number of a company's departments that profit from a new software and the experts are the respective heads of department, every head of department possesses information about the desirability of the change; jointly the heads of department are perfectly informed. Despite this shortcoming of the term "signal", we decided to adopt it and generally deviate as little as possible from the notation used in Wolinsky (2002) to avoid confusion.

old whereat they prefer doing so over maintaining the status quo; in particular the DM's and expert's respective utilities are:

$$\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma}) = \begin{cases} U_{\text{DM}}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - \alpha + \delta, & \text{if policy adopted,} \\ 0, & \text{if status quo maintained and} \end{cases}$$

$$\tilde{U}_{\text{ex}}(\boldsymbol{s}^{\Sigma}) = \begin{cases} U_{\text{ex}}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - \alpha, & \text{if policy adopted,} \\ 0, & \text{if status quo maintained.} \end{cases}$$

In order not to render the problem trivial, we assume that $\alpha > \delta > 0$, $\mathbb{E}[\boldsymbol{s}^{\Sigma}] - \alpha + \delta < 0$ and $\Pr(\boldsymbol{s}^{\Sigma} \geq \alpha) > 0$, such that it is possible for both types of players to prefer adoption of the policy, while in the absence of additional evidence, the DM maintains the status quo. Consequently, consulting the expert may sway the DM one way or the other. For sufficiently large $\boldsymbol{s}^{\Sigma}$, both the experts and the DM prefer the policy, with the latter's threshold being lower (by $\delta > 0$); that is, she is more eager to adopt the policy. Both f_S and players' utility functions are common knowledge.

The game proceeds as follows:

1. The state $\boldsymbol{s}$ is realised and each expert is privately informed about "his" dimension (i.e. i receives signal s_i).
2. The set of experts is partitioned into equally sized groups $g_1, ..., g_m$.[5] Signals can be shared within but not across groups.
3. Each group g_i decides whether to vote for ($v_{g_i} = 1$) or against the policy ($v_{g_i} = 0$), according to strategy $y_{g_i}(\boldsymbol{s}_{g_i})$.
4. The DM receives the vector of groups' votes $\boldsymbol{v}$, forms (Bayesian) beliefs regarding the state $\boldsymbol{s}$ and chooses whether to adopt the policy, according to strategy $x(\boldsymbol{v})$.
5. Payoffs are realised.

Note that it is without loss to consider groups' (rather than individuals') strategies, as all experts possess identical preferences.

The solution concept we employ is that of pure-strategy[6], symmetric Bayesian

[5] As will be shown in Section 2.3.6, the requirement of groups being equally sized is not the driver of our results.

[6] In case of indifference, the DM and experts are assumed to choose and vote in favour of the policy, respectively.

Nash Equilibrium, with optimal strategies $y^*_{g_i}(\boldsymbol{s}_{g_i}, x, \boldsymbol{y}_{-g_i})$ and $x^*(\boldsymbol{v}, \boldsymbol{y})$ for expert-groups and the DM, respectively. Lastly, let $\boldsymbol{s}^{\Sigma}_{g_i}$ be the sum of signals in group g_i and define a grouping mechanism:

Definition 3. *A **grouping mechanism** (GM) consists of*

1. *$2 \leq m \leq N$ groups $\{g_1, ..., g_m\}$ of equal size n_m,*

2. *a threshold t such that each group g_i is asked to vote $v_{g_i} = 1$ if and only if $t \leq \boldsymbol{s}^{\Sigma}_{g_i}$, where $0 < Pr(t \leq \boldsymbol{s}^{\Sigma}_{g_i})$ and*

3. *a threshold $0 < V \leq m$, announced by the DM, such that the policy is chosen if and only if $V \leq \boldsymbol{v}^{\Sigma} := \sum_i v_{g_i}$.*

We refer to a GM with m groups, threshold values V and t as GM(m, V, t).
*GM(m, V, t) is said to be **implementable** if:*

$$x^*(\boldsymbol{v}, \boldsymbol{y}^*) = 1 \iff V \leq \boldsymbol{v}^{\Sigma} \text{ and}$$
$$\text{for all } g_i\text{: } y^*_{g_i}(\boldsymbol{s}_{g_i}, x^*, \boldsymbol{y}^*_{-g_i}) = 1 \iff t \leq \boldsymbol{s}^{\Sigma}_{g_i}.$$

The above implies: in the equilibrium induced by an implementable GM, DM and experts play a threshold strategy. As shown in Lemma 12, considering symmetric equilibria, this restriction is without loss of generality.

Note that, as is the case in Wolinsky (2002), signals are verifiable insofar that group g_i cannot vote/"report" $v_{g_i} = 1$ if $\boldsymbol{s}^{\Sigma}_{g_i} < t$; nevertheless, it may vote $v_{g_i} = 0$ even if $t \leq \boldsymbol{s}^{\Sigma}_{g_i}$. Put differently, experts are able to hide/omit evidence suggestive of a profitable policy. They are, however, not able to make up any such evidence.[7]

An equilibrium with threshold values (V, t) can be understood as a simple recommendation procedure: the DM announces values V and t whereafter the experts recommend/vote for the policy ($v_{g_i} = 1$) if and only if $\boldsymbol{s}^{\Sigma}_{g_i} \geq t$. Clearly, (V, t) can only form an equilibrium if said thresholds are in fact optimal for the DM and the experts, respectively.[8] Hence, an implementable GM requires neither commitment nor transfers (by definition).

[7] Wolinsky (2002) justifies the assumption by writing: "Since the experts are less eager than DM, they naturally would not have an interest in exaggerating their reports." This argument may sound compelling but does not hold in general. For very high signal realisations, groups can profit from acting as if they had more evidence for the policy than they actually do; that is, to make sure the policy is adopted, they would prefer "exaggerating". Despite this weakness, we decided to adopt the assumption as it significantly reduces the model's complexity: the number of cases in which a group's vote may be pivotal decreases. If, for instance, s_i is normally distributed, experts may be pivotal in infinitely many ways for any signal realisation.

[8] Note that it is sufficient to determine whether groups are willing to recommend the policy if $t \leq \boldsymbol{s}^{\Sigma}_{g_i}$, as a vote for the policy in case $\boldsymbol{s}^{\Sigma}_{g_i} < t$ is not possible: we (as does Wolinsky, 2002) preclude "exaggerated" reports/votes by assumption.

2.3 Main analysis and results

2.3.1 Basics of the game

Before delving into the details of different group sizes and the games' outcomes, we derive basic properties of players' equilibrium strategies. The proof of Lemma 12 deepens the understanding of the game and lays a good ground for later discussions.

Lemma 12. *For any number of groups and parameter values, both DM and experts play a threshold strategy; that is, for some V and t:*

$$x^*(\boldsymbol{v}, \boldsymbol{y}^*) = 1 \iff V \leq \boldsymbol{v}^{\Sigma} \qquad x^*(\boldsymbol{v}, \boldsymbol{y}^*) = 0 \text{ otherwise.}$$
$$\forall g_i : \; y^*(\boldsymbol{s}_{g_i}, x^*, \boldsymbol{y}^*_{-g_i}) = 1 \iff t \leq \boldsymbol{s}^{\Sigma}_{g_i} \qquad y^*(\boldsymbol{s}_{g_i}, x^*, \boldsymbol{y}^*_{-g_i}) = 0 \text{ otherwise.}$$

Proof.
DM: Naturally, the DM chooses the policy if and only if her updated expected utility of doing so is non-negative. As experts' (and groups') signals are iid and we are considering symmetric equilibria, we can write:

$$\begin{aligned} x^*(\boldsymbol{v}, \boldsymbol{y}^*) = 1 \iff \alpha - \delta &\leq \mathbb{E}[\boldsymbol{s}^{\Sigma} \mid \boldsymbol{v}, \boldsymbol{y}^*] \\ &= \textstyle\sum_i \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid v_{g_i}, y^*_{g_i}] \\ &= \boldsymbol{v}^{\Sigma}\, \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid v_{g_i} = 1, y^*] + (m - \boldsymbol{v}^{\Sigma}) \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid v_{g_i} = 0, y^*], \end{aligned}$$

where y^* is the (symmetric) optimal strategy employed by all groups (i.e. $y^*_{g_i} = y^* \; \forall g_i$). Now let V be the smallest number of positive votes such that for $\boldsymbol{v}^{\Sigma} = V$, the above inequality holds. Then, the DM chooses the policy if and only if $\boldsymbol{v}^{\Sigma} \leq V$. This establishes the first part of the Lemma.
Experts: Upon receipt of signals $\boldsymbol{s}_{g_i}$, each group must decide whether or not to vote $v_{g_i} = 1$. As all experts have the same utility function, the process of taking this decision does not need to be modelled in detail: even if a single group member were to determine the action, none of the other members would object. As group g_i's vote only matters in case it is pivotal, the group conditions on being pivotal when determining its action; in this case (as all other groups' strategies are identical and $v \in \{0, 1\}$), there is one[9] such scenario; that is, the sum of other groups' votes $\boldsymbol{v}^{\Sigma}_{-g_i} = V - 1$. Naturally, the group recommends the policy if and only if the expected

[9] Of course, there are many ways in which $V - 1$ groups can submit a positive vote. All of them are equivalent with respect to the information experts can elicit from being pivotal.

value of adoption is non-negative:

$$\begin{aligned} & y^*_{g_i}(\boldsymbol{s}_{g_i}, x^*, \boldsymbol{y}^*_{-g_i}) = 1 \iff \\ & \quad \alpha \leq \mathbb{E}[\boldsymbol{s}^{\Sigma} \mid \boldsymbol{v}^{\Sigma}_{-g_i} = V-1, \boldsymbol{s}_{g_i}, \boldsymbol{y}^*_{-g_i}] \\ & \qquad = (V-1)\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_j} \mid v_{g_j} = 1, y^*_{g_j}] + (m-V)\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_j} \mid v_{g_j} = 0, y^*_{g_j}] + \boldsymbol{s}^{\Sigma}_{g_i} \end{aligned}$$

where $y^*_{g_j}$ is the optimal strategy of every group other than g_i (i.e. g_j is the typical "other group").[10] Again, let t be the smallest value of $\boldsymbol{s}^{\Sigma}_{g_i}$ for which the above inequality holds: group g_i votes $v_{g_i} = 1$ if and only if $t \leq \boldsymbol{s}^{\Sigma}_{g_i}$, as was to be shown. □

The very basics of the game established, we can proceed by analysing the two extrema with respect to grouping; that is, $m = 1$ (full communication) and $m = N$ (no communication). Wolinsky (2002) uses these two variants of the game as the benchmarks for his analysis and seeks to improve upon them. Similarly, we use the following results to better judge the performance of grouping mechanisms.

First, consider full communication: what happens if experts are allowed to share the value of their signals with one another before voting? The answer is simple: as the sum of all experts' signals fully determines whether the policy is profitable (i.e. the state of the world), experts encourage adoption if and only if $\boldsymbol{s}^{\Sigma} \geq \alpha$. As $\mathbb{E}[\boldsymbol{s}^{\Sigma} \mid \boldsymbol{s}^{\Sigma} < \alpha] \leq \mathbb{E}[\boldsymbol{s}^{\Sigma}] < \alpha - \delta$, it is optimal for the DM to follow the experts' advise. Hence, the policy is chosen if and only if it is profitable for the experts:

Observation 1. *For $m = 1$ (full communication), $V = 1$ and $t = \alpha$.*
Accordingly, the policy is chosen if and only if it is profitable for the experts.

Thus, full communication never results in adoption of a policy that is profitable for the DM but not the experts; experts are given full power over the outcome of the game.

As pessimistic as this result sounds, it may be better than the other extreme case, to wit, no communication: for binary signals (Bernoulli) and $\delta > 1$ (model in Wolinsky, 2002), no communication never results in the adoption of the policy. Trivially, this finding makes any mechanism that allows for adoption of the policy for *some* signal realisations with $\boldsymbol{s}^{\Sigma} > \alpha - \delta$ superior to autarky. Note that as we consider a broader set of signal distributions, this statement cannot be said to hold in general. Observation 2 represents "our version" of Wolinsky (2002)'s finding:

[10] As we are considering symmetric equilibria, $y^*_{g_j}$ is the strategy of all other groups.

Observation 2. *For any number of groups $m \in \{2,...,N\}$ and non-negative signals (i.e. $S_i \subseteq \mathbb{R}^+$), the policy is never chosen if*

$$\frac{N\bar{s}}{m} \leq \delta,$$

where $\bar{s} := \max_{s_i \in S_i}\{s_i\}$.

Clearly, the above implies Wolinsky (2002)'s finding: for $m = N$, $\bar{s} = 1$ and $\delta > 1$, the condition is satisfied and the policy is never chosen.

To see why, consider the following arguments: Experts understand that it is sufficient to decide which action to take in case a switch from $v = 0$ to $v = 1$ changes the DM's decision. Consider any number of groups greater than one and take group g_i with signal $\boldsymbol{s}^{\Sigma}_{g_i}$.[11] Note that as we consider symmetric equilibria, $\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} \geq t, \boldsymbol{y}^*] = \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_j} \mid \boldsymbol{s}^{\Sigma}_{g_j} \geq t, \boldsymbol{y}^*]$ for all j, $i \in \{1,...m\}$ (analogously for $\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} < t, \boldsymbol{y}^*]$); pivotality implies

$$\begin{aligned}
& x^*(\boldsymbol{v}_{-g_i}, v_{g_i} = 1, \boldsymbol{y}^*) = 1 > x^*(\boldsymbol{v}_{-g_i}, v_{g_i} = 0, \boldsymbol{y}^*) = 0 \\
\iff & V\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} \geq t, \boldsymbol{y}^*] + (m-V)\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} < t, \boldsymbol{y}^*] - \alpha + \delta \geq 0 \\
& > (V-1)\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} \geq t, \boldsymbol{y}^*] + (m-V+1)\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} < t, \boldsymbol{y}^*] - \alpha + \delta \\
\iff & -\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} \geq t] - \delta + \boldsymbol{s}^{\Sigma}_{g_i} \leq \mathbb{E}[U_{\text{ex}}(\boldsymbol{s}^{\Sigma}) \mid \boldsymbol{v}^{\Sigma}_{-g_i} = V-1, \boldsymbol{s}^{\Sigma}_{g_i}, \boldsymbol{y}^*] \\
& < -\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} < t, \boldsymbol{y}^*] - \delta + \boldsymbol{s}^{\Sigma}_{g_i}.
\end{aligned}$$

Accordingly, $\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid v_{g_i} = 0, \boldsymbol{y}^*] + \delta < \boldsymbol{s}^{\Sigma}_{g_i}$ is a necessary condition for a positive vote in any symmetric equilibrium. If now $S_i \subseteq \mathbb{R}^+$ and $(N\bar{s})/m < \delta$, the condition is violated and votes are always zero. If this is the case, the DM cannot infer any further information, $\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid v_{g_i} = 0, \boldsymbol{y}^*] = \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i}]$ and, as $m\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i}] - \alpha - \delta < 0$, the policy is never adopted.

Put differently, to make a positive recommendation, the experts must have acquired enough information to be sure the conflict of interest (δ) is overcome *when their vote is pivotal*: were this not the case, they would encourage the DM to choose a policy that is not profitable for themselves (conflict of interest too high).

Besides providing a more general version of Wolinsky (2002)'s finding for no communication, the result also yields insights as to why grouping may enhance information revelation:
(1) The larger the groups, the more each group can infer about the state of the

[11] As noted in Section 2.2 and the proof of Lemma 12, it is without loss of generality to consider each group's sum of signals $\boldsymbol{s}^{\Sigma}_{g_i}$ instead of each expert's signal.

world from its *own* signals. This may, as is the case under full communication, give the experts too much power such that they can control the outcome.
(2) However, if done correctly, grouping can also give them just enough information to be willing to recommend the policy in some cases but not enough to fully control the game. This stems from the fact that, at the pivotal information set, experts are less informed about *other* groups' signals: the necessary condition $\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid v_{g_i} = 0, \boldsymbol{y}^*] + \delta < \boldsymbol{s}_{g_i}^{\Sigma}$ is, if t and m are chosen correctly, not as tight anymore and experts are not able to infer as much about the the total sum of signals (i.e. the state of the world). For a detailed illustration of a simple example in which grouping improves upon both full and no communication, see Example 5 in the appendix.

To sum up, in this section, we established the benchmarks upon which we seek to improve and provided first hints as to how grouping may be beneficial for the DM. In the following section, we take a more technical and detailed perspective to further characterise GMs and the equilibria induced by them.

2.3.2 Characterisation of grouping mechanisms

In this section, we take a closer look at the details of GMs and their properties to better understand how and why they may improve upon full (and in Wolinsky (2002)'s model, no) communication.

Note that in any equilibrium induced by an implementable GM, the DM assumes all groups that voted $v_{g_i} = 0$ to have $\boldsymbol{s}_{g_i}^{\Sigma} < t$ and, hence, estimates $\boldsymbol{s}_{g_i}^{\Sigma}$ by $\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid v_{g_i} = 0, \boldsymbol{y}^*] = \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]$ (analogously for $v_{g_i} = 1$). Accordingly, we omit the former expression conditioning on strategies and actions and directly refer to the implied expected value in terms of the signal realisations (latter expression).

Proposition 1. *A combination (m, V, t) corresponds to an implementable GM(m, V, t) if and only if*

$$V = \lceil \tilde{V} \rceil := \left\lceil \frac{\alpha - \delta - m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]} \right\rceil, \tag{2.1}$$

$$\tilde{V} + \frac{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - t + \delta}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]} \leq V, \tag{2.2}$$

$$\alpha - \delta \leq m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t], \tag{2.3}$$

$$2 \leq m \text{ and } N/m \in \mathbb{N}. \tag{2.4}$$

The first statement of Proposition 1 characterises V—the minimum number of positive votes leading to adoption of the policy. Consider the fraction within the ceiling function, namely $\tilde{V}$ and rearrange it to yield:

$$\tilde{V}\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] + (m - \tilde{V})\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t] = \alpha - \delta$$

$\tilde{V}$ is the exact number of positive votes that makes the DM indifferent between the policy and the status quo. If $\tilde{V} \notin \mathbb{N}$, the actual number of positive votes needed for adoption is slightly higher ($V = \lceil \tilde{V} \rceil$), as a number of $\lfloor \tilde{V} \rfloor$ would not be sufficient evidence for the DM.

Besides that, V must also be such that experts are willing to vote for the policy upon receipt of $\boldsymbol{s}_{g_i}^{\Sigma} \geq t$. Hence, the second statement ensures the DM's and experts' constraints to be compatible.

Accordingly, the first two requirements show that implementable GMs do, by definition, neither require commitment nor transfers.

The third and fourth statements ensure that groups are equally sized and $V \leq m$; were $V > m$, the policy would never be adopted and grouping inferior to full communication.

Jointly, the conditions in Proposition 1 are thus necessary and sufficient as they ensure V and t to be incentive compatible, best responses and comply with the fundamental assumptions on GMs and the model.

Corollary 1. *In any implementable GM*(m, V, t)*,*

$$\delta < t - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]. \tag{2.5}$$

The necessary condition implied by Corollary 1 is consistent with the findings obtained in Observation 2: a positive vote is only possible if the evidence of the group is strong enough to resolve the conflict of interest (δ) when its vote is pivotal. The corollary follows from incentive compatibility: in equilibrium, the DM assumes all groups that voted $v_{g_i} = 0$ to have $\boldsymbol{s}_{g_i}^{\Sigma} < t$ and, hence, estimates $\boldsymbol{s}_{g_i}^{\Sigma}$ by $\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*] = \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]$. Accordingly, the additional policy-favourable evidence of a positive vote by group g_i for $\boldsymbol{s}_{g_i}^{\Sigma} = t$ is equal to $t - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]$. If group g_i's vote $v_{g_i} = 1$ is pivotal, it changes the DM's expected utility from being negative (for $v_{g_i} = 0$) to being positive. For the experts to be willing to vote $v_{g_i} = 1$ in this case, the change in expected utility must be greater than the conflict of interest and, hence, $\delta < t - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]$.

Besides providing a necessary condition for implementability, this is also suggestive as to how GMs may be able to improve upon no communication. To illustrate,

consider the following example:

Example 1. $N = 10$, $\delta = 1.25$, $Pr(s_i = 1) = 0.5 = Pr(s_i = 0)$.

As shown in Observation 2, for $m = 10$, the necessary condition for implementation stated above cannot be satisfied and $\Pr(x = 1) = 0$, as $\delta = 1.25 > s_i$ for all possible realisations s_i.
By decreasing the number of groups, the condition can be loosened: now, the *sum* of the group's members' signals needs to satisfy the inequality in Corollary 1. Instead of there being no t with $Pr(\boldsymbol{s}_{g_i}^{\Sigma} \geq t) > 0$ for which it holds, there may now be multiple such values. Take, for instance $m = 2$:

$$
\begin{aligned}
t = 2: &\quad \delta = 1.25 < 2 - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < 2] = 2 - 1 \cdot 0.5^5 \cdot 5 \approx 1.84 \\
t = 4: &\quad \delta = 1.25 < 4 - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < 4] \approx 2.28 \\
t = 5: &\quad \delta = 1.25 < 5 - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < 5] \approx 2.65
\end{aligned}
$$

Hence, by appropriate choice of m and t, the DM is able to loosen the requirement to an extent that allows her to elicit significantly more information than in the absence of communication. This stems from the fact that allowing for communication within groups changes the information experts can infer about the state of the world at the pivotal information set. In the example above, the information is too rich for $m = N$ and experts never vote for the policy. For $m = 2$, on the other hand, experts are not able to fully determine whether the policy is profitable and votes for the policy are possible.

So far, we have characterised implementable GMs and provided intuition as to why they may be able to improve upon both full and no communication: simply put, GMs are the perfect combination of both extrema; experts are given enough information to recommend adoption upon receipt of sufficient evidence, but, as communication is limited, they are not given full knowledge of the state of the world and cannot control the game's outcome perfectly. If chosen correctly, the threshold values and group sizes are, hence, able to allow for adoption of policies that are only profitable for the DM—not the experts; accordingly, such values (V, t) imply a relatively high probability of adoption. Given full communication, on the other hand, the policy is chosen if and only if it is profitable for both DM and experts. Put differently, full communication is the "safe" option wherein adoption of highly profitable policies ($\boldsymbol{s}^{\Sigma} \geq \alpha$) is *guaranteed*; this comes at a cost: under full communication, the total probability of adoption is relatively small. The following proposition formalises this finding and shows that a GM can only improve upon full

communication if it increases the probability of implementation. Otherwise, the DM is better off choosing the safe alternative, to wit, full communication, which, contrary to a GM, leads to guaranteed adoption of profitable policies.

Definition 4. *We write* $\boldsymbol{GM \succ_{DM}}$ ***full comm.*** *if the GM yields higher ex-ante expected utility for the DM than full communication.*

Let Pr(adopt | GM) and Pr(adopt | full comm.) denote the probability of adoption of the policy given a GM and full communication, respectively; then:

Proposition 2.

$$GM \succ_{DM} \textit{full comm.} \;\Rightarrow \textit{Pr(adopt} \mid \textit{GM)} > \textit{Pr(adopt} \mid \textit{full comm.)}$$

Proof sketch. The proposition is proven by contraposition. Let A_{full} (resp. A_{GM}) be the subset of S^N, such that under full communication (resp. GM), the policy is adopted if and only if $\boldsymbol{s} \in A_{\text{full}}$ (resp. $\boldsymbol{s} \in A_{GM}$). The DM's expected utility is:

$$\begin{aligned} E[\tilde{U}_{\text{DM}}|\text{com}] &= E[\tilde{U}_{\text{DM}}|\boldsymbol{s} \in A_{\text{com}}]\Pr(\boldsymbol{s} \in A_{\text{com}} \\ &= \frac{\int_{\boldsymbol{s}\in A_{\text{com}}} (s^{\Sigma} - \alpha + \delta) f(\boldsymbol{s}) d\boldsymbol{s}}{\Pr(\boldsymbol{s} \in A_{\text{com}})} \Pr(\boldsymbol{s} \in A_{\text{com}}) \\ &= \int_{\boldsymbol{s}\in A_{\text{com}}} s^{\Sigma} f(\boldsymbol{s}) d\boldsymbol{s} + \Pr(\boldsymbol{s} \in A_{\text{com}})(-\alpha + \delta) \end{aligned}$$

Therefore, the difference in expected utilities can be written as follows:

$$\begin{aligned} &E[\tilde{U}_{\text{DM}}|\text{full}] - E[\tilde{U}_{\text{DM}}|\text{GM}] = \\ &\int_{\boldsymbol{s}\in A_{\text{full}}} s^{\Sigma} f(s) ds - \int_{\boldsymbol{s}\in A_{\text{GM}}} s^{\Sigma} f(s) ds + \big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big)(-\alpha + \delta) = \\ &\int_{\boldsymbol{s}\in\{\boldsymbol{s}:s^{\Sigma}\geq\alpha\}\setminus A_{\text{GM}}} s^{\Sigma} f(\boldsymbol{s}) d\boldsymbol{s} - \int_{\boldsymbol{s}\in A_{\text{GM}}\cap\{\boldsymbol{s}:s^{\Sigma}<\alpha\}} s^{\Sigma} f(s) ds + \big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big)(-\alpha + \delta) \geq \\ &\alpha \int_{\boldsymbol{s}\in\{\boldsymbol{s}:s^{\Sigma}\geq\alpha\}\setminus A_{\text{GM}}} f(\boldsymbol{s}) d\boldsymbol{s} - \int_{\boldsymbol{s}\in A_{\text{GM}}\cap\{\boldsymbol{s}:s^{\Sigma}<\alpha\}} s^{\Sigma} f(\boldsymbol{s}) d\boldsymbol{s} + \big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big)(-\alpha + \delta) > \\ &\alpha \int_{\boldsymbol{s}\in\{\boldsymbol{s}:s^{\Sigma}\geq\alpha\}\setminus A_{\text{GM}}} f(\boldsymbol{s}) d\boldsymbol{s} - \alpha \int_{\boldsymbol{s}\in A_{\text{GM}}\cap\{\boldsymbol{s}:s^{\Sigma}<\alpha\}} f(\boldsymbol{s}) d\boldsymbol{s} + \big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big)(-\alpha + \delta) = \\ &\alpha\big(\Pr(\boldsymbol{s} \in A_{\text{full}} \setminus A_{\text{GM}}) - \Pr\left(\boldsymbol{s} \in A_{\text{GM}} \setminus A_{\text{full}}\right)\big) + \big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big)(-\alpha + \delta) = \\ &\delta\big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big), \end{aligned}$$

where the last equality follows from simple algebra (cf. Section 2.B for more details). Suppose now $\Pr(\boldsymbol{s} \in A_{\text{full}}) > \Pr(\boldsymbol{s} \in A_{\text{GM}})$; it follows that:

$$E[\tilde{U}_{\text{DM}}|\text{full}] - E[\tilde{U}_{\text{DM}}|\text{GM}] > \delta\big(\Pr(\boldsymbol{s} \in A_{\text{full}}) - \Pr(\boldsymbol{s} \in A_{\text{GM}})\big) > 0$$

and thus full comm. $\succ_{\text{DM}}$ GM, as was to be shown. □

While this section's implications are encouraging, the set of GMs is too general to permit precise statements. Importantly, the expected sum of all other groups' signals $\boldsymbol{s}^{\Sigma}_{-g_i}$ given a number of $V-1$ positive votes and a threshold t does not have a general and closed-form expression for most distributions. Unfortunately, this quantity is at the core of the game itself: consider, for instance, the expected value of all other groups' signals given group g_i is pivotal upon implementation of GM(m, V, t):

$$\mathbb{E}[\boldsymbol{s}^{\Sigma}_{-g_i}|\boldsymbol{v}^{\Sigma}_{-g_i} = V-1, \boldsymbol{y}^*] = (V-1)\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_j}|\boldsymbol{s}^{\Sigma}_{g_j} \geq t] + (m-V)\,\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_j}|\boldsymbol{s}^{\Sigma}_{g_j} < t]$$

It is therefore rather difficult to generally identify optimal values of m and t or to determine the conditions under which a GM is able to improve upon full communication. In light of the above, we decided to focus attention on two specific classes of GMs, the "expected value GMs" (μ-GMs) and the "$N/2$-GMs" ($N/2$-GMs):

Definition 5.
A μ***-GM*** *is a GM in which* $t = \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i}]$.
An $N/2$***-GM*** *is a GM with* $m = 2$ *and* $t = N/2$.

In a μ-GM, groups are asked to vote for the policy if and only if the sum of their signals is higher than expected. The number of groups is not fixed and will be the choice variable. This, along with the assumption of normally distributed signals, allows us to characterise optimal numbers of groups in the form of a simple maximisation problem (Section 2.3.3).[12] Note that, as will be shown in Section 2.3.4, restricting t to equal the expected sum of within-group signals does not come with much loss of generality, as results do not change qualitatively compared to those obtained for arbitrary threshold values.

In $N/2$-GMs, on the other hand, experts are divided into two groups and asked to vote for the policy if their sum of signals is equal to $N/2$, that is, the number of experts in each group. By fixing m and t, this refinement, along with the assumption of signals being Bernoulli trials, enables us to analyse comparative statics of the grouping mechanism. Furthermore, we illustrate why and how even a very simple

[12] That is, optimal μ-GMs, given a set of parameter values.

form of grouping can improve upon full and no communication without requiring complicated optimisation procedures (Section 2.3.5).

2.3.3 Expected value grouping mechanisms and optimality

As Section 2.3.2 suggests, finding the optimal GMs (i.e. combinations (m, V, t)) without any kind of restriction on the choice set is computationally intractable and results do not yield to interpretation. Hence, in order to (at least partially) address the question of optimality while still providing closed-form expressions, we need to restrict attention to a specific kind of grouping mechanism and make assumptions on the signal distribution: this section is dedicated to a closer discussion of μ-GMs given normally distributed signals. Below, we characterise optimal μ-GMs and discuss under which conditions a μ-GM is able to improve upon full communication (both given the below assumption).

Assumption normality. *Signals are normally distributed with mean μ and variance σ^2; that is, for all $i \in \{1, ..., N\}$, $s_i \sim \mathcal{N}(\mu, \sigma^2)$.*

The normal is one of the few distributions for which a closed-form of the truncated expected value exists. Accordingly, restricting attention to normally distributed signals allows us to derive closed-form expressions and, hence, characterise optimal values of m. Assuming normality and letting $t = n_m\mu$, the expected value of $\boldsymbol{s}^{\Sigma}_{g_i}$, given its being at least/less than the threshold t, can be expressed as follows:[13]

$$\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} \gtrless n_m\mu] = n_m\mu \pm \rho(m),$$

where $\rho(m) := \sigma\sqrt{\frac{2N}{m\pi}}$, and may be thought of as the expected "additional policy-favourable evidence" of a group submitting a vote indicative of a sum of signals greater than $n_m\mu$.

Now, as t is fixed and V is determined by Proposition 1, the number of groups m is the unique choice variable. Accordingly, to determine the set of optimal μ-GMs it suffices to find values of m that yield the highest expected utility. This is not only convenient from an analytical standpoint but also desirable as regards the context of the model, as the number of groups is, in fact, the key parameter that a DM

[13] cf. Greene (2003)

$$\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} < n_m\mu] = n_m\mu - \sigma\tfrac{\sqrt{N}\phi(0)}{\sqrt{m}\Phi(0)} = n_m\mu - \sigma\tfrac{\sqrt{N}2}{\sqrt{m}2\pi} = n_m\mu - \sigma\sqrt{\tfrac{2N}{m\pi}}$$
$$\mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_i} \mid \boldsymbol{s}^{\Sigma}_{g_i} \geq n_m\mu] = n_m\mu + \sigma\sqrt{\tfrac{2N}{m\pi}},$$

where $\phi(x)$ and $\Phi(x)$ are the pdf and cdf of $\mathcal{N}(0, 1)$.

may choose. To find such an optimal grouping, we need to consider the conditions implied by Proposition 1 that determine whether the DM is able to implement a μ-GM with m groups. To simplify their statement, define

$$\tilde{\alpha} := \frac{(\alpha - \delta - N\mu)\sqrt{\pi}}{\sigma\sqrt{2N}}.$$

In a similar vain to $\rho(m)$, $\tilde{\alpha}$ may be interpreted as follows: The first term in the numerator of $\tilde{\alpha}$ is the negative of DM's ex-ante expected utility from choosing the policy, which is then divided by the standard deviation of $\boldsymbol{s}^{\Sigma}$ and multiplied by a scaling factor (viz. $\sqrt{\pi/2}$). Consequently, $\tilde{\alpha}$ may be thought of as a measure of the DM's initial pessimism regarding the policy, adjusted for the degree of variability of said assessment.

$$V = V(m) := \lceil \tilde{V}(m) \rceil := \left\lceil \frac{(\alpha - \delta - N\mu)}{2\rho(m)} + \frac{m}{2} \right\rceil = \left\lceil \frac{\tilde{\alpha}\sqrt{m}}{2} + \frac{m}{2} \right\rceil \tag{2.6}$$

$$V \geq \tilde{V}(m) + \frac{1}{2} + \frac{\delta}{2\rho(m)} \Rightarrow \qquad \delta < \rho(m) \tag{2.7}$$

$$\sqrt{m} \geq \tilde{\alpha} = \frac{(\alpha - \delta - N\mu)\sqrt{\pi}}{\sigma\sqrt{2N}} \tag{2.8}$$

$$m \geq 2 \text{ and } N/m \in \mathbb{N} \tag{2.9}$$

Line 2.6 shows how the minimum required number of positive votes $V(m)$ varies in (i) the number of groups m and (ii) $\tilde{\alpha}$:
(i) a higher number of groups is associated with a decrease in the threshold value $t = (N\mu)/m$. Accordingly, the DM needs more evidence in the form of positive votes to be confident enough to choose the policy. Hence, $V(m)$ increases.
(ii) $\tilde{\alpha}$ is a measure of the DM's initial pessimism regarding the policy. As the policy becomes less profitable and $\tilde{\alpha}$ increases, more evidence is needed for adoption. Accordingly, $V(m)$ increases as well.

Recalling the aforementioned interpretation of $\rho(m)$, line 2.7 states that the additional evidence in favour of the policy must be greater than the conflict of interest (cf. Corollary 1). This, as most of our results, is closely related to the information a group may infer from being pivotal. If the added value of its positive vote is too small for the conflict of interest to be overcome, the μ-GM cannot be incentive compatible. For appropriately chosen m, this constraint can be less demanding than the condition that needs to be satisfied for a *single* expert j to vote $v_j = 1$ (cf. Section 2.3.2).

Line 2.8 ensures that $V(m) \leq m$: there are policies that do not allow for a μ-GM as the implied initial pessimism is simply too high; in such cases, $V(m) > m$ for all

admissible values of m, which never leads to adoption of the policy.

A basic understanding at hand, we are able to analyse the set of optimal values of m; that is, optimal μ-GMs:

Proposition 3. *Suppose normality and a μ-GM is used. The set of optimal m (i.e. optimal μ-GMs) are solutions to the following problem:*

$$\max_{\substack{2\leq m\leq N;\, N/m\in\mathbb{N};\\ \lceil \tilde{V}(m)\rceil \geq \tilde{V}(m)+1/2+\delta/(2\rho(m))}} \rho(m) \sum_{k=V(m)}^{m} \left(k-\tilde{V}(m)\right) 0.5^m \binom{m}{k}.$$

Let us first interpret the term to be maximised. Note that $0.5^m\binom{m}{k}$ measures the probability of receiving k positive votes as the probability of a group's sum of signals being higher than expected is exactly equal to 0.5. Accordingly, an optimal number of groups maximises the expected additional evidence in favour of the policy multiplied by the expected number of votes conditional on them being higher than the threshold $\tilde{V}(m)$. Hence, two counteracting forces need to be balanced:
(1) payoff effect: the expected additional evidence decreases in the number of groups as $\rho(m) = \sigma\sqrt{(2N)/(m\pi)}$.
(2) probability effect: the expected number of votes exceeding the threshold increases in m.
In other words, the DM has to trade off the increase in the probability of adoption with the fact that a higher number of groups may lead to relatively less profitable policies being chosen. Taking it to the extreme, full communication only allows for very profitable policies (additional evidence maximised), but the probability of adoption is comparatively small (adoption only if profitable for experts). Accordingly, small numbers of groups are similar to the "safe" option full communication with much evidence and smaller probability of adoption; higher values of m, on the other hand, trade certainty of a good policy (i.e. less evidence) for a higher probability of adoption—particularly of policies that are only profitable for the DM.

The constraint $\lceil \tilde{V}(m)\rceil \geq \tilde{V}(m) + 1/2 + \delta/(2\rho(m))$ ensures the μ-GM to be implementable (cf. Proposition 1); that is, when satisfied, experts are able to acquire enough information in favour of the policy at the pivotal information set and willing to vote for adoption upon receipt of $s_{g_i}^{\Sigma} \geq (N\mu)/m$.
To see why it is not possible to make general statements about the optimal number of groups beyond those made above, consider the below expression, which is

proportional to that stated in Proposition 3.

$$\max_{\substack{2\leq m\leq N;\, N/m\in\mathbb{N};\\ \lceil \tilde{V}(m)\rceil \geq \tilde{V}(m)+1/2+\delta/(2\rho(m))}} \sigma\sqrt{\frac{2N}{m\pi}}\sum_{k=V(m)}^{m}\left(2k-m-\tilde{\alpha}\sqrt{m}\right)0.5^m\binom{m}{k}$$

Were we to know the exact value of parameters that imply the values of $\tilde{\alpha}$ and $\sigma\sqrt{\frac{2N}{\pi}}$, the problem would be trivial. With $\tilde{\alpha}$ in its general form, it is impossible to compare expected payoffs across different values of m. In spite of this restriction, the above can help establish a (rough) intuition for the relationship between the initial pessimism and the optimal number of groups. Simply put, low values of $\tilde{\alpha}$, that is, policies that are expected to be relatively profitable, allow for higher numbers of groups: the probability effect dominates the payoff effect and the DM does not have to fear adoption of unprofitable policies. If $\tilde{\alpha}$ is high, the DM should choose smaller numbers of groups as the payoff effect dominates: the risk of choosing an undesirable policy would be too high to justify high values of m that are associated with higher probabilities of adoption.

As the above suggests, the maximisation problem is non-linear and there is no generally applicable answer to the question which number of groups maximises the DM's utility. To illustrate further, consider examples 2 and 3:

Example 2. *$N = 12$, $U_{DM}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - (83 + 2/3)$, $U_{ex}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 99$, $\mu = 4 + 2/3$ and $\sigma = 38.2$.*

Example 3. *$N = 24$, $U_{DM}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 73$, $U_{ex}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 94$, $\mu = 2.18$ and $\sigma = 37$.*

In Example 2, the set of m inducing an implementable μ-GM is equal to $\{2, 4, 6\}$. Three groups are not possible: for $m = 3$, the DM's and the experts' incentive compatibility constraints do not align (cf. line 2.7), as

$$\begin{aligned}\tilde{V}(3) &= \frac{\tilde{\alpha}\sqrt{3}}{2} + 1.5 \approx \frac{0.26\sqrt{3}}{2} + 1.5 \approx 1.73\\ 2 = V(3) &< \tilde{V}(m) + \frac{1}{2} + \frac{\delta}{2\rho(3)} \approx 2.23 + \frac{15 + 1/3}{2\rho(3)}.\end{aligned}$$

Hence, for three groups, the decision maker is not able to implement the μ-GM without commitment. The left column in Figure 2.1 depicts the values of $V(m)$, $\rho(m)$ and $\mathbb{E}[U_{\mathrm{DM}} \,|\, \mu\text{-GM}, m]$ for the parameters given in Example 2 and the implied implementable m:

Example 2 (left column) vs. Example 3 (right column)

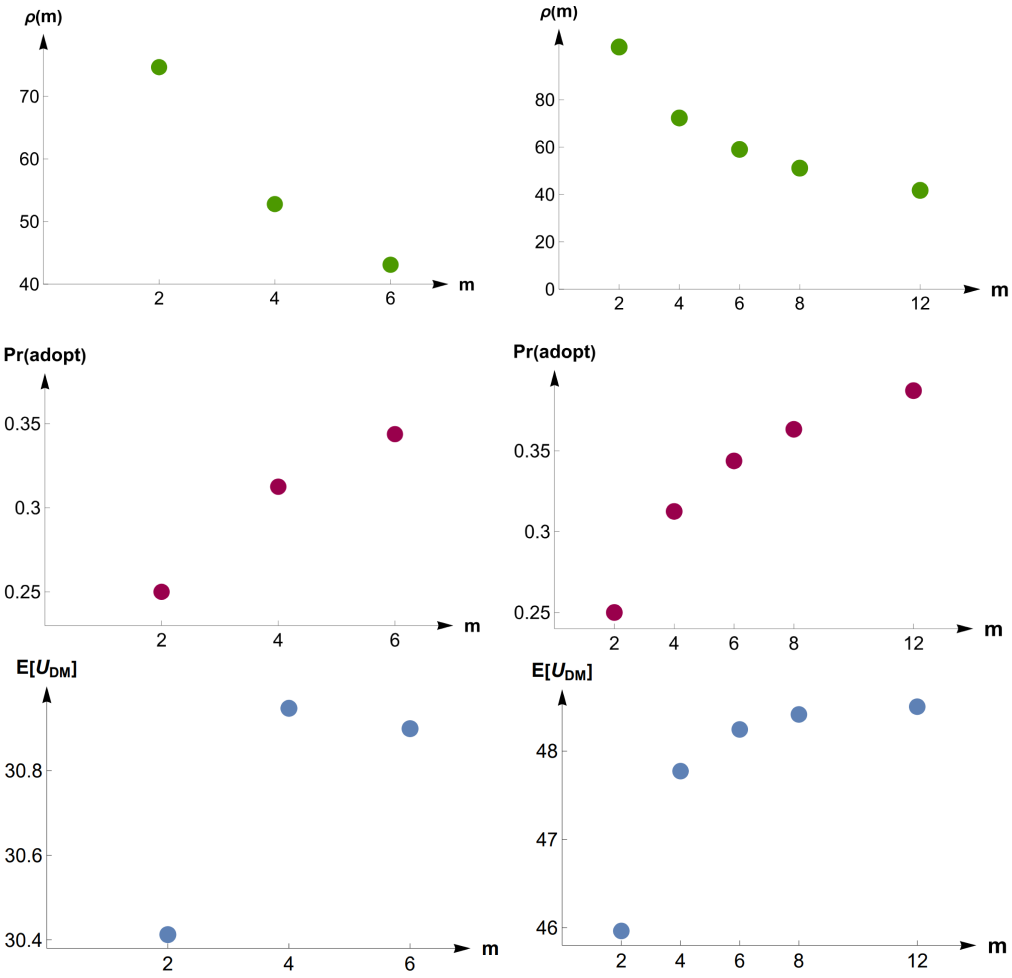

Figure 2.1: μ-GMs: comparison of group sizes

As the graphs indicate, neither the ranking of the additional evidences $\rho(m)$ nor that of the total probabilities of adoption are sufficient statistics for the optimal number of groups.[14] In this case, the best number of groups is equal to four, while $m = 4$ neither maximises $\rho(m)$ nor Pr(adopt).

Recall, the DM faces a tradeoff between probability and payoff effect: a higher probability of adoption is not necessarily superior in terms of utility, as Pr(adopt) does not contain any information about the states of the world in which the policy is chosen; the additional evidence, on the other hand, lacks consideration of the

[14]Naturally, the additional evidence ($\rho(m) = \sigma\sqrt{(2N)/(m\pi)}$) decreases in the number of groups. Furthermore, the share of positive group-votes needed for adoption ($V/m = \lceil \tilde{\alpha}\sqrt{m}/2 + m/2 \rceil / m$) weakly decreases in m, implying the positive relationship between Pr(adopt) and the number of groups.

frequency of adoption—were the DM to care about $\rho(m)$ exclusively, full communication would be the best alternative. In Example 2, the median number of groups $m = 4$ perfectly balances both effects and represents the best choice. However, note that there are examples in which the highest or lowest possible number of groups are the best choices.

Now, consider Example 3: the set of m inducing implementable μ-GMs is equal to $\{2, 4, 6, 8, 12\}$. As in Example 2, three groups are not possible due to the experts' incentive-compatibility constraints. The optimal number of groups, however, is different: twelve groups yield the highest ex-ante expected utility (cf. right column of Figure 2.1). The plots corresponding to Example 3 may suggest there being a direct relationship between the probability of adoption and the expected utility of the respective μ-GMs; as Example 2 indicates, this is not the case.

Comparing the two examples, we can establish that in the former, $\tilde{\alpha}$ is higher (≈ 0.26) than in Example 3 (≈ 0.14). This confirms the intuition of a smaller initial pessimism being associated with higher optimal numbers of groups. As the initial pessimism decreases, the DM is willing to take more risk in order to maximise the probability of adoption. Accordingly, she is satisfied with less additional evidence ($\rho(m)$) and the number of optimal groups increases. For higher values of $\tilde{\alpha}$ (i.e. more initial pessimism), the optimal number of groups decreases and the μ-GM "converges" to full communication.
Taken together, the above illustrate the non-linearity of the optimisation problem and the negative relationship between the initial pessimism and the optimal number of groups.

So far, we have considered grouping mechanisms in isolation; that is, we did not compare them to alternative mechanisms (without commitment or transfers). As high as the GM-payoff may be, it is of no use if full communication is more profitable for the DM. To address this concern, the following paragraphs are dedicated to the comparison of partial (GM) and full communication.

Let $f_{\text{BIN}}(k; m, p)$ denote the probability mass function of a binomial random variable with m trials and success probability p, evaluated at k; $F_{\text{BIN}}(k; m, p)$ denotes the respective cumulative distribution function, where $\hat{F}_{\text{BIN}}(k; m, p) = 1 - F_{\text{BIN}}(k-1; m, p)$.

Proposition 4. *Assuming normality, there exists a μ-GM such that μ-GM $\succ_{DM}$ full comm. if and only if there exists a value of m such that GM$(m, V(m), n_m\mu)$ is implementable and*

$$\frac{V(m)}{\sqrt{m}} f_{BIN}\big(V(m); m, 0.5\big) - \sqrt{\tfrac{\pi}{2}}\,\phi\big(\tfrac{\alpha - N\mu}{\sigma\sqrt{N}}\big) > \tilde{\alpha}\left(\hat{F}_{BIN}\big(V(m); m, 0.5\big) - \Big(1 - \Phi\big(\tfrac{\alpha - N\mu}{\sigma\sqrt{N}}\big)\Big)\right).$$

The expression stated in Proposition 4 can also be written as follows:

$$\underbrace{\frac{V(m)}{\sqrt{m}}}_{\approx 0.5(\tilde{\alpha}+\sqrt{m})} \Pr(\boldsymbol{\nu}^{\Sigma} = V(m)) - \sqrt{\tfrac{\pi}{2}}\Pr(\boldsymbol{s}^{\Sigma} = \alpha) > \tilde{\alpha}\,\big(\Pr(\text{adopt} \mid \text{GM}) - \Pr(\text{adopt} \mid \text{full comm.})\big).$$

To understand the above, we need to analyse the two counteracting forces that determine whether a GM is better than full communication: Under full communication, very profitable policies (even experts prefer them over the status quo) are adopted for sure. The likelihood of adoption of policies that are only profitable for the DM, however, is 0. Accordingly, the payoff terms in $\mathbb{E}[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{full comm.}]$ tend to be high, while the corresponding probability terms tend to be low.[15]

When a GM is used, even policies that are only profitable for the DM may be chosen and (if m is chosen appropriately; cf. Proposition 2) the total likelihood of adoption is higher. This comes at a cost: in some cases, policies that are not profitable for both experts and DM may be adopted.

The RHS of the condition represents the comparison of probabilities of adoption weighted by the DM's initial pessimism ($\tilde{\alpha}$). For comparatively unprofitable policies ($\tilde{\alpha}$ high), a "good" μ-GM should not lead to adoption with a much higher likelihood than full communication (term in brackets small). The LHS is related to the "worst-case adoption": the first term represents the probability of adoption at $V(m)$ positive votes and the second to adoption at the minimal sum of signals (i.e. "worst" state of the world) that allow for adoption under full communication—both weighted by a payoff term. If the weighted payoff of the μ-GM is relatively high even if only $V(m)$ positive votes were submitted, the μ-GM is likely to be superior to the full communication outcome.

[15] Roughly speaking, this is due to the fact that under full communication, the policy is adopted *for sure* if $\boldsymbol{s}^{\Sigma} \geq \alpha$ and the status quo is maintained *for sure* otherwise. In a GM, the probability of adoption is not concentrated on states of the world in which $\boldsymbol{s}^{\Sigma} \geq \alpha$, but adoption may, hence, not be guaranteed for all $\boldsymbol{s}^{\Sigma} \geq \alpha$.

In light of the expression in Proposition 4 being rather simple and easy to interpret, one may wonder whether it is possible to derive similar results for normally distributed signals without any restrictions on the value of t: in the following section, we provide an analogous result that does not impose constraints on the threshold values. Unfortunately, the terms contained therein are rather difficult to work with and require implicit definitions. Intuition and qualitative results, however, are analogous, suggesting the above restriction to serve a good and not overly limited simplification. Accordingly, the next section may be skipped by readers who are satisfied with the analysis provided in this section and less interested in the technical details of more general grouping mechanisms.

2.3.4 A short note on normally distributed signals and arbitrary values of t

To determine the conditions under which, given normally distributed signals and arbitrary threshold values, GMs are able to improve upon full communication, we use the so-called beta-normal distribution (BND). To explain said distribution and understand why it appears in the context of this model, it is instructive to first relate the binomial distribution, used thus far, to the ordinary beta distribution; to this end, consider the following expression relating their respective pmf and pdf:

$$f_{\text{BIN}}(k;\ m,p) = m^{-1} f_\beta(p; k+1, m-k+1)$$

Apart from superficial differences[16] such as the first factor on the RHS (m^{-1}), the most substantive difference between f_β and f_{BIN} is that in the former p is a variable, while in the latter it is a parameter; conversely, the latter's variable is k, while, in the former, $k+1$ is a parameter. Consequently, both functions can be thought of as alternative interpretations of the underlying data generating process, but may be exchanges for one another according to the previous expression.

To motivate going from the beta- to the beta-normal distribution, we first note that, in the present model, $p = \Pr(v_{g_i} = 1) = \Pr(\mathbf{s}^\Sigma_{g_i} \geq t)$. As, by assumption, $s_i \sim \mathcal{N}(\mu,\ \sigma^2)$, it follows that $\mathbf{s}^\Sigma_{g_i} \sim \mathcal{N}(n_m\mu,\ n_m\sigma^2)$ and thus:

$$p = \Pr(\mathbf{s}^\Sigma_{g_i} \geq t) = 1 - \Phi\left(\frac{t - n_m\mu}{\sigma\sqrt{n_m}}\right) = \hat{\Phi}\left(\frac{t - n_m\mu}{\sigma\sqrt{n_m}}\right) = \Phi\left(\frac{n_m\mu - t}{\sigma\sqrt{n_m}}\right) =: \Phi(-\bar{t}),$$

[16] The pdf of the beta distibution is usually defined as $f_\beta(p;a,b) := \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$, where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(x)$ denotes the gamma-function, which extends the factorial function, equalling $(x-1)!$ for any $x \in \mathbb{N}$ (cf.Zelen and Severo, 1972).

where Φ is the cdf of the standard normal distribution and $\hat{\Phi} := 1-\Phi$.

Accordingly, p is ultimately a function of t (and the distribution-parameters of $\mathbf{s}^{\Sigma}_{g_i}$). Consequently, it is desirable to alter f_β in such way as to account for this fact, which is precisely what is accomplished by the beta-normal distribution $\beta\mathcal{N}\left(x=-t;\ a=V+1, b=m-V+1, \eta=n_m\mu,\ \varsigma^2=n_m\sigma^2\right)$, with corresponding pdf and cdf (cf. Eugene, Lee and Famoye, 2002):[17]

$$f_{\beta\mathcal{N}}\left(-t;\ V+1, m-V+1, n_m\mu, n_m\sigma^2\right) := \frac{\Phi(-\bar{t})^V \hat{\Phi}(-\bar{t})^{m-V}\ \phi(-\bar{t})}{\sigma\sqrt{n_m}B(V+1, m-V)}$$

$$F_{\beta\mathcal{N}}\left(-t;\ V+1, m-V+1, n_m\mu, n_m\sigma^2\right) = \frac{\int_0^{\Phi(-\bar{t})} z^V(1-z)^{m-V}dz}{\sigma\sqrt{n_m}B(V+1, m-V)},$$

where $B(a,b)$ denotes the beta-function.[18] With this delineation in mind, we may state the following proposition:

Proposition 5. *Assuming normality, there exists a GM such that GM $\succ_{DM}$ full comm. if and only if there exists an implementable GM(m, V, t) such that*

$$\sqrt{\tfrac{\pi}{m2}} f_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - \sqrt{\tfrac{m\pi}{2}} f_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right) > \\ \tilde{\alpha}\left(F_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - \sqrt{m} F_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right)\right)$$

Note that in the above, the distribution functions $f_{\beta\mathcal{N}}\left(-\alpha; 1, 1, N\mu, N\sigma^2\right)$ and $F_{\beta\mathcal{N}}\left(-\alpha; 1, 1, N\mu, N\sigma^2\right)$ may be expressed in terms of the distribution function of $\mathbf{s}^{\Sigma}$, as follows:

$$f_{\beta\mathcal{N}}\left(-\alpha; 1, 1, N\mu, N\sigma^2\right) = \frac{\Phi\left(\frac{N\mu-\alpha}{\sigma\sqrt{N}}\right)^{1-1} \hat{\Phi}\left(\frac{N\mu-\alpha}{\sigma\sqrt{N}}\right)^{1-1}\ \phi\left(\frac{N\mu-\alpha}{\sigma\sqrt{N}}\right)}{\sigma\sqrt{N}\ B(1,1)} = \frac{\phi\left(\frac{N\mu-\alpha}{\sigma\sqrt{N}}\right)}{\sigma\sqrt{N}} \\ = f_{\mathcal{N}}\left(\alpha; N\mu, N\sigma^2\right)$$

$$F_{\beta\mathcal{N}}\left(-\alpha; 1, 1, N\mu, N\sigma^2\right) = \frac{\int_0^{\Phi\left(\frac{N\mu-\alpha}{\sigma\sqrt{N}}\right)} z^{1-1}(1-z)^{1-1}dz}{\sigma\sqrt{N}\ B(1,1)} = \frac{\Phi\left(\frac{N\mu-\alpha}{\sigma\sqrt{N}}\right)}{\sigma\sqrt{N}} \\ = \frac{1-F_{\mathcal{N}}\left(\alpha; N\mu, N\sigma^2\right)}{\sigma\sqrt{N}}$$

The technical details covered, we can interpret the condition, which, perhaps sur-

[17] For general $\beta\mathcal{N}\left(x;\ a, b, \eta,\ \varsigma^2\right)$ the pdf and cdf are:

$$f_{\beta\mathcal{N}}\left(x;\ a, b, \eta,\ \varsigma^2\right) := \frac{\Phi(x)^{a-1}\hat{\Phi}(x)^{b-1}\ \phi(x)}{\varsigma B(a,b)};\ F_{\beta\mathcal{N}}\left(x;\ a, b, \eta,\ \varsigma^2\right) = \frac{\int_0^{\Phi(z)} z^{a-1}(1-z)^{b-1}dz}{\varsigma B(a,b)}.$$

[18] For the definition of the beta-function, see footnote 16.

prisingly, is rather similar to its counterpart in Section 2.3.3 (wherein we require $t = \mu$): the terms are analogous to those in Proposition 4. The LHS measures the difference in the weighted worst-case adoption payoffs: under the GM, adoption for V positive votes; under full communication, adoption for $\boldsymbol{s}^{\Sigma} = \alpha$. The difference between these payoffs must be strictly greater than the net increase in the probability of adoption weighted by the initial pessimism (RHS). Note that the cdf corresponding to full communication is multiplied by $\sqrt{m}$ as the variance of $\boldsymbol{s}^{\Sigma}$ is equal to $\sigma\sqrt{N}$, while that of groups' signal-sums $\boldsymbol{s}^{\Sigma}_{g_i}$ amounts to $\sigma\sqrt{N/m}$. Accordingly, the GM can only improve upon full communication if it allows the DM to adopt more frequently (RHS) but does not do so for "too unprofitable" policies (LHS). Furthermore, the higher the initial pessimism, the smaller the probability of adoption of a "good" GM, as relatively unprofitable policies ($\tilde{\alpha}$ high) increase the attractiveness of the safe option full communication. Evidently, the interpretation is analogous to that in the previous sections' result Proposition 4; this suggests that restricting t to be equal to μ does not change results qualitatively and serves a good simplification.

As the multitude of definitions needed for its statement and the implicitly defined probability density functions suggest, Proposition 5 is rather impractical to work with; deriving results on the optimal values of t and V is, other than in Section 2.3.3, almost impossible. Accordingly, we decided to focus on the less involved but more restricted grouping mechanisms discussed in the previous and subsequent sections and leave the less constrained analysis for future research.

2.3.5 Comparative statics of two groups

In Section 2.3.3, we characterised the set of optimal μ-GMs given normally distributed signals and analysed the conditions under which they improve upon full communication. Considering the lengthy expressions involving implicitly defined variables, comparative statics are not very intuitive and rather involved. To nevertheless give a flavour of how the picture changes in the game's fundamentals, in this section, we consider a rather simple pair of GM and signal distribution: there are two groups, $t = N/2$ and experts receive binary signals (Bernoulli, cf. Wolinsky, 2002). Doing so, we are able to, at least partially, assess comparative statics and show that even a rather simple form of grouping can benefit the DM without requiring lengthy and complicated optimisation procedures.

Assumption Bernoulli. *Suppose signals are Bernoulli trials with success probability p. Furthermore, $\delta > 1$.*

The above corresponds to Wolinsky (2002)'s model, wherein $\delta > 1$ ensures the

conflict of interest to be "sufficiently severe" for there to be a problem worthy of being studied.

Naturally, not all parameter constellations allow for an $N/2$-GM to be implementable and potentially preferable for the DM: on the one hand, the conditions stated in Proposition 1 need to be satisfied (inter alia, due to incentive compatibility); on the other hand, such GM can never improve upon full communication if parameters are such that V must be equal to two for the $N/2$-GM to be implementable. In this case, the DM chooses the policy if and only if $\boldsymbol{s}^{\Sigma} = N$, which yields (weakly) lower utility than full communication (adoption of the policy for $\boldsymbol{s}^{\Sigma} \geq \lceil \alpha \rceil$). The below definition formalises these requirements:

Definition 6. *An $N/2$-GM is said to be a **potential improvement** if N is even,*

$$N/2 + \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < N/2] \geq \alpha \text{ and}$$
$$\alpha - \delta > 2\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < N/2].$$

Let $\Pr^{GM}$ be the probability that exactly one of the groups in the GM receives t positive signals and $\Pr_{\alpha}^{\text{full}}$ be that of sums of signals between $\lceil \alpha \rceil$ and $N-1$; that is,

$$\Pr^{GM} := 2(1-p^{N/2})p^{N/2} \text{ and } \Pr_{\alpha}^{\text{full}} := \Pr(N > \boldsymbol{s}^{\Sigma} \geq \lceil \alpha \rceil).$$

Consequently, $\Pr^{GM}$ and $\Pr_{\alpha}^{\text{full}}$ are the probabilities of adoption for the GM and full communication, respectively, less that of $\boldsymbol{s}^{\Sigma} = N$, where, trivially, the policy is adopted in both. Given these definitions and their interpretations, we can state the next result, which compares an $N/2$-GM to full communication.

Lemma 13. *Suppose Bernoulli. $N/2$-GM $\succ_{DM}$ full comm. if and only if $N/2$-GM is a potential improvement and*

$$Pr^{GM}\left(N/2 + \mathbb{E}\left[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < N/2\right]\right) - Pr_{\alpha}^{full}\,\mathbb{E}\left[\boldsymbol{s}^{\Sigma} \mid N > \boldsymbol{s}^{\Sigma} \geq \lceil \alpha \rceil\right] >$$
$$(\alpha - \delta)(Pr^{GM} - Pr_{\alpha}^{full}).$$

Lemma 13 characterises the requirements on parameter values that allow for the $N/2$-GM to be preferable for the DM: as one would expect, the $N/2$-GM needs to be a potential improvement; any $N/2$-GM that violates this condition is either not implementable or has the DM choose the policy if and only if all experts receive a positive signal—clearly not superior to full communication.

The inequality stated in the Lemma, on the other hand, can be derived from a comparison of the respective ex-ante expected utilities and can be interpreted as follows: The first term on the LHS is the expected value of signals given one positive

vote, multiplied by the GM's net probability of implementation. The latter factor (first parenthesis) is comprised of a part that is guaranteed (viz. $N/2$), in virtue of the positive vote by one group, and another that is the expected signal-sum of the group that voted against the policy. Similarly, the second term is the probability of a signal-sum between α and $N-1$, times its expected value in said case; naturally, it corresponds to full communication.

The first term on the RHS can be interpreted as the DM's "baseline" loss from the policy; that is, the loss in utility if all experts received a signal of zero. The second term measures the difference in the probability of adoption.

Given these individual interpretations, the inequality in Lemma 13 asserts that a potentially improving $N/2$-GM is preferable (for the DM) to full communication if and only if the gain (going from full communication to $N/2$-GM) in adoption-probability-weighted expected signal-sums exceeds the increase in magnitude of the baseline loss. For example, if the policy is likelier to be adopted under the $N/2$-GM (i.e. $\Pr^{GM} > \Pr^{\text{full}}_{\alpha}$), it can only be an improvement over full communication if the expected gain (LHS) from adoption is higher than the loss (RHS) compared to those given full communication. Put differently, a higher probability of adoption is not sufficient for the $N/2$-GM to be superior if it leads to frequent adoption of unprofitable policies. Hence, a "good" $N/2$-GM balances probability and payoff effects.

Given the above, we can establish Proposition 6, which provides insights into the comparative statics of an $N/2$-GM: let $N/2\text{-GM}(\alpha, \delta, p, N)$ be the $N/2$-GM given parameters (α, δ, p, N), then:

Proposition 6. *Suppose Bernoulli and $N/2$-GM$(\alpha, \delta, p, N) \succ_{DM}$ full comm.:*

(1) For any α' such that $\lceil \alpha' \rceil > \lceil \alpha \rceil$ and $N/2$-GM(α', δ, p, N) is a potential improvement,

$$N/2\text{-GM}(\alpha', \delta, p, N) \succ_{DM} \text{full comm.}.$$

(2) For any $\delta' > \delta$ such that $N/2$-GM(α, δ', p, N) is a potential improvement,

$$N/2\text{-GM}(\alpha, \delta', p, N) \succ_{DM} \text{full comm.}.$$

(1): An increase in α decreases the term $\Pr^{\text{full}}_{\alpha} \mathbb{E}\left[\boldsymbol{s}^{\Sigma} \mid N > \boldsymbol{s}^{\Sigma} \geq \lceil \alpha \rceil\right]$ on the LHS while keeping those corresponding to the GM constant. However, it also changes $-\Pr^{\text{full}}_{\alpha}(\alpha - \delta)$ on the RHS. In other words, it makes the policy less profitable (in all states of the world), while also decreasing the probability of adoption induced by full communication. Accordingly, it is not clear which one of these two effects is stronger. The statement shows: as long as the increase is small enough for the

$N/2$-GM to not require $V = 2$ to be implementable (i.e. remains to be a potential improvement), full communication is still inferior. The decrease in the probability of adoption under full communication is stronger than the effect on ex-post utilities and the DM profits from limiting the experts' power by partitioning them into groups.

(2): Changes in δ increase the gains from grouping: as the policy is more likely to be chosen under the $N/2$-GM than under full communication (cf. Proposition 2), increases in δ are associated with more group-favourable conditions. The now higher loss from leaving the decision to the experts under full communication increases the relative attractiveness of grouping. The higher the conflict of interest, the better the alternative solution in which experts are not able to fully determine the DM's decision.

In the Appendix (Example 5), we provide a detailed discussion of the benefits of an $N/2$-GM in a very simple framework. The example is supplementary; it may be skipped by readers who prefer not to delve into technical details and are satisfied by the explanations provided above.

To summarise, even very simple grouping mechanisms, such as the $N/2$-GM, may be able to improve upon full communication. Furthermore, the more evidence experts require to choose the policy, the higher the benefits of the grouping mechanism. Lastly, increases in the conflict of interest are associated with higher gains from grouping.

2.3.6 Heterogeneous group sizes

This section is dedicated to a "robustness check". In particular we discuss whether the beneficial effect of grouping relies on groups being equally sized.

Our findings are encouraging: Take, for instance, a set of four experts, assume GM$(2, 1, 2)$ is implementable and signals are Bernoulli trials (cf. Example 5); that is, the two groups vote for the policy if and only if they receive two positive signals and the DM adopts if and only if she receives at least one vote for the policy. Now suppose one expert is added to the set-up (now $N = 5$). Were we to only consider groupings covered by Definition 3, we would not be able to improve upon no communication, as groups must, by definition, be equally sized. Fortunately, we are able to apply the underlying concept even in cases in which N is a prime number. As the below result and example illustrate, the DM may still be able to make use of the threshold values used for $N = 4$ and improve upon both full and no communication. In other words, the concept of our mechanisms does not rely on

the assumption of groups being equally sized. To fix ideas, consider the following definition and subsequent result:

Definition 7. *An* ***implementable GM***$_{m+1}(m, V, t)$ *consists of*

1. *m groups of equal size N/m (typical group g_i) and one group of size 1 (labelled g_o),*
2. *a threshold value t such that $y^*_{g_i}(x^*, \boldsymbol{y}^*_{-g_i}, \boldsymbol{s}^{\Sigma}_{g_i}) = 1 \iff \boldsymbol{s}^{\Sigma}_{g_i} \geq t$, where $Pr(\boldsymbol{s}^{\Sigma}_{g_i} \geq t) > 0$ and*
3. *a threshold value $0 < V \leq m$ such that $x^*(\boldsymbol{v}, \boldsymbol{y}^*) = 1 \iff \boldsymbol{v}^{\Sigma} \geq V$.*

To simplify the statement of the below Proposition, we enrich the notation: Let GM$(m, V, t; N, \alpha, \delta, f_{S_i})$ be a GM(m, V, t) in a world in which there are N experts, α and δ are the utility-parameters and signals are distributed according to f_{S_i} (analogously for GM$_{m+1}(m, V, t; N+1, \alpha, \delta, f_{S_i})$), then:

Proposition 7. *Suppose $\bar{s} < t$. Let $\alpha' = \alpha - \mathbb{E}[s_i]$, then*

$$GM_{m+1}(m, V, t; N+1, \alpha, \delta, f_{S_i}) \text{ implementable}$$
$$\iff$$
$$GM(m, V, t; N, \alpha', \delta, f_{S_i}) \text{ implementable.}$$

Recall, in a GM$_{m+1}$, there are m groups of equal size and one group of size one—the "single expert". As $\bar{s} < t$, the single expert cannot vote for the policy (as is the case in Wolinsky, 2002); hence, the equally sized groups' conditions for incentive compatibility remain, essentially, unchanged compared to those given a GM$(m, V, t; N, \alpha, \delta, f_{S_i})$. The additional expert simply shifts their conditions by the expected value of his signal. This shifting is analogous to a decrease in α (if $\mathbb{E}[s_i] > 0$, increase otherwise). Hence, the GM$_{m+1}$ is implementable if and only if in a world in which there are N experts and α has been decreased (increased) by $\mathbb{E}[s_i]$, the standard GM is implementable.

Proposition 7 provides a simple method to deal with settings in which the number of experts does not allow for equally sized groups and extends the applicability of our results to a broader set of parametrisations. However, note that, given the above assumptions, the information contained in the single expert's signal is not used by the DM. Accordingly, GMs$_{m+1}$ are likely inferior to other forms of grouping: one could, for instance, simply add the $N+1^{\text{st}}$ expert to one of the m equally sized groups; below, we refer to this grouping as a GM$_{n_m+1}$. This type of GM is more complicated than the GMs$_{m+1}$ discussed above when it comes to incentive-compatibility

constraints: now, every group can be pivotal in many ways; accordingly, one has to consider various combinations of group-votes that lead to a tie to determine whether the proposed grouping is implementable. Hence, the GM_{m+1} can be understood as a computationally less involved alternative that may—in spite of the likely lower expected utility compared to a GM_{n_m+1}—still allow the DM to improve upon full and no communication.

To illustrate, consider the following example that compares the performances of a GM_{m+1}, a GM_{n_m+1} and full communication in a simple setting:

Example 4. *$N = 5$, $U_{DM}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 1.1$, $U_{ex}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 2.3$, $Pr(s_i = 1) = 0.19$ and $Pr(s_i = 0) = 0.81$.*

First option: a GM_{n_m+1}.
Suppose the DM partitions experts into two groups, of which one has three members. Let $t = 2$ and $\boldsymbol{\nu}^{\Sigma} = 1$ and refer to this type of grouping as a $\text{GM}_{n_m+1}(2,1,2)$. To see why the $\text{GM}_{n_m+1}(2,1,2)$ is incentive compatible, consider the following calculations: The small group, (w.l.o.g.) labelled 1, is willing to vote for the policy if $\boldsymbol{s}^{\Sigma}_{g_1} = 2$, as, assuming the DM and the other group play as required by the GM,

$$\mathbb{E}[\boldsymbol{s}^{\Sigma} | \boldsymbol{s}^{\Sigma}_{g_1} = 2,\ \nu_{g_2} = 0] = 2 + (3 \cdot 0.19 \cdot 0.81^2)/(0.81^3 + 3 \cdot 0.19 \cdot 0.81^2) \approx 2.41 > \alpha = 2.3.$$

The same holds for group 2:

$$\mathbb{E}[\boldsymbol{s}^{\Sigma} | \boldsymbol{s}^{\Sigma}_{g_2} = 2,\ \nu_{g_1} = 0] = 2 + (2 \cdot 0.19 \cdot 0.81)/(1 - 0.19^2) \approx 2.32 > 2.3.$$

Furthermore, given the groups play as required by the GM, the DM chooses to adopt the policy if and only if $\boldsymbol{\nu}^{\Sigma} \geq 1$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{s}^{\Sigma} | \nu_{g_1} = 0, \nu_{g_2} = 0] &= (2 \cdot 0.19 \cdot 0.81)/(1 - 0.19^2) \\
&\quad + (3 \cdot 0.19 \cdot 0.81^2)/(0.81^3 + 3 \cdot 0.19 \cdot 0.81^2) \\
&\approx 0.73 < \alpha - \delta = 1.1 \\
\mathbb{E}[\boldsymbol{s}^{\Sigma} | \nu_{g_1} = 1, \nu_{g_2} = 0] &= \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_2} | \boldsymbol{s}^{\Sigma}_{g_2} \geq 2] + \mathbb{E}[\boldsymbol{s}^{\Sigma}_{g_1} | \boldsymbol{s}^{\Sigma}_{g_1} < 2] \approx 2.39 > 1.1. \\
\mathbb{E}[\boldsymbol{s}^{\Sigma} | \nu_{g_1} = 0, \nu_{g_2} = 1] &= 2 + (3 \cdot 0.19 \cdot 0.81^2)/(0.81^3 + 3 \cdot 0.19 \cdot 0.81^2) \approx 2.41 > 1.1
\end{aligned}$$

By Observation 2 ($\delta > 1 = \bar{s}$), without communication, the policy is never adopted and the $\text{GM}_{n_m+1}(2,1,2)$ improves upon no communication. It is superior to full

communication as well, as

$$\begin{aligned}\mathbb{E}[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma})\,|\,\mathrm{GM}_{n_m+1}(2,1,2)] &= (5-1.1)\cdot 0.19^5 + (4-1.1)\cdot 0.19^4\cdot 0.81\cdot(2+3)\\ &\quad + (3-1.1)\cdot 0.19^3\cdot 0.81^2\cdot(1+6+3)\\ &\quad + (2-1.1)\cdot 0.19^2\cdot 0.81^3\cdot(3+1)\approx 0.1708,\\ \mathbb{E}[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma})\,|\,\text{full comm.}] &= (5-1.1)\cdot 0.19^5 + (4-1.1)\cdot 5\cdot 0.19^4\cdot 0.81\\ &\quad + (3-1.1)\cdot 10\cdot 0.19^3\cdot 0.81^2\approx 0.1018.\end{aligned}$$

Second option: a GM_{m+1}.
Alternatively, the DM could create an additional group (w.l.o.g. labeled 0) with only one member; that is, implement a $\mathrm{GM}_{m+1}(2,1,2)$. Recall, as the additional expert is never able to vote for the policy, the $\mathrm{GM}_{m+1}(2,1,2)$ can be understood as the DM's ignoring the fifth expert's information. Accordingly, his signal is estimated by $\mathbb{E}[s_5]=0.19$.
The $\mathrm{GM}_{m+1}(2,1,2)$ is incentive compatible: The groups of equal size (labeled i and j) are willing to vote for the policy upon receipt of two positive signals as, given the strategies implied by the GM,

$$\begin{aligned}\mathbb{E}[\boldsymbol{s}^{\Sigma}|\boldsymbol{v}^{\Sigma}_{g_j}=0,\boldsymbol{s}^{\Sigma}_{g_i}=2] &= (2\cdot 0.19\cdot 0.81)/(1-0.19^2)+2+\mathbb{E}[s_5]+2\\ &\approx 2.51>\alpha=2.3.\end{aligned}$$

Furthermore, assuming groups to act according to the strategies implied by the GM, the DM is not tempted to implement for less than one positive vote as

$$\mathbb{E}[\boldsymbol{s}^{\Sigma}|\boldsymbol{v}^{\Sigma}=0]=2\cdot(2\cdot 0.19\cdot 0.81)/(1-0.19^2)+\mathbb{E}[s_5]\approx 0.83<\alpha-\delta=1.1.$$

One positive vote, on the other hand, is sufficient evidence for the DM:

$$\mathbb{E}[\boldsymbol{s}^{\Sigma}|\boldsymbol{v}^{\Sigma}=1]=(2\cdot 0.19\cdot 0.81)/(1-0.19^2)+2+\mathbb{E}[s_5]\approx 2.51>1.1$$

As one would expect, the $\mathrm{GM}_{m+1}(2,1,2)$ is inferior to the $\mathrm{GM}_{n_m+1}(2,1,2)$ but superior to full communication, as

$$\begin{aligned}\mathbb{E}[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma})\,|\,\mathrm{GM}_{m+1}(2,1,2)] &= (5-1.1)\cdot 0.19^5 + (4-1.1)\cdot 0.19^4\cdot 0.81\cdot(1+4)\\ &\quad + (3-1.1)\cdot 0.19^3\cdot 0.81^2\cdot(4+2)\\ &\quad + (2-1.1)\cdot 0.19^2\cdot 0.81^3\cdot 2\approx 0.1021.\end{aligned}$$

This is due to the fact that, other than in the $\mathrm{GM}_{n_m+1}(2,1,2)$, the DM "ignores" the

additional expert's signal in the $GM_{m+1}(2,1,2)$. Accordingly, if possible (incentive compatible), it is likely preferable for the DM to add the expert to one of the existing groups and make use of the information contained in his signal. Put differently, a GM_{m+1} may represent a good alternative in cases in which GMs_{n_m+1} are either not incentive compatible or computationally too involved (multiple ways in which votes can be pivotal).

Furthermore and most importantly, the existence of implementable GMs_{m+1} and Proposition 7 imply: the assumption of equally sized groups is not the driver of our results.

2.4 Related literature

Clearly, our paper's closest relative is Wolinsky (2002), which, given its extensive discussion in previous sections, is not covered explicitly in this section.

Generally speaking, this paper relates to two strands of literature: (1) information transmission between a decision maker and multiple experts and (2) strategic voting.

The literature falling into the first category is vast and many solutions to the resolution of conflicts of interest have been proposed. Those studies differ in experts' preferences (homogeneous: e.g. Krishna and Morgan, 2001; heterogeneous: e.g. Gradwohl and Feddersen, 2018), the relationship between the state space and the signal (signals are suggestive of state: e.g. Gradwohl and Feddersen, 2018; signals determine the state: e.g. Quement, 2016) and the verifiability of signals (verifiable: e.g. Bhattacharya and Mukherjee, 2013; both verifiable and unverifiable: e.g. Bhattacharya and Mukherjee, 2013). As an example, consider Krishna and Morgan (2001) who analyse a game in which a decision maker consults two biased experts: As was the case in our model (Observation 2), consultation of experts in isolation is likely not optimal as they withhold substantial information. Comparing the single-expert case to sequential consultation of two experts, they find that full revelation can never occur if the experts have similar preferences. If, on the other hand, experts are biased in opposing directions, sequential consultation is always beneficial. Krishna and Morgan (2001)'s findings align with our results: common preferences among experts impede information revelation.[19]

Perhaps, the most important reference pertaining to work on strategic juries (the second relevant strand of literature) is Feddersen and Pesendorfer (1998). The pa-

[19] For more related analyses cf. e.g. Ekmekci and Lauermann (2022), Gilligan and Krehbiel (1989), Battaglini (2002) and Austen-Smith (1993).

per does an excellent job in illustrating the effect and drivers of strategic voting: a finite number of jurors receive a private and noisy signal about the state of the world whereafter they are asked to vote on two alternatives. By comparing different voting rules, the authors show that unanimity voting is likely to be inferior to majority voting under many circumstances. While the topic itself may not be closely related to our analysis, the core of Feddersen and Pesendorfer (1998)'s results is: if rational, voters condition their votes on the election being tight (they are "pivotal"). The two voting rules differ in the state of the world in which a vote may be pivotal and, hence, differ in their strategies. Under unanimity, a voter's decision matters only if all other voters have voted for the same alternative. Accordingly, there is much evidence for said alternative being the preferable one. Under majority, the picture is less clear and the voter may focus on his private information when deciding which option to choose. This difference in the "pivotal information set" is exactly what makes GMs profitable.
Gradwohl and Feddersen (2018) and Feddersen and Gradwohl (2020) provide a different interpretation of the effect of grouping on voters' strategies: analysing strategic voting in "small" committees, the authors show that transparency (i.e. making committee members' actions observable) harms information transmission in the presence of a conflict of interest and non-verifiable signals. Accordingly, oftentimes both DM and committee members prefer privacy or, as they call it, opacity. This finding is in line with our results on the beneficial effect of allowing for partial and unobserved communication between experts: the partitioning of experts into groups could be understood as a means to provide privacy. Instead of containing information about one signal in isolation and therefore being relatively precise (as is the case without communication), a group-vote maps multiple *vectors* of signal realisations to the same vote; that is, roughly speaking, it allows for more ambiguous messages which, in light of a conflict of interest, can enhance information transmission.

To conclude, we would like to mention Maug and Yilmaz (2002) who analyse a model in which a finite number of voters are asked to vote for one of two alternatives. Before doing so, they receive a signal suggestive of the state of the world. Other than in our model, voters differ in their preferences: there are two types of voters whose interests do not align. Thus, the conflict of interest does not arise between informed and uninformed parties but voters only. The authors show that partitioning them into two groups and requiring majority in each group improves upon the outcome of a standard "one-group" election if the conflict of interest is sufficiently severe. While their mechanism makes use of pivotal inference as well,

their set-up is rather distinct from ours and does not consider higher numbers of groups. Furthermore, most of their results rely on a "sufficiently large" number of voters, while we concentrate on results that hold for small N as well.

2.5 Conclusion

Building on work by Wolinsky (2002), we analyse a model in which a decision maker (DM) consults a finite number of experts to determine whether to adopt a policy. Each expert receives an iid signal; jointly, the experts' signals determine the state of the world and the desirability of the policy. There is a conflict of interest as the experts are less eager for the policy to be adopted than the DM.

For sufficiently severe conflicts of interest, the case in which experts are not allowed to communicate among one another (no communication) can never (or very rarely) lead to adoption of the policy—even for rather profitable policies. The conflict of interest is simply too profound. Allowing for (full) communication among all experts may be somewhat of a remedy: now, the policy is chosen if and only if it is profitable for the experts; the conflict of interest, however, persists. Analysing the intermediate case, partial communication, we elaborate on a simple mechanism proposed by Wolinsky (2002) that may allow the DM to elicit more information than in either of the extreme cases full and no communication. Partitioning the experts into smaller groups and allowing for intra- but not inter-group communication, the DM may be able to beneficially change the information experts can infer from being pivotal. The mechanism makes use of the fact that voters, if rational, condition their action on the event of being pivotal; that is, being able to change the DM's decision. An adequately chosen group size alters the events in which votes are pivotal and may, hence, be able to change experts' best responses to the benefit of the DM. Such grouping mechanism is particularly useful as it requires neither commitment nor transfers.

We characterise the set of optimal group sizes for a specific type of grouping mechanism and determine under which conditions such grouping is able to improve upon full communication. Thereafter, we discuss conditions under which even simple grouping mechanisms can improve upon full communication and analyse how changes in the conflict of interest or the gains from the policy influence the relative attractiveness of grouping. To conclude, we show that grouping can enhance information transmission even if, other than in pervious parts of the paper, groups are not equally sized.

Appendix 2.A

Detailed discussion of an example

To illustrate how GMs (i.e. an $N/2$-GM) may improve upon no and full communication, it is best to consider an example:[20]

Example 5. *$N = 4$, $U_{DM}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 1.1$, $U_{ex}(\boldsymbol{s}^{\Sigma}) = \boldsymbol{s}^{\Sigma} - 2.4$, $Pr(s_i = 1) = 0.25$ and $Pr(s_i = 0) = 0.75$.*

Clearly, as $s_i < \delta = 2.4 - 1.1$, votes cannot be positive without communication: the evidence a single expert may receive is never enough for him to vote for the policy (cf. Observation 2). Full communication, on the other hand, leads to adoption of the policy if and only if at least three experts have a positive signal.

A GM, however, may improve upon both outcomes: the experts could be divided into two groups ($m = 2$) and asked to vote $v_{g_i} = 1$ if and only if $\boldsymbol{s}_{g_i}^{\Sigma} = 2$; $x = 1$ if and only if $\boldsymbol{v}^{\Sigma} \geq 1$. Accordingly, in some states of the world, the policy is chosen even though it is only profitable for the DM. This improves upon both full and no communication. To see why the GM is implementable, consider the following arguments:
(1) Group 1 is willing to vote for the policy if $\boldsymbol{s}_{g_1}^{\Sigma} = 2$ as, assuming the DM and g_2 to play according to the strategies implied by the GM, such vote would be pivotal only if $v_{g_2} = 0$. In this case, $\boldsymbol{s}^{\Sigma} = 2$ or $\boldsymbol{s}^{\Sigma} = 3$. In expectation, this amounts to

$$\mathbb{E}[\boldsymbol{s}^{\Sigma} | \boldsymbol{s}_{g_1}^{\Sigma} = 2,\ \boldsymbol{s}_{g_2}^{\Sigma} < 2] = 2 + 1 \cdot (2 \cdot 0.25 \cdot 0.75)/(1 - 0.25^2) = 2.4 = \alpha.$$

By symmetry, the same holds for group 2.
(2) If the groups act as implied by the GM, the DM is willing to adopt the policy if and only if $\boldsymbol{v}^{\Sigma} \geq 1$ as

$$\mathbb{E}[\boldsymbol{s}^{\Sigma} | \boldsymbol{v}^{\Sigma} = 0] = 2 \cdot 1 \cdot (2 \cdot 0.25 \cdot 0.75)/(1 - 0.25^2) = 0.8 < 1.1 \text{ and}$$
$$\mathbb{E}[\boldsymbol{s}^{\Sigma} | \boldsymbol{v}^{\Sigma} = 1] = 2 + 1 \cdot (2 \cdot 0.25 \cdot 0.75)/(1 - 0.25^2) = 2.4 > 1.1.$$

To see why the GM improves upon full communication, consider the respective ex-

[20] We borrowed this example from Wolinsky (2002) and added some further explanation.

ante expected utilities:

$$\begin{aligned}\mathbb{E}[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma})\,|\,\text{GM}] &= (4-1.1)\cdot 0.25^4 + (3-1.1)\cdot 2(2\cdot 0.25\cdot 0.75\cdot 0.25^2)\\ &\quad + (2-1.1)\cdot 2\cdot 0.75^2\cdot 0.25^2 \approx 0.16\\ \mathbb{E}[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma})\,|\,\text{full comm.}] &= (4-1.1)\cdot 0.25^4 + (3-1.1)\cdot 4\cdot 0.25^3\cdot 0.75 \approx 0.10.\end{aligned}$$

Hence, as without communication, the policy is never adopted (cf. Observation 2), the GM improves upon both full and no communication.

Appendix 2.B

Proofs and omitted results

Proof of Observation 1. As the sum of all experts' signals fully determines whether the policy is profitable, experts encourage adoption if and only if $\boldsymbol{s}^{\Sigma} \geq \alpha$. As $\mathbb{E}[\boldsymbol{s}^{\Sigma} \mid \boldsymbol{s}^{\Sigma} < \alpha] \leq \mathbb{E}[\boldsymbol{s}^{\Sigma}] < \alpha - \delta$, the DM cannot do other than follow experts' advise. Accordingly, $t = \alpha$ for all distributions and $\boldsymbol{v}^{\Sigma} = 1$. Hence, the policy is chosen if and only if it is profitable for the experts. □

Proof of Observation 2. Group g_i with signal $\boldsymbol{s}_{g_i}^{\Sigma}$ is pivotal at a vote vector $\boldsymbol{v}_{-g_i}$ if the policy is chosen for vote $v_{g_i} = 1$ and not chosen for $v_{g_i} = 0$. Note that as we consider symmetric equilibria, $\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t, \boldsymbol{y}^*] = \mathbb{E}[\boldsymbol{s}_{g_j}^{\Sigma} \mid \boldsymbol{s}_{g_j}^{\Sigma} \geq t, \boldsymbol{y}^*]$ for all j, $i \in \{1, \ldots m\}$ (analogously for $\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*]$); pivotality implies

$$\begin{aligned}&x^*(\boldsymbol{v}_{-g_i}, v_{g_i} = 1, \boldsymbol{y}^*) = 1 > x^*(\boldsymbol{v}_{-g_i}, v_{g_i} = 0, \boldsymbol{y}^*) = 0 \iff\\ &V\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t, \boldsymbol{y}^*] + (m-V)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*] - \alpha + \delta \geq 0 >\\ &(V-1)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t, \boldsymbol{y}^*] + (m-V+1)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*] - \alpha + \delta \iff\\ &\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t, \boldsymbol{y}^*] + \delta - \boldsymbol{s}_{g_i}^{\Sigma} \geq\\ &-(V-1)\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t, \boldsymbol{y}^*] - (m-V)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*] + \alpha - \boldsymbol{s}_{g_i}^{\Sigma} >\\ &\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*] + \delta - \boldsymbol{s}_{g_i}^{\Sigma} \iff\\ &-\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \delta + \boldsymbol{s}_{g_i}^{\Sigma} \leq\\ &\mathbb{E}[U_{\text{ex}}(\boldsymbol{s}^{\Sigma}) \mid \boldsymbol{v}_{-g_i}^{\Sigma} = V-1, \boldsymbol{s}_{g_i}^{\Sigma}, \boldsymbol{y}^*] < -\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*] - \delta + \boldsymbol{s}_{g_i}^{\Sigma}.\end{aligned}$$

Accordingly, if $\boldsymbol{s}_{g_i}^{\Sigma} \leq \delta + \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t, \boldsymbol{y}^*]$ for all t and $\boldsymbol{s}_{g_i}^{\Sigma}$, groups never recommend the policy. This implies:

(1) Proposition 1 in Wolinsky (2002) (for the characterisation of utility functions we are using): he assumes $\delta > 1$ and $s \in \{0, 1\}$; hence, for a group size of 1 ("no

communication"), the policy is never chosen.
(2) If $S_i \in \mathbb{R}^+$, the policy is never chosen for all group sizes n such that $n \cdot \bar{s} < \delta$. □

Proof of Proposition 1. First, note that if GM(m, t, V) is implementable, groups vote $v_{g_i} = 1$ if and only if $\boldsymbol{s}_{g_i}^{\Sigma} \geq t$. Accordingly, $\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t, \boldsymbol{y}^*] = \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t]$. Hence, for the decision maker to choose the policy if and only if $\boldsymbol{v}^{\Sigma} \geq V$, the following inequalities need to be satisfied:

$$V\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] + (m-V)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t] \geq \alpha - \delta \tag{2.10}$$

$$(V-1)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] + (m-V+1)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t] < \alpha - \delta \tag{2.11}$$

Let $V = \lceil \tilde{V} \rceil$, the smallest integer greater than $\tilde{V}$—the number of votes for adoption that make the DM indifferent between the policy and the status quo; then:

$$\tilde{V} = \frac{\alpha - \delta - m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}. \tag{2.12}$$

Now, consider the experts' incentives: Note that a deviation in the form of a vote $v_{g_i} = 1$ if $\boldsymbol{s}_{g_i}^{\Sigma} < t$ is not possible due to the verifiability of "overreporting". Hence, we need to only ensure experts are willing to vote $v_{g_i} = 1$ if $\boldsymbol{s}_{g_i}^{\Sigma} \geq t$. Accordingly, we can establish that

$$\begin{aligned}
&\alpha \leq (V-1)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] + t + (m-V)\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t] \iff \\
&V \geq \frac{\alpha - m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t] + \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - t}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]} \\
&\quad = \tilde{V} + \frac{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - t + \delta}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}.
\end{aligned}$$

Accordingly, for the strategies to be consistent, V must satisfy the experts' constraint as well; this implies the above inequality, which is equal to that in the proposition.

Furthermore, V has to be weakly smaller than m. Hence, a necessary condition for m, V and t to induce an equilibrium is the following:

$$\begin{aligned}
&m \geq V \geq \tilde{V} \iff m - \frac{\alpha - \delta - m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]} \geq 0 \iff \\
&m\left(\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]\right) \geq \alpha - \delta - m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t] \iff \\
&m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] \geq \alpha - \delta
\end{aligned}$$

The remaining condition ensures group sizes to be equal and rules out full commu-

nication.

Lastly, consider the DM's beliefs, which are assumed to be formed by Bayes' rule. As long as t is a possible realisation of $\boldsymbol{s}_{g_i}^{\Sigma}$, $v_{g_i} = 1$ cannot be off path. Neither can $v_{g_i} = 0$, as this would imply t to be the smallest possible group signal realisation. This, on the other hand, would imply the policy to be chosen for any combination of signals which is neither optimal for the DM (second constraint violated) nor the experts (first constraint violated).

By the definition of the GM ($V \leq m$, $\Pr(\boldsymbol{s}_{g_i}^{\Sigma} \geq t) > 0$) and, as $V = 0$ does not constitute an equilibrium, the DM does not have off-path actions; hence, off path beliefs do not need to be defined.

Accordingly, we have shown that as long as all stated conditions are met, the grouping is possible, the DM's strategy is consistent with V and that of experts with t. Lastly, V is neither too high nor too low. Any GM(m, V, t) that does not satisfy one of these conditions is either not incentive compatible or m is not an admissible group number. Hence, the conditions are both necessary and sufficient. □

Proof of Corollary 1. The proof follows from the fact that a necessary condition for

$$V \geq \frac{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - t + \alpha - m\,\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]}$$

(cf. proof of Proposition 1) can be written as

$$\frac{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - t + \delta}{\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} \geq t] - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]} < 1 \iff \delta < t - \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \mid \boldsymbol{s}_{g_i}^{\Sigma} < t]. \quad \square$$

Proof of Proposition 2 by contraposition. Let A_{full} (resp. A_{GM}) be the subset of S^N, such that under full communication (resp. GM), the policy is adopted if and only if $\boldsymbol{s} \in A_{\text{full}}$ (resp. $\boldsymbol{s} \in A_{GM}$). The DM's expected utility is:

$$\begin{aligned} E[\tilde{U}_{\text{DM}}|\text{com}] &= E[\tilde{U}_{\text{DM}}|\boldsymbol{s} \in A_{\text{com}}]\Pr(\boldsymbol{s} \in A_{\text{com}}) \\ &= \frac{\int_{\boldsymbol{s} \in A_{\text{com}}} (s^{\Sigma} - \alpha + \delta) f(\boldsymbol{s}) d\boldsymbol{s}}{\Pr(\boldsymbol{s} \in A_{\text{com}})} \Pr(\boldsymbol{s} \in A_{\text{com}}) \\ &= \int_{\boldsymbol{s} \in A_{\text{com}}} s^{\Sigma} f(\boldsymbol{s}) d\boldsymbol{s} + \Pr(\boldsymbol{s} \in A_{\text{com}})(-\alpha + \delta) \end{aligned}$$

Therefore:

$$
\begin{aligned}
&E[\tilde{U}_{\mathrm{DM}}|\mathrm{full}] - E[\tilde{U}_{\mathrm{DM}}|\mathrm{GM}] = \\
&\int_{\boldsymbol{s}\in A_{\mathrm{full}}} s^{\Sigma} f(s)ds - \int_{\boldsymbol{s}\in A_{\mathrm{GM}}} s^{\Sigma} f(s)ds + \big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)(-\alpha+\delta) = \\
&\int_{\boldsymbol{s}\in A_{\mathrm{full}}\setminus A_{\mathrm{GM}}} s^{\Sigma} f(\boldsymbol{s})d\boldsymbol{s} - \int_{s\in A_{\mathrm{GM}}\setminus A_{\mathrm{full}}} s^{\Sigma} f(s)ds + \big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)(-\alpha+\delta) = \\
&\int_{\boldsymbol{s}\in\{\boldsymbol{s}:s^{\Sigma}\geq\alpha\}\setminus A_{\mathrm{GM}}} s^{\Sigma} f(\boldsymbol{s})d\boldsymbol{s} - \int_{\boldsymbol{s}\in A_{\mathrm{GM}}\cap\{\boldsymbol{s}:s^{\Sigma}<\alpha\}} s^{\Sigma} f(s)ds + \big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)(-\alpha+\delta) \geq \\
&\alpha \int_{\boldsymbol{s}\in\{\boldsymbol{s}:s^{\Sigma}\geq\alpha\}\setminus A_{\mathrm{GM}}} f(\boldsymbol{s})d\boldsymbol{s} - \int_{\boldsymbol{s}\in A_{\mathrm{GM}}\cap\{\boldsymbol{s}:s^{\Sigma}<\alpha\}} s^{\Sigma} f(\boldsymbol{s})d\boldsymbol{s} + \big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)(-\alpha+\delta) > \\
&\alpha \int_{\boldsymbol{s}\in\{\boldsymbol{s}:s^{\Sigma}\geq\alpha\}\setminus A_{\mathrm{GM}}} f(\boldsymbol{s})d\boldsymbol{s} - \alpha \int_{\boldsymbol{s}\in A_{\mathrm{GM}}\cap\{\boldsymbol{s}:s^{\Sigma}<\alpha\}} f(\boldsymbol{s})d\boldsymbol{s} + \big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)(-\alpha+\delta) = \\
&\alpha\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}\setminus A_{\mathrm{GM}}) - \Pr\left(\boldsymbol{s}\in A_{\mathrm{GM}}\setminus A_{\mathrm{full}}\right)\big) \\
&+ \big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)(-\alpha+\delta) = \\
&\alpha\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}\setminus A_{\mathrm{GM}}) + \Pr(\boldsymbol{s}\in A_{\mathrm{GM}}) - \Pr(\boldsymbol{s}\in (A_{\mathrm{GM}}\setminus A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{full}})\big) \\
&+ \delta\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big) = \\
&\alpha\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}\cup A_{\mathrm{GM}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{full}}\cup A_{\mathrm{GM}})\big) + \delta\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big) = \\
&\delta\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big)
\end{aligned}
$$

Suppose now that adoption is more probable under full communication than the GM; that is, $\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) > \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})$. It follows that:

$$E[\tilde{U}_{\mathrm{DM}}|\mathrm{full}] - E[\tilde{U}_{\mathrm{DM}}|\mathrm{GM}] > \delta\big(\Pr(\boldsymbol{s}\in A_{\mathrm{full}}) - \Pr(\boldsymbol{s}\in A_{\mathrm{GM}})\big) > 0$$

and thus full comm. $\succ_{\mathrm{DM}}$ GM, as was to be shown. □

Proof of Proposition 3. The DM's ex-ante expected utility takes the following form:

$$\mathbb{E}\Big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \mathrm{GM}(m, V(m), n_m\mu)\Big] =$$
$$\sum_{k=V(m)}^{m} \Bigg(k\, \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \,|\, \boldsymbol{s}_{g_i}^{\Sigma} \geq n_m\mu]$$
$$+ (m-k)\, \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \,|\, \boldsymbol{s}_{g_i}^{\Sigma} < n_m\mu] - \alpha + \delta \Bigg) f_{\mathrm{BIN}}\big(k; m, p = \mathbb{P}(\boldsymbol{s}_{g_i}^{\Sigma} \geq n_m\mu)\big) =$$
$$\sum_{k=V(m)}^{m} \left(k\left(n_m\mu + \sigma\sqrt{\frac{2N}{m\pi}} \right) + (m-k)\left(n_m\mu - \sigma\sqrt{\frac{2N}{m\pi}} \right) - \alpha + \delta \right) f_{\mathrm{BIN}}\big(k; m, 0.5\big) =$$
$$2\sigma\sqrt{\frac{2N}{m\pi}} \sum_{k=V(m)}^{m} k\, f_{\mathrm{BIN}}\big(k; m, 0.5\big)$$
$$+ \left(m\left(n_m\mu - \sigma\sqrt{\frac{2N}{m\pi}} \right) - \alpha + \delta \right) \hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big) =$$
$$2\sigma\sqrt{\frac{2N}{m\pi}} \sum_{k=V(m)}^{m} k\, f_{\mathrm{BIN}}\big(k; m, 0.5\big) + \left(N\mu - \sigma\sqrt{\frac{2mN}{\pi}} - \alpha + \delta \right) \hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big) =$$
$$2\rho(m) \sum_{k=V(m)}^{m} k\, f_{\mathrm{BIN}}\big(k; m, 0.5\big) + \Big(N\mu - \alpha + \delta - m\,\rho(m) \Big) \hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big)$$

Note that $N\mu - \alpha + \delta - m\,\rho(m) = -\tilde{V}(m)2\rho(m)$:

$$-\tilde{V}(m)2\rho(m) = -2\rho(m)\left(\frac{\alpha - \delta - N\mu}{2\rho(m)} + \frac{m}{2} \right) = N\mu - \alpha + \delta - m\rho(m)$$

Hence,

$$\mathbb{E}\Big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \mathrm{GM}(m, V(m), n_m\mu)\Big] = 2\rho(m) \sum_{k=V(m)}^{m} \big(k - \tilde{V}(m)\big)\, f_{\mathrm{BIN}}\big(k; m, 0.5\big) =$$
$$\rho(m) \sum_{k=V(m)}^{m} \big(2k - m - \tilde{\alpha}\sqrt{m}\big)\, 0.5^m \binom{m}{k},$$

which is proportional to the expression in the proposition. □

Lemma 14. *Suppose normality, then*

$$\mathbb{E}\Big[\tilde{U}_{DM}(\boldsymbol{s}^{\Sigma}) \,|\, GM(m, V(m), n_m\mu)\Big] =$$
$$0.5^m m!\left(\frac{\rho(m)}{(V(m)-1)!(m-V(m))!} - (\alpha - \delta - N\mu) \sum_{k=V(m)}^{m} \frac{1}{k!(m-k)!} \right).$$

Proof. By the proof of Proposition 3,

$$\mathbb{E}\big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \mathrm{GM}(m, V(m), n_m\mu)\big] =$$
$$0.5^m \sum_{k=V(m)}^{m} \big(k\rho(m) - (m-k)\rho(m) - (\alpha - \delta - N\mu)\big)\frac{m!}{k!(m-k)!} =$$
$$0.5^m\Bigg(\rho(m)m! \sum_{k=V(m)}^{m-1}\left(\frac{k}{k!(m-k)!} - \frac{(m-k)}{k!(m-k)!}\right) + \rho(m)m!\left(\frac{m}{m!} - \frac{0}{m!}\right) -$$
$$(\alpha - \delta - N\mu)\sum_{k=V(m)}^{m}\frac{m!}{k!(m-k)!}\Bigg).$$

Note that

$$\sum_{k=V(m)}^{m-1}\left(\frac{1}{(k-1)!(m-k)!} - \frac{1}{k!(m-k-1)!}\right) = \frac{1}{(V(m)-1)!(m-V(m))!} +$$
$$\sum_{k=V(m)+1}^{m-1}\left(-\frac{1}{(k-1)!(m-k)!} + \frac{1}{(k-1)!(m-k)!}\right) -$$
$$\frac{1}{(m-1)!0!} = \frac{1}{(V(m)-1)!(m-V(m))!} - \frac{1}{(m-1)!}.$$

Hence,

$$\mathbb{E}\big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \mathrm{GM}(m, V(m), n_m\mu)\big] =$$
$$0.5^m\Bigg(\rho(m)m!\left(\frac{1}{(V(m)-1)!(m-V(m))!} - \frac{1}{(m-1)!}\right) +$$
$$\rho(m)m!\frac{m}{m!} - (\alpha - \delta - N\mu)\sum_{k=V(m)}^{m}\frac{m!}{k!(m-k)!}\Bigg) =$$
$$0.5^m m!\left(\frac{\rho(m)}{(V(m)-1)!(m-V(m))!} - (\alpha - \delta - N\mu)\sum_{k=V(m)}^{m}\frac{1}{k!(m-k)!}\right). \quad \square$$

Proof of Proposition 4. First, note that by Lemma 14

$$\begin{aligned}
&\mathbb{E}\big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \mathrm{GM}(m, V(m), n_m\mu)\big] = \\
&0.5^m m!\left(\frac{\rho(m)}{(V(m)-1)!(m-V(m))!} - (\alpha-\delta-N\mu)\sum_{k=V(m)}^{m}\frac{1}{k!(m-k)!}\right) = \\
&0.5^m V(m)\rho(m)\binom{m}{V(m)} - (\alpha-\delta-N\mu)\sum_{k=V(m)}^{m} 0.5^m\binom{m}{k} = \\
&0.5^m \frac{V(m)}{\sqrt{m}}\frac{\sigma\sqrt{2N}}{\sqrt{\pi}}\binom{m}{V(m)} - (\alpha-\delta-N\mu)\sum_{k=V(m)}^{m} 0.5^m\binom{m}{k} = \\
&\frac{V(m)}{\sqrt{m}}\frac{\sigma\sqrt{2N}}{\sqrt{\pi}} f_{\mathrm{BIN}}\big(V(m); m, 0.5\big) - (\alpha-\delta-N\mu)\hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big).
\end{aligned}$$

Furthermore, by Greene (2003)

$$\begin{aligned}
&\mathbb{E}\big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{full comm.}\big] = \mathbb{E}\big[\boldsymbol{s}^{\Sigma} - \alpha + \delta \,|\, \boldsymbol{s}^{\Sigma} \geq \alpha\big]\Pr(\boldsymbol{s}^{\Sigma} \geq \alpha) = \\
&\Big(N\mu + \frac{\sigma\sqrt{N}\phi\big(\frac{\alpha-N\mu}{\sigma\sqrt{N}}\big)}{1-\Phi\big(\frac{\alpha-N\mu}{\sigma\sqrt{N}}\big)} - \alpha + \delta\Big)\Big(1-\Phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big)\Big) = \\
&(N\mu-\alpha+\delta)\Big(1-\Phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big)\Big) + \sigma\sqrt{N}\phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big).
\end{aligned}$$

Hence, the difference between the DM's expected utility under partial and full communication is positive if and only if

$$\begin{aligned}
&\mathbb{E}\big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \mathrm{GM}(m, V(m), n_m\mu)\big] - \mathbb{E}\big[\tilde{U}_{\mathrm{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{full comm.}\big] = \\
&\frac{V(m)}{\sqrt{m}}\frac{\sigma\sqrt{2N}}{\sqrt{\pi}} f_{\mathrm{BIN}}\big(V(m); m, 0.5\big) - (\alpha-\delta-N\mu)\hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big) + \\
&+ (\alpha-\delta-N\mu)\Big(1-\Phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big)\Big) - \sigma\sqrt{N}\phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big) = \\
&\frac{V(m)}{\sqrt{m}}\frac{\sigma\sqrt{2N}}{\sqrt{\pi}} f_{\mathrm{BIN}}\big(V(m); m, 0.5\big) - \sigma\sqrt{N}\phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big) + \\
&+ (\alpha-\delta-N\mu)\Big(1-\Phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big) - \hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big)\Big) \propto \\
&\frac{V(m)}{\sqrt{m}} f_{\mathrm{BIN}}\big(V(m); m, 0.5\big) - \phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big)\sqrt{\tfrac{\pi}{2}} + \\
&\tilde{\alpha}\left(1-\Phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big) - \hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big)\right) > 0 \iff \\
&\frac{V(m)}{\sqrt{m}} f_{\mathrm{BIN}}\big(V(m); m, 0.5\big) - \phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big)\sqrt{\tfrac{\pi}{2}} > \tilde{\alpha}\left(\Phi\big(\tfrac{\alpha-N\mu}{\sigma\sqrt{N}}\big) + \hat{F}_{\mathrm{BIN}}\big(V(m); m, 0.5\big) - 1\right),
\end{aligned}$$

as was to be shown. □

Proof of Proposition 5. For the normal distribution with zero-mean and unit-variance, let ϕ and Φ denote the pdf and cdf, respectively, as well as $\hat{\Phi} := 1-\Phi$.

Additionally, the following distributions will be used (cf. Zelen and Severo, 1972 and Eugene, Lee and Famoye, 2002):

- Binomial distribution: $B(m,p)$, with:

  pdf: $f_{\text{BIN}}(k;m,p) := \binom{m}{k} p^k (1-p)^{m-k}$.

  cdf: $F_{\text{BIN}}(V;m,p) := \sum_{k=0}^{V} \binom{m}{k} p^k (1-p)^{m-k} = I_{1-p}(m-V, V+1)$, where $I_{1-p}(m-V,V+1) := \frac{1}{B(m-V,V+1)} \int_0^{1-p} t^{m-V-1}(1-t)^V dt$ is the regularised incomplete beta function and $B(m-V,V+1) := \frac{\Gamma(m-V)\Gamma(V+1)}{\Gamma(m+1)}$ the beta function.

- Beta-normal distribution: $\beta\mathcal{N}(a,b,\mu,\sigma^2)$ with:

  pdf: $f_{\beta\mathcal{N}}(x;a,b,\mu,\sigma^2) := \frac{1}{\sigma B(a,b)} \Phi\left(\frac{x-\mu}{\sigma}\right)^{a-1} \hat{\Phi}\left(\frac{x-\mu}{\sigma}\right)^{b-1} \phi\left(\frac{x-\mu}{\sigma}\right)$

  cdf: $F_{\beta\mathcal{N}}(x;a,b,\mu,\sigma^2) := \frac{I_{\Phi\left(\frac{x-\mu}{\sigma}\right)}(a,b)}{\sigma}$

For later convenience, define the following normalised counterparts to α and t:

$$\overline{\alpha} := \frac{\alpha - N\mu}{\sigma\sqrt{N}} \qquad\qquad \overline{t} := \frac{t - n_m\mu}{\sigma\sqrt{n_m}}$$

Moreover, we state the following straightforward/well-known results without proof (cf. Zelen and Severo, 1972 and Greene, 2003):

$$\Pr(\#\{g : s_g^\Sigma \geq t\} = k) = f_{\text{BIN}}(k;m,\hat{\Phi}(\overline{t})) \tag{2.13}$$

$$\mathbb{E}\left[s_g^\Sigma | s_g^\Sigma \geq t\right] = n_m\mu + \sigma\sqrt{n_m}\frac{\phi(\overline{t})}{\hat{\Phi}(\overline{t})} \tag{2.14}$$

$$\mathbb{E}\left[s_g^\Sigma | s_g^\Sigma < t\right] = n_m\mu - \sigma\sqrt{n_m}\frac{\phi(\overline{t})}{\Phi(\overline{t})} \tag{2.15}$$

$$\mathbb{E}\left[s^\Sigma | s^\Sigma \geq \alpha\right] = N\mu + \sigma\sqrt{N}\frac{\phi(\overline{\alpha})}{\hat{\Phi}(\overline{\alpha})} \tag{2.16}$$

$$I_p(a,b) = 1 - I_{1-p}(b,a) \tag{2.17}$$

$$I_p(a+1,b) = I_p(a,b) - \frac{p^a(1-p)^b}{a\,B(a,b)} = I_p(a,b) - \frac{p^a(1-p)^b}{a}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{2.18}$$

We begin by deriving an expression for the DM's expected (ex-ante) utility from

GM(m, V, t), denoted $\mathbb{E}[\tilde{U}_{\text{DM}}|\text{GM}]$:

$$
\begin{aligned}
&\mathbb{E}[\tilde{U}_{\text{DM}}|\text{GM}] = \Pr(\text{adopt}|\text{GM})\mathbb{E}[\tilde{U}_{\text{DM}}|\text{GM}, \text{adopt}] \\
&= \Pr(\#\{g : s_g^\Sigma \geq t\} \geq V)\mathbb{E}[\tilde{U}_{\text{DM}}|\#\{g : s_g^\Sigma \geq t\} \geq V] \\
&= \sum_{k=V}^{m} \mathbb{E}[\tilde{U}_{\text{DM}}|\#\{g : s_g^\Sigma \geq t\} = k]\Pr(\#\{g : s_g^\Sigma \geq t\} = k) \qquad \text{by 2.13} \\
&= \sum_{k=V}^{m} \left(k\mathbb{E}\left[s_g^\Sigma | s_g^\Sigma \geq t \right] \right. \\
&\left. + (m-k)\mathbb{E}\left[s_g^\Sigma | s_g^\Sigma < t \right] - \alpha + \delta \right) \binom{m}{k} \hat{\Phi}\left(\bar{t}\right)^k \Phi\left(\bar{t}\right)^{m-k} \qquad \text{by 2.14 and 2.15} \\
&= \sum_{k=V}^{m} \left(k\left(n_m\mu + \sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\hat{\Phi}(\bar{t})} \right) \right. \\
&\left. + (m-k)\left(n_m\mu - \sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})} \right) - \alpha + \delta \right) \binom{m}{k} \hat{\Phi}(\bar{t})^k \Phi(\bar{t})^{m-k} \\
&= \sum_{k=V}^{m} \left(k\left(\sigma\sqrt{n_m}\left(\frac{\phi(\bar{t})}{\hat{\Phi}(\bar{t})} + \frac{\phi(\bar{t})}{\Phi(\bar{t})} \right) \right) \right. \\
&\left. + m\left(n_m\mu - \sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})} \right) - \alpha + \delta \right) \binom{m}{k} \hat{\Phi}(\bar{t})^k \Phi(\bar{t})^{m-k} \\
&= \sum_{k=V}^{m} (k - \hat{\Phi}(\bar{t})m)\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\hat{\Phi}(\bar{t})\Phi(\bar{t})} \binom{m}{k} \hat{\Phi}(\bar{t})^k \Phi(\bar{t})^{m-k} \\
&+ (N\mu - \alpha + \delta) \sum_{k=V}^{m} \binom{m}{k} \hat{\Phi}(\bar{t})^k \Phi(\bar{t})^{m-k} \\
&= \sum_{k=V}^{m} (k - \hat{\Phi}(\bar{t})m)\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\hat{\Phi}(\bar{t})\Phi(\bar{t})} \binom{m}{k} \hat{\Phi}(\bar{t})^k \Phi(\bar{t})^{m-k} + (N\mu - \alpha + \delta)\hat{F}_{\text{BIN}}(V; m, \hat{\Phi}(\bar{t})) \\
&= \sigma\sqrt{n_m}\phi(\bar{t}) \sum_{k=V}^{m} (k - \hat{\Phi}(\bar{t})m) \binom{m}{k} \hat{\Phi}(\bar{t})^{k-1} \Phi(\bar{t})^{m-k-1} + (N\mu - \alpha + \delta)\hat{F}_{\text{BIN}}(V; m, \hat{\Phi}(\bar{t}))
\end{aligned}
$$

The first summand in this last line can be rewritten as follows:

$$
\begin{aligned}
&\sqrt{n_m}\phi(\bar{t})\sum_{k=V}^{m}(k-\hat{\Phi}(\bar{t})m)\binom{m}{k}\hat{\Phi}(\bar{t})^{k-1}\Phi(\bar{t})^{m-k-1}\\
&=\sigma\sqrt{n_m}\phi(\bar{t})\left(\sum_{k=V}^{m}k\binom{m}{k}\hat{\Phi}(\bar{t})^{k-1}\Phi(\bar{t})^{m-k-1}-m\sum_{k=V}^{m}\binom{m}{k}\hat{\Phi}(\bar{t})^{k}\Phi(\bar{t})^{m-k-1}\right)\\
&=\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})}m\left(\sum_{k=V}^{m}\binom{m-1}{k-1}\hat{\Phi}(\bar{t})^{k-1}\Phi(\bar{t})^{m-k}-\sum_{k=V}^{m}\binom{m}{k}\hat{\Phi}(\bar{t})^{k}\Phi(\bar{t})^{m-k}\right)\\
&=\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})}m\left(\sum_{k=V-1}^{m-1}\binom{m-1}{k}\hat{\Phi}(\bar{t})^{k}\Phi(\bar{t})^{m-1-k}-\sum_{k=V}^{m}\binom{m}{k}\hat{\Phi}(\bar{t})^{k}\Phi(\bar{t})^{m-k}\right)\\
&=\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})}m\left(\hat{F}_{\mathrm{BIN}}(V-1;m-1,\hat{\Phi}(\bar{t}))-\hat{F}_{\mathrm{BIN}}(V;m,\hat{\Phi}(\bar{t}))\right)\text{ by def of }F_{\mathrm{BIN}}\text{ and 2.17}\\
&=\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})}m\left(I_{\hat{\Phi}(\bar{t})}(V,m-V)-I_{\hat{\Phi}(\bar{t})}(V+1,m-V)\right)\text{ by 2.18}\\
&=\sigma\sqrt{n_m}\frac{\phi(\bar{t})}{\Phi(\bar{t})}m\frac{\hat{\Phi}(\bar{t})^{V}\Phi(\bar{t})^{m-V}}{V\ B(V,m-V)}\\
&=\sigma\sqrt{n_m}\phi(\bar{t})\frac{\hat{\Phi}(\bar{t})^{V}\Phi(\bar{t})^{m-V-1}}{B(V+1,m-V)}\text{ by def of }f_{\beta\mathcal{N}}\\
&=\sigma\sqrt{n_m}f_{\beta\mathcal{N}}(-\bar{t};V+1,m-V,0,1)
\end{aligned}
$$

Therefore, $\mathbb{E}[\tilde{U}_{\mathrm{DM}}|\mathrm{GM}]$ can be written as:

$$
\begin{aligned}
&\mathbb{E}[\tilde{U}_{\mathrm{DM}}|\mathrm{GM}]=\sigma\sqrt{n_m}f_{\beta\mathcal{N}}(-\bar{t};V+1,m-V,0,1)+(N\mu-\alpha+\delta)F_{\beta\mathcal{N}}(-\bar{t};V+1,m-V,0,1)\\
&=\sigma^2 n_m f_{\beta\mathcal{N}}(-t;V+1,m-V,n_m\mu,n_m\sigma^2)\\
&+(N\mu-\alpha+\delta)\sigma\sqrt{n_m}F_{\beta\mathcal{N}}(-t;V+1,m-V,n_m\mu,n_m\sigma^2)
\end{aligned}
$$

In a similar manner, we can express the DM's expected utility from full communication, $\mathbb{E}[\tilde{U}_{\mathrm{DM}}|\mathrm{full}]$, as:

$$
\begin{aligned}
&\mathbb{E}[\tilde{U}_{\mathrm{DM}}|\mathrm{full}]=\Pr(\mathrm{adopt}|\mathrm{full})\mathbb{E}[\tilde{U}_{\mathrm{DM}}|\mathrm{adopt},\mathrm{full}]\\
&\text{by 2.16}\\
&=\Pr(s^{\Sigma}\geq\alpha)\mathbb{E}[\tilde{U}_{\mathrm{DM}}|s^{\Sigma}\geq\alpha]\\
&=\hat{\Phi}(\bar{\alpha})\left(N\mu+\frac{\sigma\sqrt{N}\phi(\bar{\alpha})}{\hat{\Phi}(\bar{\alpha})}-\alpha+\delta\right)\\
&=\sigma\sqrt{N}\phi(\bar{\alpha})+(N\mu-\alpha+\delta)\hat{\Phi}(\bar{\alpha})\\
&=\sigma\sqrt{N}f_{\beta\mathcal{N}}(-\bar{\alpha},1,1,0,1)+(N\mu-\alpha+\delta)F_{\beta\mathcal{N}}(-\bar{\alpha},1,1,0,1)\\
&=\sigma^2 N f_{\beta\mathcal{N}}(-\alpha,1,1,N\mu,N\sigma^2)+(N\mu-\alpha+\delta)\sigma\sqrt{N}F_{\beta\mathcal{N}}(-\alpha,1,1,N\mu,N\sigma^2)
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
&\text{GM} \succ_{\text{DM}} \text{full} \\
&\iff \mathbb{E}[\tilde{U}_{\text{DM}}|\text{GM}] > \mathbb{E}[\tilde{U}_{\text{DM}}|\text{full}] \\
&\iff \frac{\alpha-\delta-N\mu}{\sigma\sqrt{n_m}} = \\
&\tilde{\alpha}\sqrt{\frac{m2}{\pi}} < \frac{f_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - m f_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right)}{F_{\beta\mathcal{N}}(-t, V+1, m-V, n_m\mu, n_m\sigma^2) - \sqrt{m} F_{\beta\mathcal{N}}(-\alpha, 1, 1, N\mu, N\sigma^2)} \\
&\iff f_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - m f_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right) > \\
&\qquad \tilde{\alpha}\sqrt{\tfrac{m2}{\pi}}\left(F_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - \sqrt{m} F_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right)\right) \\
&\iff \sqrt{\tfrac{\pi}{m2}} f_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - \sqrt{\tfrac{m\pi}{2}} f_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right) > \\
&\qquad \tilde{\alpha}\left(F_{\beta\mathcal{N}}\left(-t, V+1, m-V, n_m\mu, n_m\sigma^2\right) - \sqrt{m} F_{\beta\mathcal{N}}\left(-\alpha, 1, 1, N\mu, N\sigma^2\right)\right),
\end{aligned}
$$

as was to be shown. □

Proof of Lemma 13. The expected utility of full communication $\mathbb{E}\left[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{full comm.}\right]$ and that of the GM $\mathbb{E}\left[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{GM}(2,1,N/2)\right]$ can be expressed as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{full comm.}\right] = \sum_{k=\lceil\alpha\rceil}^{N} (k-\alpha+\delta)\binom{N}{k} p^k (1-p)^{N-k} = \\
&(\delta-\alpha)\sum_{k=\lceil\alpha\rceil}^{N}\binom{N}{k} p^k (1-p)^{N-k} + \sum_{k=\lceil\alpha\rceil}^{N} k\binom{N}{k} p^k (1-p)^{N-k}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{U}_{\text{DM}}(\boldsymbol{s}^{\Sigma}) \,|\, \text{GM}(2,1,N/2)\right] = \\
&(\delta-\alpha)\Pr(\boldsymbol{v}^{\Sigma} \geq 1) + N\Pr(\boldsymbol{v}^{\Sigma} = 2) + \left(N/2 + \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \,|\, \boldsymbol{s}_{g_i}^{\Sigma} < N/2]\right)\Pr(\boldsymbol{v}^{\Sigma} = 1) = \\
&(\delta-\alpha)\left((p^{N/2})^2 + 2(1-p^{N/2})p^{N/2}\right) + \\
&N(p^{N/2})^2 + \left(N/2 + \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \,|\, \boldsymbol{s}_{g_i}^{\Sigma} < N/2]\right)2(1-p^{N/2})p^{N/2} = \\
&(\delta-\alpha)\left(p^N + 2(1-p^{N/2})p^{N/2}\right) + Np^N + \\
&\left(N/2 + \mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma} \,|\, \boldsymbol{s}_{g_i}^{\Sigma} < N/2]\right)2(1-p^{N/2})p^{N/2}
\end{aligned}
$$

Note that parameters that require $V = 2$ given $t = N/2$ and $m = 2$ can be disregarded as candidates for an improvement compared to full communication as those would imply adoption only for $\boldsymbol{s}^{\Sigma} = N$ (cf. Definition 13 and its discussion).

Accordingly, if the $N/2$-GM yields higher utility, the following needs to hold:

$$
\begin{aligned}
&(\delta-\alpha)\big(p^N+2(1-p^{N/2})p^{N/2}\big)+Np^N+\\
&\big(N/2+\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma}\,|\,\boldsymbol{s}_{g_i}^{\Sigma}<N/2]\big)2(1-p^{N/2})p^{N/2}>\\
&(\delta-\alpha)\sum_{k=\lceil\alpha\rceil}^{N}\binom{N}{k}p^k(1-p)^{N-k}+\sum_{k=\lceil\alpha\rceil}^{N}k\binom{N}{k}p^k(1-p)^{N-k}\iff\\
&2(1-p^{N/2})p^{N/2}\big(N/2+\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma}\,|\,\boldsymbol{s}_{g_i}^{\Sigma}<N/2]\big)>\\
&\sum_{k=\lceil\alpha\rceil}^{N-1}k\binom{N}{k}p^k(1-p)^{N-k}+(\alpha-\delta)\big(p^N+2(1-p^{N/2})p^{N/2}-\\
&\sum_{k=\lceil\alpha\rceil}^{N}\binom{N}{k}p^k(1-p)^{N-k}\big)=\\
&\mathbb{E}[\boldsymbol{s}^{\Sigma}\,|\,N>\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil]\Pr(N>\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil)+\\
&(\alpha-\delta)\big(2(1-p^{N/2})p^{N/2}-\Pr(N>\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil)\big)\iff\\
&\Pr^{GM}\big(N/2+\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma}\,|\,\boldsymbol{s}_{g_i}^{\Sigma}<N/2]\big)-\Pr_{\alpha}^{\text{full}}\mathbb{E}[\boldsymbol{s}^{\Sigma}\,|\,N-1\geq\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil]>\\
&(\alpha-\delta)(\Pr^{GM}-\Pr_{\alpha}^{\text{full}})
\end{aligned}
$$

Note that the $N/2$-GM needs to be a potential improvement as the conditions defining potential improvements are (1) necessary conditions for implementability following from Proposition 1 and (2) conditions ensuring that $V=1$: if $V=2$, the DM would choose the policy if and only if $\boldsymbol{s}^{\Sigma}=N$, which yields weakly lower utility than full communication (adoption for $\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil$).

Together, the above establish the Lemma. □

Proof of Proposition 6. Note that the below must hold if the $N/2$-$GM(\alpha,\delta,p,N)$ is superior to full communication:

$$
\begin{aligned}
&\Pr^{GM}\big(N/2+\mathbb{E}[\boldsymbol{s}_{g_i}^{\Sigma}\,|\,\boldsymbol{s}_{g_i}^{\Sigma}<N/2]\big)>\\
&\Pr_{\alpha}^{\text{full}}\mathbb{E}[\boldsymbol{s}^{\Sigma}\,|\,N>\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil]+(\alpha-\delta)(\Pr^{GM}-\Pr_{\alpha}^{\text{full}})
\end{aligned}
$$

As the LHS is constant in α, it suffices to elicit how the RHS changes if α increases. If α increases by Δ and $\lceil\alpha\rceil$ by one, the RHS decreases:

$$
\begin{aligned}
&\mathbb{E}[\boldsymbol{s}^{\Sigma}\,|\,N>\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil+1]\big(\Pr_{\alpha}^{\text{full}}-\Pr(\boldsymbol{s}^{\Sigma}=\lceil\alpha\rceil)\big)+\\
&(\alpha+\Delta-\delta)\big(\Pr^{GM}-\Pr_{\alpha}^{\text{full}}+\Pr(\boldsymbol{s}^{\Sigma}=\lceil\alpha\rceil)\big)-\\
&\mathbb{E}[\boldsymbol{s}^{\Sigma}\,|\,N>\boldsymbol{s}^{\Sigma}\geq\lceil\alpha\rceil]\Pr_{\alpha}^{\text{full}}-(\alpha-\delta)(\Pr^{GM}-\Pr_{\alpha}^{\text{full}})=\\
&-\lceil\alpha\rceil\Pr(\boldsymbol{s}^{\Sigma}=\lceil\alpha\rceil)+\Delta\Pr(\boldsymbol{s}^{\Sigma}=\lceil\alpha\rceil)<0,
\end{aligned}
$$

as $\alpha > N/2 > \Delta$ if the $N/2$-$GM(\alpha, \delta, p, N)$ is superior to full communication.

The comparative statics with respect to δ are trivial, as δ only influences the term $\alpha - \delta$. □

Proof of Proposition 7. Consider the $\text{GM}_{m+1}(m, V, t; N+1, \alpha, f_{S_i})$. As the single expert cannot vote for the policy ($\bar{s} < t$), the groups' conditions for incentive compatibility remain, essentially, unchanged. The additional expert simply shifts their conditions by the expected value of his signal. This could be understood as a decrease in α by $\mathbb{E}[s_i]$ if $\mathbb{E}[s_i] > 0$ (increase otherwise). The same holds for the DM. Accordingly, the $\text{GM}_{m+1}(m, V, t; N+1, \alpha, f_{S_i})$ needs to satisfy the incentive compatibility constraints of a $\text{GM}(m, V, t; N, \alpha', f_{S_i})$ in a world in which there are N experts and the experts utility upon implementation is equal to $\boldsymbol{s}^{\Sigma} - \alpha' = \boldsymbol{s}^{\Sigma} - (\alpha - \mathbb{E}[s_i])$. Hence, the $\text{GM}_{m+1}(m, V, t; N+1, \alpha, f_{S_i})$ is implementable if and only if the $\text{GM}(m, V, t; N, \alpha', f_{S_i})$ is. □

References

Austen-Smith, David. 1993. "Interested Experts and Policy Advice: Multiple Referrals under Open Rule." *Games and Economic Behavior*, 5(1): 3–43.

Battaglini, Marco. 2002. "Multiple Referrals and Multidimensional Cheap Talk." *Econometrica*, 70(4): 1379–1401.

Bhattacharya, Sourav, and Arijit Mukherjee. 2013. "Strategic information revelation when experts compete to influence." *RAND Journal of Economics*, 44(3): 522–544.

Ekmekci, Mehmet, and Stephan Lauermann. 2022. "Informal Elections with Dispersed Information: Protests, Petitions, and Nonbinding Voting." *Working Paper*.

Eugene, Nicholas, Carl Lee, and Felix Famoye. 2002. "Beta-normal distribution and its applications." *Communications in Statistics-Theory and Methods*, 31(4): 497–512.

Feddersen, Timothy, and Ronen Gradwohl. 2020. "Decentralized advice." *European Journal of Political Economy*, 63(November 2019): 101871.

Feddersen, Timothy, and Wolfgang Pesendorfer. 1998. "Convicting the Innocent: The Inferiority of Unanimous Jury Verdicts under Strategic Voting." *American Political Science Review*.

Gilligan, Thomas W., and Keith Krehbiel. 1989. "Asymmetric Information and Legislative Rules with a Heterogeneous Committee." *American Journal of Political Science*, 33(2): 459–490.

Gradwohl, Ronen, and Timothy Feddersen. 2018. "Persuasion and transparency." *Journal of Politics*, 80(3): 903–915.

Greene, William. 2003. "Econometric Analysis." *Prentice-Hall*, 5th edition: p. 759.

Krishna, Vijay, and John Morgan. 2001. "A Model of Expertise." *The Quarterly Journal of Economics*, 116(2): 747–775.

Maug, Ernst, and Bilge Yilmaz. 2002. "Two-Class Voting: A Mechanism for Conflict Resolution." *American Economic Review*, 92(5): 1448–1471.

Quement, By Mark Thordal-le. 2016. "The (Human) Sampler's Curses." *American Economic Journal: Microeconomics*, 8(4): 115–48.

Wolinsky, Asher. 2002. "Eliciting information from multiple experts." *Games and Economic Behavior*, 41(1): 141–160.

Zelen, Marvin, and Norman C Severo. 1972. "Probability Functions. Handbook of mathematical functions with formulas, graphs, and mathematical tables." *National Bureau of Standards*, 10th edition: pp. 944.

Chapter 3

Climate clubs: adverse effects and how to avoid them*

3.1 Introduction

The term "climate club" refers to a group of firms, individuals or even countries whose aim lies in the joint implementation of climate policies, such as emission-reducing regulations. By sharing the costs of a public good (e.g. research on environmentally friendly energy sources), clubs facilitate the fulfilment of their members' joint goal. The concept of climate clubs gained recognition as renowned economist William Nordhaus received the Nobel Prize for his research on the externalities of climate change. One of his most important contributions is Nordhaus (2018), an extensive analysis of climate clubs as a potential remedy.[1] Perhaps in response to the increasing popularity of the concept, in 2022, the G7 founded the "G7 Climate Club" whose main objective lies in promoting the implementation of the Paris Agreement. One of the proposed measures is a reduction in the production of emission-intensive goods, such as steel and cement.

While based on good intentions, such intervention does not come without risk: analysing a model in which a finite number of firms compete in a market for an emission-intensive good, this paper sheds light on potential adverse effects of said measure. To illustrate, suppose all G7 countries decide to heavily cut back on their

*This project was funded by the Bonn Graduate School of Economics and the German Research Foundation (DFG) through CRC TR 224 (project B03).
I am grateful for valuable comments by Daniel Krähmer, Philipp Hamelmann, Dezsö Szalay, Michael Krause, Lina Uhe, Silvio Sorbera, Carl-Christian Groh, Amelie Schiprowski and the participants of the YEP Workshop.

[1] While Nordhaus (2015) is not the first analysis on characteristics of club-like structures in the context of climate economics (cf. Barrett, 1994; Finus, Altamirano-Cabrera and Van Ierlan, 2005; Bosetti et al., 2013), William Nordhaus seems to have been the first to introduce the term "climate club". For more details on his contribution to the literature, see Section 3.4.

steel production. The worldwide demand for steel will, at least in the short run, not be changed by such decision and need to be served by non-members, such as China. I show that this effect may compensate for member countries' decreased emissions, leading to *higher* aggregate emissions (Proposition 8). Surprisingly, there are parametrisations for which both the emission-minimising and the optimal[2] club-production levels imply an *increase* in the club's emissions (Proposition 8): especially when non-members react heavily to changes in competitors' production levels, it can be optimal to increase the club's production and, hence, emissions. Due to the now lower price level, non-members' quantities decrease (heavily) and aggregate emissions do so as well. Unfortunately, however, it is neither possible to preclude the risk of adverse effects of a reduced club-production nor to determine optimal club-production levels without detailed knowledge of each non-member's production and emission function.

To address this shortcoming, I suggest two interventions (Proposition 9) that are free of any such risk and do not require club members to possess detailed knowledge of non-members' cost and emission functions. The first intervention efficiently reallocates production within the club without reducing the total quantity supplied by club members. Accordingly, the club can, for virtually all parametrisations, make sure to reduce within-club emissions. As the club does not reduce its market share, the market is not left to the goodwill of non-members; hence, non-members neither increase their production nor emissions. The second intervention makes use of the fact that the club is able to increase its supply without changing within-club emissions if it reallocates quantities among members emission-efficiently. This allows the club to reduce the price level, which, in turn, leads to lower supply and emissions by non-members. Taken together, aggregate emissions decrease. Both interventions are independent of non-members' characteristics and, hence, do not require any knowledge thereof. I characterise the two "incomplete-information robust" interventions for convex, concave and linear emission functions (Proposition 4). Thereafter, I discuss their implementation via tax schemes and the corresponding comparative statics (Proposition 10 and Theorems 5 and 6).

Taken together, this paper sheds light on risks of (uninformed) market interventions and proposes methods to avoid potential adverse effects that do not require in-depth knowledge of non-members' characteristics.

The remainder of the paper is structured as follows: in Section 3.2, I describe the model; Section 3.3 constitutes the main part of my analysis wherein I present and discuss results; Section 3.4 relates my work to the literature and Section 3.5

[2]That is, optimal in terms of the club's objective.

concludes.

3.2 Model

Demand function

Let $\mathcal{N} = \{1, 2, ..., n\}$ be the set of n firms, whereof a typical firm i produces a quantity q_i of a homogeneous and perfectly divisible good. The inverse demand function of said good is common knowledge and of the following form:

$$P\Big(\sum_i q_i\Big) = \alpha - \beta \sum_i q_i, \text{ where } \alpha, \beta > 0.$$

Firms

In the absence of regulation, firm i seeks to maximise its payoff π_i by producing quantity $q_i \geq 0$:

$$\max_{q_i \geq 0} \pi_i(q_1, ..., q_n) = q_i P\big(q_i + \textstyle\sum_{l \neq i} q_l\big) - c_i(q_i),$$

where the cost function $c_i(q_i) : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing (strictly for all $q_i > 0$) and strictly convex with $c_i(0) = 0$. The cost measured by c_i is not limited to "pure" costs of production but could also contain emission taxes or emission-related reputation concerns. Note that the above assumptions imply existence and uniqueness of the equilibrium before the founding of the club/in the absence of regulation (Szidarovszky and Yakowitz, 1982). Note that, while this paper analyses imperfect competition, the results presented in the following sections do not change qualitatively if, instead, firms are assumed to be price takers.

Firm i's emissions are measured by an increasing, differentiable and continuous emission function $e_i(q_i) : \mathbb{R}_+ \to \mathbb{R}_+$ with $e_i(0) = 0$ and $e_i'(q_i) > 0 \ \forall q_i > 0$. Emissions are not observable; they can, however, be inferred via knowledge of emission functions and quantities.

In later sections of the paper, attention will be restricted to convex emission functions which I consider to be fairly appropriate—especially given convex production costs: if, for instance, the efficiency of machines decreases in the quantity produced due to a built up of heat or longer per-unit production times, increasing average emission levels do not seem far-fetched. An overproportional increase in the reliance on emission-intensive energy sources or foreign resources (e.g. raw materials) would also likely be associated with a convex emission function. For a closer discussion of this assumption and related literature, consider Section 3.A in the appendix.

Climate club

The set of $0 < m < n$ firms located in countries that are members of the climate club is denoted by $\mathcal{M} \subset \mathcal{N}$ with typical element k. Without loss of generality, firms are labelled such that $\mathcal{M} = \{1, 2, ..., m\}$. Note that I do not impose assumptions on the number of firms in each country: as demand does not distinguish between firms' locations, only the set of firms located in member countries matters—not the set of countries themselves. Accordingly, I refer to firms as members (non-members) if they are located in member countries (non-member countries).

To simplify the analysis, I assume $e_k(q_k)$ to be strictly convex, strictly concave or linear for all $k \in \mathcal{M}$. Furthermore, if $e_k(q_k)$ is convex/concave/linear, then so is $e_l(q_l)$ for all $l, k \in \mathcal{M}$; that is, emission functions may differ, but their second derivatives are either all positive, all negative or all zero for all positive quantities.

Non-member-firms are not regulated; the set of non-members is denoted by $\mathcal{M}^C$ with typical element j.

Timing and equilibrium concept

Throughout, I compare equilibria of two games; namely, those *before* and *after* the club's founding, respectively. Before the founding, I consider the standard Cournot oligopoly in which firms move simultaneously and choose individually optimal quantities; that is, none of the firms are regulated. In the game after the founding, the directorate of the climate club D[3] decides on (i.e. regulates) the production levels of all member-firms and commits to its decision; the decision is observable. Thereafter, members produce as dictated by D and non-members respond optimally to the club's output. Accordingly, the Nash equilibrium *before* is compared to the equilibrium *after* the founding that results from non-members' subgame perfect response to the club's regulated quantities.

One may wonder why I chose to provide D with the power to decide on quantities before non-members choose theirs. In short, this shall reflect that D takes an informed decision and has access to more resources than a single Cournot-firm that chooses its quantity at the same time as his competitors. For a detailed discussion of my reasoning behind this choice, see page 116 (Section 3.3.1).

The club's objective is to reduce *aggregate emissions* taking into account changes in consumers' surplus $CS(p)$. D is assumed to know the shape of member-firms' cost and emission functions. Considering the timing described above, the directorate's

[3] D is not a set of firms but a player itself.

problem can be expressed as follows:

$$\min_{q_i \geq 0 \, \forall i \in \mathcal{N}} \quad (1-\omega)\Bigg(\sum_{k\in\mathcal{M}} e_k(q_k) + \sum_{j\in\mathcal{M}^C} e_j(q_j)\Bigg) - \omega\, CS\Big(P\big(\textstyle\sum_{i\in\mathcal{N}} q_i\big)\Big) \qquad (*)$$
$$\text{s.t. } P\big(\textstyle\sum_{i\in\mathcal{N}} q_i\big) = c_j'(q_j) + \beta q_j \;\; \forall j \in \mathcal{M}^C,$$

where the exogenous parameter $0 \leq \omega < 1$ measures the relative importance D assigns to consumers' welfare.
One may wonder why the objective neither considers the cost of implementation (e.g. via taxes) nor firms' profits. Lower post-club profits could incentivise firms to engage in research on emission-reducing technologies and internalise the social costs of pollution.
Naturally, any market intervention will unavoidably come at a cost for some party. Costs at the firm level are, compared to those incurred by consumers, simple to be compensated for by central institutions such as the club or governments: As will be shown in later sections, the disadvantages the club's regulation may have for member-firms[4] can be balanced by monetary transfers via taxes; for a broad set of parameter constellations, there exist tax functions that incentivise club members to endogenously choose to produce as wanted by D without reducing their profits.[5] Furthermore, as in many cases, consumers are voters and the decision to join the club is likely made by member countries' governments (not modeled here), D may have an incentive to prioritise consumers' welfare over firms' profits.

Further notation
A market is defined as a set $\{c_1, ..., c_n, e_1, ...e_n, \alpha, \beta, m\}$.

A capital letter with subscript is the sum over its corresponding lower case ones over the set specified by the subscript. For instance, the sum of entries in firms' pre-club equilibrium quantities $\left(q_i^0\right)_{i\in\mathcal{N}}$ is referred to as $Q^0_{\mathcal{N}}$, where the superscript is equivalent to that of the vectors' entries and the subscript indicates the set of firms whose quantities are contained in the vector. The same holds for emissions; for instance, $E^*_{\mathcal{M}^C}$ is the sum of non-members' emissions given quantities $\left(q_j^*\right)_{j\in\mathcal{M}^C}$.

The first derivative of some function $f(x)$, $x \in \mathbb{R}$ with respect to x is denoted by f', while f^{-1} represents its inverse.

[4]Naturally, non-members should not receive compensation for losses caused by the club's regulation: the losses could serve as a penalty for their not internalising the social cost of emission.
[5]That is, not reducing their profits compared to pre-founding equilibrium profits.

3.3 Main analysis and results

3.3.1 Adverse effects and the optimal intervention given complete information

At first sight, a reduction in club members' production—maybe even a market exit—may appear to be an effective and natural intervention leading to reduced emissions. As the title of this paper suggests, for some parametrisations, the exact opposite is the case. Importantly, there are markets in which any reduction in the club's production level leads to higher rather than lower aggregate emissions (than before the founding). Furthermore, it can be optimal for the club to *increase* its emissions in order to minimise aggregate emissions: emission-minimising member-production levels trade off reactions by non-members to changes in their competitors' supply against changes in the club's emission level. In some markets, it may therefore be preferable to *in*crease production such that the price level and, hence, non-members' emissions decrease. By optimally reallocating production within the club, the increased within-club emissions may be relatively harmless compared to non-members' reduced emissions:

Proposition 8. *There exist markets in which at least one of the below statements is true:*
(1) For a club-production level $Q^1_{\mathcal{M}}$ and the corresponding aggregate equilibrium emission level $E^1_{\mathcal{N}}$:

$$Q^1_{\mathcal{M}} < Q^0_{\mathcal{M}} \Rightarrow E^0_{\mathcal{N}} < E^1_{\mathcal{N}}.$$

(2) The emission-minimising member-quantities q_k^{min} are such that

$$E^0_{\mathcal{M}} < \sum_{k \in \mathcal{M}} e_k(q_k^{min}).$$

To better understand the above, consider Observation 3, which—given some simplifying assumptions—characterises the set of parameter values for which both statements apply.

Observation 3. *Suppose*
$c_j(q_j) = 0.5c_{\mathcal{M}^C}\, q_j^2$, $e_j(q_j) = \eta_{\mathcal{M}^C}\, q_j^2$ *for all* $j \in \mathcal{M}^C$ *and*
$c_k(q_k) = 0.5c_{\mathcal{M}}\, q_k^2$, $e_k(q_k) = \eta_{\mathcal{M}}\, q_k^2$ *for all* $k \in \mathcal{M}$.
Both statements made in Proposition 8 hold true if and only if

$$\underbrace{\frac{\eta_{\mathcal{M}^C}}{\eta_{\mathcal{M}}}}_{(i)} \quad \underbrace{\frac{q_j^0}{q_k^0}}_{(ii)} \quad \underbrace{\frac{\beta(n-m)}{\beta(n-m+1)+c_{\mathcal{M}^C}}}_{(iii)} > 1,$$

where q_k^0 *and* q_j^0 *denote the quantities produced by members and non-members in the pre-club equilibrium, respectively.*

In the market described in Observation 3, all firms have both quadratic cost and emission[6] functions. Furthermore, all members' cost and emission functions are equal; the same holds for non-members.

Consider now the condition stated in the observation. Recall, if the condition holds true, (1) any reduction in the clubs' production increases the total emission level and (2) it is necessary to increase within-club emissions to minimise aggregate emissions. Roughly speaking, the condition compares members' and non-members' emission parameters (i) and their supply in the absence of regulation (ii) weighted by an expression measuring the change in the sum of non-members' quantities as a reaction to a unit change in the club's production (iii).[7] Accordingly: if non-members are (i) comparatively "dirty", (ii) used to supply comparatively high quantities before the founding or (iii) change their quantities relatively strongly in response to changes in the club's production, then it is likely better for the club to *in*crease rather than decrease its production in order to lower emissions.
Note that (ii) also makes a statement about non-members' reaction/cost functions: High non-member-quantities pre club imply comparatively flat non-member-cost functions; that is, non-members react relatively strongly to changes in parameters or their competitors' supply.

Accordingly, if the condition holds, reductions in the club's production lead to relatively strong increases in non-members' quantities (ii and iii). As non-members are comparatively emission-intensive (i), this leads to an overall increase in emissions. Hence, D is better off *increasing* members' quantities: non-members strongly reduce their supply (ii and iii) and, as non-members are comparatively "dirty" (i), the increase in members' emissions does not compensate for the lower non-member-

[6] For a closer discussion of arguments that suggest convex emissions to be an appropriate assumption, see Section 3.A in the appendix.

[7] Non-member j's best response to members' quantity q_k takes the following form: $q_j^{\mathrm{BR}}(q_k) = (\alpha - \beta m q_k)/(\beta(n-m+1)+c_{\mathcal{M}^C})$ (cf. proof in Appendix 3.B for more details).

emissions. Hence, aggregate emissions decrease.

Naturally, the reverse holds if the condition is violated: in this case, emission-minimising club-quantities are smaller than before the founding, as an increase in non-members' supply and, hence, emissions is relatively harmless compared to the decrease in members' emissions.

As Lemma 15 shows, information about every *individual* non-member's cost and emission function is necessary to determine optimal actions and elicit whether a reduced production may have adverse effects—a, potentially, rather strong requirement. Before going into detail about alternative solutions that do not require such knowledge, I discuss what would be the *optimal* action to be taken by the club given full information; that is, the solution to its objective $(*)$.

To do so, I shortly elaborate on why I chose to give the club's directorate the power to decide on members' quantities before any production occurs: The central difference between the game before and after the founding is the fact that D internalises non-members' reactions, while given Cournot competition (before the founding), none of the players do so. Accordingly, the club's directory is "strategically" more sophisticated in its decision process. This could be due to access to better resources, such as teams of experts, market analyses and funds. Furthermore, the club's decision is likely observable as it must be communicated to all member countries/firms and may be subject to public interest. Hence, the assumption of D being able to take an informed and public decision before production occurs does not seem far-fetched.

Pivoting back to the production levels that best achieve the club's intentions, recall: D's objective is the reduction of aggregate emissions taking into account changes in consumers' surplus $CS(p)$. Parameter $0 \leq \omega < 1$ measures the relative importance D assigns to consumers' welfare:

$$\min_{q_i \geq 0 \ \forall i \in \mathcal{N}} \quad (1-\omega)\left(\sum_{k\in\mathcal{M}} e_k(q_k) + \sum_{j\in\mathcal{M}^C} e_j(q_j)\right) - \omega\, CS\Big(P\big(\textstyle\sum_{i\in\mathcal{N}} q_i\big)\Big) \qquad (*)$$
$$\text{s.t. } P\big(\textstyle\sum_{i\in\mathcal{N}} q_i\big) = c_j'(q_j) + \beta q_j \ \ \forall j \in \mathcal{M}^C$$

Let $(q_1^*, ..., q_N^*)^T$ be the solutions to $(*)$, then:

Lemma 15. *For all $k \in \mathcal{M}$, $j \in \mathcal{M}^C$ such that $q_k^*, q_j^* > 0$:*

$$e_k'(q_k^*) = e_j'(q_j^*) - \frac{c_j''(q_j^*) + \beta}{1-\omega}$$

As is readily apparent, the above does not fully characterise the optimal quantities. The problem's solutions are highly sensitive to exact parametrisations; the generality of the set-up does not allow for closed-form expressions. However, as the purpose of this part of the analysis is to establish a general understanding of the comparative statics of $(q_1^*, ..., q_N^*)^T$ and the information required for its implementation, a full characterisation is not necessary at this point. In other words, Lemma 15 simply serves as a motivation for "incomplete-information robust" interventions discussed in the following sections.

For illustration of the above result, assume e_i to be strictly convex and $e_i'(0) = 0$ for all i; this implies all members' and non-members' optimal quantities to be positive-valued.[8]

Trivially, Lemma 15 implies all members' marginal emissions to be equal. This ensures an emission-efficient allocation of quantities within the club; roughly speaking, "clean" members produce more than "dirty" ones—up until the point whereat their relative cleanliness is compensated for by the inefficiencies induced by the greater production level. As D is able to change non-members' production only *indirectly* (i.e. non-members' first order conditions need to be satisfied), this statement does not hold for non-members, whose marginal emissions may differ. This reflects the limited power D has over the market.

Furthermore, members' quantities are increasing in non-members' marginal emissions: roughly speaking, if non-members produce at relatively inefficient levels (e_j' high), members need to increase their production to crowd out that of non-members (i.e. decrease their emissions).

Now, consider the term $c_j''(q_j^*) + \beta$. Note that firm j's first order condition takes the following form:

$$P(Q_{-j}^* + q_j^*) = \alpha - \beta Q_{-j}^* - \beta q_j^* = c_j'(q_j^*) + \beta q_j^*$$

Accordingly, $c_j''(q_j^*) + \beta$ can be understood as an approximate measure for the sensitivity of firm j's reaction to changes in the price level. Firms with comparatively high values of $c_j''(q_j^*)$ react strongly to changes in competitors' supply; firms with "flatter" first order conditions (small $c_j''(q_j^*)$) make more moderate adjustments to their supply as a response to such changes. Referring back to Lemma 15: if non-members react strongly to changes in the club's production (high $c_j''(q_j^*) + \beta$), members do not need to take extreme actions to indirectly change their competitors' production

[8] For a closer discussion of arguments that suggest convexity to be an appropriate assumption, consider Section 3.A (appendix).

levels. If, on the other hand, non-members are insensitive to members' production, the club needs to take stronger measures. The weight $1/(1-\omega)$ addresses the club's concerns with respect to consumers' surplus: if more weight is placed on CS, then $1/(1-\omega)$ is large in magnitude as is the second summand on the RHS of the expression in Lemma 15. Consequently, the difference in marginal emissions between club- and non-club-members is greater than would be the case for small values of ω. A significant reduction in the aggregate output may be beneficial when it comes to emissions; it is, however, associated with a reduction in consumers' welfare. Accordingly, for a higher weight on consumers' welfare, a more moderate reduction in output/emissions is optimal than for smaller values of ω.

The above illustrates: the optimal quantities *perfectly* balance both non-members' reactions and changes in consumers' welfare in response to variations in members' production levels. To determine them, the club needs to perfectly predict changes in non-members' emission levels, necessitating detailed knowledge of every non-member's respective cost and emission function. To my mind, it would be bold to assume the club to have such profound information—especially in global markets with a high number of participants. In light of this shortcoming, the next section is dedicated to the analysis of interventions that do not require such knowledge and are, for virtually all parametrisations, guaranteed to reduce emissions while never harming consumers.

3.3.2 Emission reduction under incomplete information

Clearly, to best achieve the club's objective, members should produce according to $(q_1^*, ..., q_m^*)^T$. However, as alluded to previously, implementation thereof requires detailed information about non-members' maximisation problems and emission functions. To address this shortcoming, I propose two interventions that (1) do away with said requirement and (2) are, for virtually all parametrisations, guaranteed to reduce emissions while not decreasing consumers' surplus no matter the exact properties of non-members' production technologies.

Before going into detail about the alternative interventions, I need to define what is meant by "incomplete" information:[9] to determine the subsequent solutions, D needs to only possess knowledge about *members'* emission functions; that is, it can be ignorant of non-members' emission and cost functions.

[9]Note that in the game after the founding, non-members are able to observe the club's decision; this assumption is not altered in the current section. In a classic Cournot oligopoly (pre-club equilibrium), on the other hand, information is always imperfect, as firms act simultaneously; that is, they cannot observe other players' actions.

The following table compares the assumptions made on D's information in this section to those in Section 3.3.1 and can be read as follows: Under complete information (second column), D has knowledge of non-members' and members' exact cost and emission functions and the inverse demand function. Given incomplete information (third column), on the other hand, D possesses less knowledge about non-members' production technologies; that is, it knows that non-members' cost functions are increasing and convex and their emission functions increase in the quantities produced.[10]

Information...	complete (Section 3.3.1).	incomplete (following sections).
D knows...	c_j, e_j for all $j \in \mathcal{M}^C$, c_k, e_k for all $k \in \mathcal{M}$, $P\left(\sum_i q_i\right)$.	c_j increasing and convex, e_j increasing for all $j \in \mathcal{M}^C$, c_k, e_k for all $k \in \mathcal{M}$, $P\left(\sum_i q_i\right)$.

Table 3.1: Complete vs. incomplete information

Note that the two interventions are not sensitive to the exact timing of the game: the equilibrium emission levels do not change if non-members are not able to observe the club's output before setting their production levels; that is, non-members' best responses are not altered by such variation.

The two "incomplete-information robust" interventions correspond to solutions of optimisation problems (minE) and (maxQ) characterised below:

$$\min_{q_k \geq 0 \ \forall k \in \mathcal{M}} \quad \sum_{k \in \mathcal{M}} e_k(q_k) \quad \text{s.t.} \sum_{k \in \mathcal{M}} q_k = Q^0_{\mathcal{M}} \tag{minE}$$

$$\max_{q_k \geq 0 \ \forall k \in \mathcal{M}} \quad \sum_{k \in \mathcal{M}} q_k \quad \text{s.t.} \sum_{k \in \mathcal{M}} e_k(q_k) = E^0_{\mathcal{M}} \tag{maxQ}$$

Let $E^{\text{minE}}_{\mathcal{N}}$ and $E^{\text{maxQ}}_{\mathcal{N}}$ denote the equilibrium emissions upon implementation of solutions $\boldsymbol{q}^{\text{minE}}_{\mathcal{M}}$ and $\boldsymbol{q}^{\text{maxQ}}_{\mathcal{M}}$ to optimisation problems (minE) and (maxQ), respectively; then:

Proposition 9.
$E^{minE}_{\mathcal{N}} \leq E^0_{\mathcal{N}}$ and $E^{maxQ}_{\mathcal{N}} \leq E^0_{\mathcal{N}}$.
The statements hold with equality if and only if $e'_k(q^0_k) = e'_l(q^0_l)$ for all $k, l \in \mathcal{M}$ and club members' emission functions are not strictly concave.

[10] Note that knowledge of the inverse demand function and members' cost functions given incomplete information is only necessary for implementation of the desired quantities via taxes. In case D has the power to set members' production levels without incentivising them to choose them endogenously, it suffices to know that demand is decreasing in the price level.

For now, assume that there exist $l, k \in \mathcal{M}$ such that $e'_k(q_k^0) \neq e'_l(q_l^0)$ or emission functions are strictly concave.

First, consider problem (minE). The idea is simple: solution $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$ reallocates quantities within the club without changing the aggregate club-production. As emission-intensive members produce less and "clean" members produce more, within-club emissions are guaranteed to decrease. Furthermore, the constraint ensures non-members' quantities to stay constant: non-members' best responses are functions of the club's aggregate (and constant) production; hence, the pre-club non-member-quantities remain optimal. Accordingly, non-members' emission are constant, members' emissions decrease and the aggregate emission level decreases. Consumers are not harmed by the intervention either: post- and pre-founding price levels are identical and consumers equally well off.

Other than the solutions of (minE), those of (maxQ) holds constant the level of club-emissions. Put differently, the club is allowed to emit as much as it did *before* the founding. Accordingly, were the aggregate non-member-production not to change ($Q_{\mathcal{M}^C} = Q^0_{\mathcal{M}^C}$), emissions would not be altered either. The reduction in emissions via (maxQ) is implied by the maximisation of the club's production: by increasing its supply, the club is able to "crowd out" non-members' supply, thereby inducing reduced non-member-emissions. The constraint ensures the club's increased production not to compensate for decreased non-member-emissions. To paraphrase, the club optimally reallocates production among members, enabling them to produce more without changing within-club emissions. As higher club-quantities imply smaller non-member-quantities, non-members decrease their production/emissions. Importantly, in the new equilibrium, the aggregate price decreases as do total emissions.
This is not trivial: Non-member j's quantity is the best response to the clubs' *and* all other non-members' quantities; that is, non-members' quantities are interdependent. Accordingly, it is not clear whether the aggregate production post implementation of $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$ decreases or increases: could the increase in members' production be compensated by non-members? The following Lemma answers this question in the negative:

Lemma 16. *For an equilibrium club-production level $Q^1_{\mathcal{M}}$ and the corresponding aggregate equilibrium production level $Q^1_{\mathcal{N}}$:*

$$Q^0_{\mathcal{M}} < Q^1_{\mathcal{M}} \iff Q^0_{\mathcal{N}} < Q^1_{\mathcal{N}}.$$

To establish a rough understanding of this finding, consider firm j's first order con-

dition along with the subsequent rough but intuitive explanation:

$$\alpha - \beta Q_{-j} = c_j'(q_j) + 2\beta q_j$$

Suppose, Q_{-j} changes by one unit. This leads to a change in the LHS by $-\beta$. As the cost function is convex and the RHS contains the term $2\beta q_j$, it is not possible for j's quantity to change by more than Q_{-j} does. Heuristically, this ensures the change in Q_{-j} to not be overcompensated by a change in q_j. Accordingly, given implementation of $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$, the total quantity supplied is higher than that before the founding, as was to be established.

Note that in case emissions are strictly concave, equal marginal pre-club emissions do not present a problem, as, by the definition of concavity, an improvement can be achieved by letting the cleanest[11] member produce the club's total output. For a closer discussion of strictly concave emissions, see Observation 4.

Now, consider the case in which $e_k'(q_k^0) = e_l'(q_l^0)$ for all $k, l \in \mathcal{M}$ and club members' emission functions are not strictly concave; that is, in the absence of regulation, club members' quantities are distributed emission-efficiently within the club. As both interventions decrease emissions via an (emission-)efficient reallocation of production within the club, they are, in this case, not able to improve upon the pre-club equilibrium. To my mind, however, the requirement is unlikely to be met: for the above to be satisfied, all club members' emission *and* cost functions need to perfectly align such that all marginal emissions are equal in the equilibrium before the founding. To illustrate, under the assumption of uniform emission functions, this requires all cost functions to be exactly equal—a rather strong premise.

Last and importantly, both optimisation problems do not require any knowledge about non-members' maximisation problems or the exact shape of the demand function, making them rather attractive solution candidates in the presence of incomplete information.[12]

Having discussed the effects on emissions, I pivot to Corollary 2, which considers consumers' welfare. Define p^0 and $p^{\bar{\tau}}$ as the equilibrium price before the founding of the club and upon introduction of a uniform tax on carbon $\bar{\tau} > 0$ to be paid by all

[11] If there are multiple cleanest members, only one of them produces.

[12] Note that it is not possible to make general statements about their relative effectiveness: non-members' best response and emission functions determine which of the procedures is more effective in reducing emissions.

member countries, respectively. Similarly, p^{minE} and p^{maxQ} denote the prices upon implementation of $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$ and $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$, respectively; then:

Corollary 2.
$CS(p^0) = CS(p^{minE})$,
$CS(p^0) < CS(p^{maxQ})$,
$CS(p^{\bar{\tau}}) < CS(p^0)$.

As discussed above, the implementation of $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$ does, by definition, not change the aggregate supply and, hence, the price level. Accordingly, it does not affect consumers.
$\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$, on the other hand, reduces emissions due to the fact that members' increased production crowds out that of non-members. The fundamentals of the model imply the aggregate quantity to increase—in spite of the decreased non-member-production level. Clearly, this increases consumers' welfare compared to the equilibrium before the founding of the club.

These findings might seem obvious, but they are of relevance when comparing (minE) and (maxQ) to other standard climate policies: Not only is it unclear whether a uniform carbon tax decreases or increases the total level of emissions (see statement (1) in Proposition 8), such tax is also guaranteed to *decrease* consumers' welfare: $\bar{\tau}$ decreases the production by members for every price level. Accordingly, the club reduces its supply. This leads to an increased price level that, in turn, harms consumers. This phenomenon is also known as *carbon cost pass-through*.[13]

Taken together, Proposition 9 and Corollary 2 indicate that the implementation of either (minE)'s or (maxQ)'s solutions is, for virtually all parametrisations, sure to reduce emissions *and* does not harm consumers—no matter non-members' characteristics. Furthermore, their implementation does, contrary to that of $\boldsymbol{q}_{\mathcal{N}}^{*}$, not require in-depth knowledge of non-members' characteristics.

A general understanding at hand, one may wonder about the exact form of the respective solutions. Unsurprisingly, the implied best response functions depend on members' emission functions: both $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$ and $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$ reduce emissions, in part, via an emission-efficient reallocation of the quantities produced by club members; naturally, such optimal reallocation differs significantly across various types of emission functions. The following observation characterises (minE)'s and (maxQ)'s solutions conditional on the shapes of members' e_k.

[13] Ganapati, Shapiro and Walker (2020) provide a detailed literature review not limited to but also on pass-through due to climate policies such as carbon taxes.

Note that the following assumption is made to simplify the statement of the case with convex costs.

Assumption (Conv). *For all $k \in \mathcal{M}$: e_k is strictly convex, e_k' is continuous and invertible on $\mathbb{R}_+$; c_k' exists and is continuous.*

Let

$$Q_{\mathcal{M}}^{e'}(x) := \sum_{k\in\mathcal{M}} e'^{-1}_k(x) = \sum_{k\in\mathcal{M}} \{q_k : e_k'(q_k) = x\}$$
$$\left[E_{\mathcal{M}} \circ Q_{\mathcal{M}}^{e'}\right](x) := \sum_{k\in\mathcal{M}} e_k\left(e'^{-1}_k(x)\right) = \sum_{k\in\mathcal{M}} e_k(\{q_k : e_k'(q_k) = x\}),$$

where $x \in \mathbb{R}^+$. Accordingly, $Q_{\mathcal{M}}^{e'}(x)$ measures the sum of produced quantities given all members' marginal emissions are equal to x. $\left[E_{\mathcal{M}} \circ Q_{\mathcal{M}}^{e'}\right](x)$, on the other hand, measures the emissions upon production of quantities leading to marginal emissions x for all members. Note that, given Assumption (Conv), both functions are invertible.[14]

Observation 4.
If (Conv) for all $k \in \mathcal{M}$, then
(minE): q_k^{minE} such that $e_k'(q_k^{minE}) = \left[Q_{\mathcal{M}}^{e'}\right]^{-1}\left(Q_{\mathcal{M}}^0\right)\ \forall k$.
(maxQ): q_k^{maxQ} such that $e_k'(q_k^{maxQ}) = \left[E_{\mathcal{M}} \circ Q_{\mathcal{M}}^{e'}\right]^{-1}(E_{\mathcal{M}}^0)\ \forall k$.

If e_k is strictly concave and unbounded for all $k \in \mathcal{M}$, then
(minE): $\exists!\, k$ such that $q_k^{minE} = Q_{\mathcal{M}}^0$, where
$$e_k\left(Q_{\mathcal{M}}^0\right) \in \min\left\{e_1\left(Q_{\mathcal{M}}^0\right), ..., e_m\left(Q_{\mathcal{M}}^0\right)\right\}.$$
(maxQ): $\exists!\, k$ s.t. $q_k^{maxQ} = e_k^{-1}(E_{\mathcal{M}}^0)$, where
$$e_k^{-1}(E_{\mathcal{M}}^0) \in \max\left\{e_1^{-1}(E_{\mathcal{M}}^0), ..., e_m^{-1}(E_{\mathcal{M}}^0)\right\}.$$

If $e_k(q) = \eta_k q$ for all $k \in \mathcal{M}$, then $\forall Q_{\mathcal{M}^C}$:
$$q_k^{minE} = q_k^{maxQ} = 0\ \forall k \text{ such that } \eta_k \notin \min\left\{\eta_1, ..., \eta_m\right\}.$$

First, consider convex emissions—the, to my mind (see Section 3.2), most realistic case: Given convexity, it would be harmful to only have one member produce the aggregate club-output. Instead, output is distributed among members such that all members' marginal emissions are equal. Were this not the case, the club could improve the outcome by reducing one member's quantity (i.e. lower its marginal emissions) and let another member produce more instead (i.e. increase its marginal emissions). Doing so, the club could decrease emissions while keeping

[14] This follows from Lemmas 20 and 21.

the output constant (minE) or increase its output while keeping emissions constant (maxQ). The exact production quantities can be interpreted as follows: If all members produce such that their marginal emissions equal $\left[Q^{e'}_{\mathcal{M}}\right]^{-1}\left(Q^0_{\mathcal{M}}\right)$, members' quantities exactly sum up to $Q^0_{\mathcal{M}}$. Accordingly, for $e'_k(q_k^{\text{minE}}) = \left[Q^{e'}_{\mathcal{M}}\right]^{-1}\left(Q^0_{\mathcal{M}}\right)$ for all k, quantities are distributed emission-efficiently (as marginal emissions are uniform) and the aggregate club-output is equal to $Q^0_{\mathcal{M}}$—this perfectly achieves the objective of (minE). Similarly, $\left[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\right]^{-1}(E^0_{\mathcal{M}})$ is the value of marginal emissions that guarantees the club to emit $E^0_{\mathcal{M}}$, provided that members' marginal emissions are uniform. This aligns with optimisation problem (maxQ), as it ensures emission-efficient distribution of quantities and a club-emission of exactly $E^0_{\mathcal{M}}$ units of carbon. Together, the above imply: if, for instance, all members have quadratic cost functions and a uniform quadratic emission function (cf. Observation 3), q_k^{minE} is equal to the average member-production level before the founding $1/m \sum_{k\in\mathcal{M}} q_k^0$ and q_k^{maxQ} equals the quadratic mean of the pre-founding member-production levels $\sqrt{1/m \sum_{k\in\mathcal{M}} (q_k^0)^2}$.

In the second statement, I consider unbounded[15] concave functions: for (minE), only the member with the lowest emission level upon production of the total quantity $Q^0_{\mathcal{M}}$ is allowed to produce. Even if there are two members with the exact same emission function, only one of them is asked to supply a positive quantity; this follows from the fact that any concave function f with $f(0) = 0$ is subadditive. For (maxQ), only the member with the highest output (given emission level $E^0_{\mathcal{M}}$) is allowed to produce. Note that, as emission functions may intersect, $q_k^{\text{minE}} > 0$ does not necessarily imply $q_k^{\text{maxQ}} > 0$ as well.[16]

Lastly, consider linear emission functions—a rather intuitive and simple case: for both (minE) and (maxQ), the cleanest member produces the total quantity. As the member with the lowest level of emissions given a level of production, it is also able to produce the highest output, given a fixed level of emissions. Accordingly, the solutions are almost equivalent; that is, they differ in the level of production but not—other than for strictly concave emission functions—in the set of members allowed to produce.

So far, this analysis has (1) highlighted the risk of adverse effects, (2) emphasised the importance of complete information for the implementation of the optimal

[15] Note that the above does not consider bounded concave emission functions as these would imply there being levels of $E^0_{\mathcal{M}}$ for which at least one member produces $q_k = \infty$.

[16] That is, the unique producing member may not be the same in $\boldsymbol{q}^{\text{minE}}_{\mathcal{M}}$ as in $\boldsymbol{q}^{\text{maxQ}}_{\mathcal{M}}$.

quantities and (3) discussed alternative solutions that are robust to incomplete information about non-members. The existence of such interventions is good news in principle but remains agnostic regarding their implementability: How could club members be incentivised to produce as suggested? Do they need to be forced to supply the exact targeted quantities or could they be nudged to do so via taxes or certificates? The next section is dedicated to the analysis and comparative statics of tax schemes that may be used to implement the quantities implied by (minE) and (maxQ); it shows that for many parametrisations, the two incomplete-information robust interventions can be implemented via tax schemes harming neither consumers nor club members.

3.3.3 Implementation via tax schemes

A rather natural approach to regulating club members' production is the introduction of a *tax scheme* (i.e. a vector of tax functions). Besides a uniform price on carbon, such tax scheme could also entail a member-specific tax system or level: depending on the club's targeted production quantity, some members may have to be provided with more, others with less incentives to produce as demanded by D. A tax scheme that internalises differences in members' resources and production technologies is likely to be more efficient than a uniform one. Accordingly, for most parametrisations, the quantities implied by (minE) and (maxQ) cannot be implemented via "uniform" tax schemes.[17]

Note that Assumption (Conv) ensures the existence of such tax schemes. However, a violation thereof does not imply implementation via tax schemes to be impossible but simply more involved.

Proposition 10. *Suppose (Conv). There exist vectors* $\left(\tau_k\right)_{k\in\mathcal{M}}$ *of continuous tax functions* $\tau_k(q_k, Q_{-k})$ *(linear in* q_k*) for all* $k \in \mathcal{M}$ *that implement the quantities implied by problems (minE) and (maxQ) in a Cournot Nash Equilibrium, respectively.*

Note that a similar result can be obtained for implementation of the optimal quantities $(q_1^*, ..., q_N^*)$ and is not stated explicitly, as I am mainly interested in the incomplete-information set-up.

One might wonder why I highlight the continuity of the tax functions. This is to exclude a type of quantity regulation wherein firms have to pay fines that render unprofitable all bar the precise production quantity demanded by the club's policy. Such a scheme leaves firms with no choice but to produce the desired quantity,

[17] Most tax systems differentiate between subjects: income taxes, for instance, vary significantly across income classes, individuals' characteristics and sources of income.

which may prove highly inefficient if said production target has been set suboptimally. In stark constrast to this, continuous tax functions induce firms to choose output levels that are more robust to tax-misspecifications as the cost of adjustment to new targets is likely smaller than in the former case.

Note that, from the perspective of the tax payers (i.e. members), the functions are not only continuous but also *linear*: below, τ_k will be shown to be a function of q_k and Q_{-k} for all $k \in \mathcal{M}$; as firms cannot influence Q_{-k}, the tax functions are univariate from their point of view. Furthermore, treating Q_{-k} as a parameter, $\tau_k(q_k|Q_{-k})$ is linear: Consider problem (minE).[18] As I am seeking to define a tax function that changes firm k's maximisation problem such that it endogenously chooses q_k^{minE}, I need to find τ_k^{minE} such that

$$\begin{aligned}
&\frac{\partial\left(\pi_k(q_k,Q_{-k})-\tau_k^{\text{minE}}(q_k,Q_{-k})\right)}{\partial q_k} = \\
&\alpha-\beta Q_{-k}-2\beta q_k-c_k'(q_k)-\frac{\partial \tau_k^{\text{minE}}(q_k,Q_{-k})}{\partial q_k}=0 \\
&\iff q_k = q_k^{\text{minE}}\ \forall Q_{-k}\geq 0.
\end{aligned}$$

To do so, I determine the value of $\alpha-\beta Q_{-k}$ at which member k would produce q_k^{minE} *in the absence of regulation*. By the first order conditions, this is the case for $\alpha-\beta Q_{-k} = c_k'(q_k^{\text{minE}})+2\beta q_k^{\text{minE}}$. Hence, defining the tax functions accordingly, the firms' first order conditions can be manipulated such that members produce exactly as needed:[19]

Definition 8. *Define two tax schemes:*

$$\begin{aligned}
\boldsymbol{\tau}^{minE} &:= \left(\tau_k^{minE}(q_k,Q_{-k})\right)_{k\in\mathcal{M}} \text{ with} \\
&\tau_k^{minE}(q_k,Q_{-k}) := q_k(\alpha-\beta Q_{-k}-c_k'(q_k^{minE})-2\beta q_k^{minE}).
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\tau}^{maxQ} &:= \left(\tau_k^{maxQ}(q_k,Q_{-k})\right)_{k\in\mathcal{M}} \text{ with} \\
&\tau_k^{maxQ}(q_k,Q_{-k}) := q_k(\alpha-\beta Q_{-k}-c_k'(q_k^{maxQ})-2\beta q_k^{maxQ}).
\end{aligned}$$

[18] Derivations for (maxQ) are analogous.

[19] In case Assumption (Conv) is violated *and* no continuous tax function can be determined, the procedure may be more involved. In some cases, the best the club can do is to supply members with a table indicating the targeted quantity for every level of Q_{-k}. Any deviation from said level would be penalised with a fine which makes compliance the best response. Such fines would either be very high for every firm (likely rather inefficient) or different for every firm and price level. By contrast, in markets that allow for continuous tax functions, firms are simply informed about their respective tax function, include it in their profit function and pay/get paid just enough to endogenously choose the correct production level.

The process of determining above tax functions may seem rather involved. Luckily, it can be shown that even implementation of quantity vectors that are similar to, but do not equal the solutions of (minE) and (maxQ) can decrease emissions. If, for instance, emission functions are convex and do not vary "too much" across members, it suffices to incentivise all club members to produce the arithmetic or quadratic mean of the pre-club quantities, which are equal to the respective solutions if all members have the *same* emission function. Accordingly, even an "approximate" implementation of $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$ and $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$ is likely to decrease emissions.

Especially in light of the result on adverse effects (Proposition 8), the existence of tax schemes that are sure to reduce emissions (for virtually all parametrisations) and do not require knowledge of non-members' production and emission functions is encouraging. Unsurprisingly, however, such reduction comes at a cost. Assuming (Quad), tax schemes $\boldsymbol{\tau}^{\text{minE}}$ and $\boldsymbol{\tau}^{\text{maxQ}}$ generate negative revenues, with the latter being more costly than the former; that is:

Assumption (Quad). *$e_k(q_k) = \eta q_k^2$, $\eta > 0$ and $c_k(q_k) = 0.5\gamma_k q_k^2$, $\gamma_k > 0$ for all $k \in \mathcal{M}$. Furthermore, $\exists\, l, k \in \mathcal{M}$ such that $\gamma_l \neq \gamma_k$.*

Theorem 5. *Suppose (Quad).*

$$\sum_{k\in\mathcal{M}} \tau_k^{maxQ}(q_k^{maxQ}, Q_{-k}^{maxQ}) < \sum_{k\in\mathcal{M}} \tau_k^{minE}(q_k^{minE}, Q_{-k}^{minE}) < 0.$$

Recall, while $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$ keeps the total quantity produced by members constant, $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$ ensures the club's emissions to remain unchanged. To achieve this without increasing emissions, quantities need to be reallocated (emission-)efficiently among members. As I am looking for tax functions that make firms endogenously choose the correct production levels and members themselves do not care about emissions but costs only, tax functions need to be tailored to members' cost functions. Roughly speaking, this implies the above inequalities.

To illustrate, note that, given Assumption (Quad), all members have the same emission function. In such scenario, scheme $\boldsymbol{\tau}^{\text{minE}}$ incentivises all members to produce the same quantity (i.e. the average pre-club quantity $\bar{q}_{\mathcal{M}}$). Now, compare members l and k and assume $q_l^0 - \varepsilon = \bar{q}_{\mathcal{M}}$, $q_k^0 + \varepsilon = \bar{q}_{\mathcal{M}}$ for $\varepsilon > 0$. To implement the optimal quantity, member l needs to pay taxes, while member k's production must be subsidised. Member l must have smaller marginal costs at $\bar{q}_{\mathcal{M}}$ than k—otherwise, it would have chosen a smaller, not a higher quantity in the equilibrium before the founding. As all cost functions are assumed to be strictly convex, the value of the tax needed to increase member l's marginal costs at $\bar{q}_{\mathcal{M}}$ is smaller than

the value of member k's subsidy. Taken together, the balance of $\boldsymbol{\tau}^{\text{minE}}$ is negative.

For $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$, which increases the total club-production, a strictly negative balance is even less surprising: on aggregate, production must be subsidised as firms need to be compensated for the individually suboptimally high production levels. Note that there may still be members with $\tau_k(q_k^{\text{maxQ}}, Q_{-k}^{\text{maxQ}}) > 0$; on average, however, production is subsidised. Roughly speaking, the sum of subsidies paid to firms is higher than that given implementation of $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$, as the club produces more in the former than the latter case.

Unfortunately, it is not possible to make general statements about the ranking of the interventions with respect to their effect on emissions: the aggregate level of emissions given $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$ depends on non-members' reaction to the now higher club-production. Luckily, a ranking with respect to firms' profits, on the other hand, is possible: given Assumption (Quad), for both $\boldsymbol{\tau}^{\text{minE}}$ and $\boldsymbol{\tau}^{\text{maxQ}}$, members' net profits increase compared to the pre-club equilibrium with the latter implying strictly higher profits than the former.

Theorem 6. *Suppose (Quad).*

$$
\begin{aligned}
&0 < \sum_{k\in\mathcal{M}} \pi_k(q_k^0, Q_{-k}^0) < \\
&\sum_{k\in\mathcal{M}} \pi_k(q_k^{minE}, Q_{-k}^{minE}) - \sum_{k\in\mathcal{M}} \tau_k^{minE}(q_k^{minE}, Q_{-k}^{minE}) < \\
&\sum_{k\in\mathcal{M}} \pi_k(q_k^{maxQ}, Q_{-k}^{maxQ}) - \sum_{k\in\mathcal{M}} \tau_k^{maxQ}(q_k^{maxQ}, Q_{-k}^{maxQ}).
\end{aligned}
$$

Simply put, club members profit from their increased market share given $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$: subsidised by the tax system, they are able to produce more and increase their profits. Given $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$, on the other hand, the aggregate club supply remains constant. Profits are, hence, smaller in this case compared to those given $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$. They do, however, increase compared to the pre-club equilibrium, as, on aggregate, members receive subsidies (cf. Theorem 5).[20] This also implies that the introduction of a lump-sum tax on club members could recuperate some of the losses due to the unbalanced tax schemes. For $\boldsymbol{\tau}^{\text{minE}}$, this would fully compensate for the subsidies needed to incentivise members without reducing their profits. $\boldsymbol{\tau}^{\text{maxQ}}$, on the other hand, does not allow for such general statement, as the price p^{maxQ} depends on non-members' cost functions; accordingly, the lump sum taxes may or may not fully recuperate the losses without reducing members' profits. In light of this finding and given the limited knowledge the club may have, scheme $\boldsymbol{\tau}^{\text{minE}}$ could be preferred

[20] Given that both tax schemes increase firms' profits, D could introduce a lump-sum tax on club members to recuperate some of the losses due to the unbalanced tax schemes.

over τ^{maxQ}.

To conclude, this section illustrated that, for many parametrisations, it is possible to incentivise club members to endogenously choose the quantities consistent with (minE) and (maxQ), respectively, while not decreasing the sum of member-firms' profits.[21] This finding is of importance when it comes to welfare considerations, as it implies neither consumers nor the club members to be harmed by the implementation of the two incomplete-information robust interventions.

3.4 Related literature

Arguably, one of the most acclaimed works on climate clubs is Nobel laureate William Nordhaus' "Climate Clubs: Overcoming Free-Riding in International Climate Policy." (Nordhaus, 2015). His analysis incorporates both theoretical and empirical arguments and leads him to propose that members of climate clubs should not only jointly implement climate policies but also penalise non-members that do not comply with them. Take, for instance, the introduction of a price on carbon: a climate club à la Nordhaus would require all its members to introduce a tariff on emission-intensive goods from countries in which there is either no or a too small price on carbon—be they members or non-members. While Nordhaus (2015) is not the first analysis on characteristics of club-like structures in the context of climate economics (cf. Barrett, 1994; Finus, Altamirano-Cabrera and Van Ierlan, 2005; Bosetti et al., 2013), William Nordhaus seems to have been the first to introduce the term "climate club". Importantly, he emphasised a mechanism-design component that had not been the main focus of existing studies. Instead of mainly characterising climate alliances/coalitions, he motivated a shift towards a more solution-oriented, normative rather than merely descriptive strand of literature. To my mind, this might well have been one of the reasons his ideas were heard by a very broad audience and even implemented (cf. G7 Climate Club). In a way, he presented the club as an opportunity/a solution instead of describing existing climate alliances and pointing at potential flaws thereof.

Nordhaus' work has inspired many analyses since, most of which incorporate empirical estimates into model-based simulations (eg. Hovi et al., 2019; Sælen, 2016). In line with Nordhaus (2015), these analyses do not consider markets separately but evaluate the effects of climate clubs based on aggregate variables. To capture effects and risks on the micro level, a complementary strand of literature has evolved: studies such as Hoel (1991), Babiker (2005), Yomogida and Tarui

[21] That is, not leading to lower aggregate club-profits than those in the pre-club equilibrium.

(2013) and Baccianti and Schenker (2021) analyse carbon leakage as a result of competition on a given market in isolation. Note that, other than one might expect, carbon leakage is defined as the reduction in regulated entities' emissions accompanied by an increase in emissions by unregulated ones; importantly, this does *not* imply an increase in aggregate emissions.

The articles in this strand of literature related closest to my paper—a working paper by Robert Ritz (2009)[22] and a published article by Meredith Fowlie (2009)—analyse Cournot competition in the presence of a carbon tax that has to be paid by a *subset* of firms only ("incomplete" regulation). The central aspects distinguishing the two from my analysis lie in (1) the papers' main objective and (2) the respective properties of emission and cost functions.

First, consider the most important aspect (1): Ritz (2009)'s and Fowlie (2009)'s main focus lies in providing (simulation-based) estimates for the effect of regulation in specific markets. Importantly, both analyses put emphasis on scenarios in which regulation leads to carbon leakage, not necessarily an increase in *total* emissions. This paper, on the other hand, is qualitative in nature and focuses on adverse effects on said *total* emission level.[23] Furthermore, its main contribution lies in the discussion of interventions that alleviate the risk of adverse effects, while Ritz (2009) and Fowlie (2009) take the type of regulation as given (i.e. uniform tax on carbon for all members) and estimate effects thereof. Put differently, in this paper, the type of regulation is a choice variable, not, as in Ritz (2009) and Fowlie (2009), a parameter with restricted range.

Besides that, (2) both studies assume carbon emissions to be linear in quantities produced,[24] while my most important results are not restricted to such specific functional form of costs and emission functions; in Fowlie (2009), costs of production are linear as well. As one would expect, such strong assumptions deliver strong results:[25] in Fowlie (2009) for instance, the occurrence of adverse effects depends only on the relationship between regulated and unregulated entities' emission functions, not on the policy (i.e. the level of the carbon tax). Put differently, given a set of regulated firms, either all (incomplete) carbon tax systems reduce emissions or

[22] There is another working paper (Neuhoff and Ritz, 2019) that contains parts of the working paper Ritz (2009) but focuses on the (mostly empirical) analysis of carbon cost pass-through.

[23] In fact, in my model, carbon leakage is always positive if there is a uniform price on carbon to be paid by members (Observation 5). This, however, does not imply total emissions to be higher after than before the founding.

[24] Besides that, Ritz (2009) assumes firms subject to carbon taxation (cf. club members) and those unaffected by carbon policies (cf. non-members) to have uniform and linear emission functions, respectively.

[25] Similarly, in Ritz (2009), post-regulation emissions can exceed pre-regulation emissions only if unregulated entities have "dirtier" production technologies. In my model, this is not the case; even if all entities face the same emission function, leakage can exceed 100 percent.

none of them does. My findings are more encouraging as I am able to characterise interventions that never increase emissions.[26] Furthermore, in her model, complete regulation (cf. all firms are club members) can be inferior to incomplete regulation—a somewhat counterintuitive result that does not apply to the interventions I suggest; in my framework, more club members are always preferable.[27]

Lastly, I would like to mention the literature on regulation in the presence of uncertain externalities: the two tax schemes suggested in Section 3.3.3 could be understood as solutions to a mechanism design problem with an uncertain social choice correspondence. In case the club (i.e. the regulator or mechanism designer) has no power over and limited knowledge about non-members, the exact level of aggregate emissions (i.e. the externality) resulting from an intervention is uncertain. Accordingly, tax schemes $\boldsymbol{\tau}^{\text{minE}}$ and $\boldsymbol{\tau}^{\text{maxQ}}$ could be understood as mechanisms that implement a socially acceptable outcome *regardless of* the uncertainty. Such environments have been studied before, for instance by Lee and Park (2010) who analyse a Cournot game with free entry. They show that if the externality varies exogenously in aggregate output, a combination of output taxes and entrance fees is able to implement the social optimum. Note that their findings cannot be applied to the model above as the authors assume all firms' objective functions to be equal and market entry to be possible. For further similar analyses, see Koenig (1985) and McKitrick (1999).

3.5 Conclusion and discussion

A climate club is a group of countries whose aim lies in the joint implementation of climate policies, such as emission-reducing regulations. I analyse the effect of the founding of such climate club in a market for an emission-intensive good (Cournot competition) and highlight the risk of interventions that reduce the club's production; that is, reduced production by club members can increase the total level of carbon emissions: in the new equilibrium after the founding of the climate club, non-member countries raise their supply and, hence, emissions; such increase may compensate for club members' reduced emissions, resulting in a higher total emission level after than before the intervention. For some parametrisations, implemen-

[26] Furthermore, it can be shown that, given a fixed set of parameter values, a uniform tax on carbon to be paid by members can, depending on its value, increase or decrease the total level of emissions. In Fowlie (2009), this would not be possible as parameters—not the price itself—determine whether emissions are higher after implementation than before.

[27] That is, more club members are always preferable given one of the interventions discussed in Section 3.3 is used (Observation 6).

tation of the emission-minimising club-quantities leads to *higher* within-club emissions than before the founding: especially when non-members are highly reactive to their competitors' supply, the club may be better off increasing its own emissions and output, thereby decreasing the price and indirectly causing non-members to (heavily) cut back on theirs.
Both the risk of adverse effects and the optimal intervention are highly dependent on the exact shape of non-members' emission and production functions, which are likely unknown to the club itself. As a remedy, I characterise two interventions that are, for virtually all parametrisations, sure to reduce emissions, not harm consumers and do not require club members to possess such detailed information. To close the main part of the analysis, I discuss the implementation of the two "incomplete-information robust" interventions via tax schemes.
My results do not hinge on the assumption of imperfect competition; assuming firms to be price takers does not change the findings qualitatively.

To summarise, this paper highlights the importance of extensive market analyses preceding interventions and emphasises the need for accurate, closely screened and reviewed greenhouse gas emission reporting to avoid adverse effects of market interventions—particularly in the absence of effective demand-targeted regulation. Furthermore, it characterises interventions that are robust to both adverse effects and the lack of such detailed analyses and reporting schemes. These could be employed until the needed infrastructure is in place and data readily available.

Appendix 3.A

Convex emission functions

In this section, I discuss the assumption of convex emissions and relate it to both empirical and theoretical literature.

Roughly speaking, convex emissions are an indication of increasing inefficiencies in the production process: if, for instance, the efficiency of machines decreases in the quantity produced due to a built up of heat or longer per-unit production times, increasing average emission levels are fairly plausible. Accordingly, when costs are convex and there exists a close relationship between costs and emissions, increasing marginal emissions seem to be appropriate.

Furthermore, as emission-intensive goods are typically energy-intensive (cf. International Energy Association, 2020), a higher production level requires a higher input of energy. Given the limited supply of sustainable energy sources, firms need to resort to fossil fuels and the like, indicating increasing marginal emissions. For instance, Holland et al. (2022) find that marginal CO_2 emissions are increasing in the US electricity sector. This is due to an increased reliance on coal to satisfy the elevated demand for electricity, they argue. Their finding is of particular importance for the evaluation of policies that aim to increase the use of electric vehicles: they find that, without complementary policies decarbonising the electricity sector, the increased demand for electricity may offset more than half of the emission reductions caused by the decreased use of gasoline powered vehicles.

Furthermore, a firm may need to increase the share of imported resources when raising its output which, naturally, leads to higher transportation-related emissions.

Moreover, note that emissions in both steel and iron production, for instance, are decreasing in the quota of recycled material ("scrap") used. As the demand for both materials has seen significant increases, such recycled material is scarce. Naturally, the more a firm increases its production, the more it relies on other (not recycled) inputs. This implies convexity in the emission function. For a detailed description of determinants of emissions in the production of steel and iron, see International Energy Association (2020).[28]

Lastly, the assumption of convex emissions is not new to economic modelling (cf. e.g. David and Sinclair-Desgagné, 2005; Lazkano, Marrouch and Nkuiya, 2016; Mason, Polasky and Tarui, 2017; Dietz and Venmans, 2019, etc.):[29]

[28] The report is the source for all information contained in the paragraph.

[29] Lazkano, Marrouch and Nkuiya (2016) and Mason, Polasky and Tarui (2017) model the positive effect of emissions on the output of a production process to be concave. This is (David and Sinclair-Desgagné, 2005, and as argued above) essentially equivalent to the assumption of convex emissions.

Generally speaking, there are two approaches in modelling emissions: (1) emissions as an input of a production process, (2) emissions as an output. As shown by Ebert and Welsch (2007), considering the "materials balance principle", the two approaches are equivalent; that is, taking into account that matter can neither be wasted nor created. Simply put, if the production technology is concave in inputs and emissions and the products to be sold are the unique outputs, emissions are convex: an increase in inputs does not correspond to a proportional increase in production; accordingly, marginal emissions must be increasing—otherwise the materials balance would not hold. Similarly, emissions can be modelled as an input that—as all other inputs—has decreasing marginal effects on the output.

Appendix 3.B

Proofs and omitted results

Proof of Proposition 8. Assuming $c_j(q_j) = 0.5c_{\mathcal{M}^C}q_j^2$, $e_j(q_j) = \eta_{\mathcal{M}^C}q_j^2$ for all $j \in \mathcal{M}^C$ and $c_k(q_k) = 0.5c_{\mathcal{M}}q_k^2$, $e_k(q_k) = \eta_{\mathcal{M}}q_k^2$ for all $k \in \mathcal{M}$, members' pre-club quantities and non-members' best response functions to members' quantity q_k take the following form:

$$q_k^0 = \frac{\alpha(c_{\mathcal{M}^C} + \beta)}{(c_{\mathcal{M}} + \beta m + \beta)(\beta(n-m+1) + c_{\mathcal{M}^C}) - \beta^2(n-m)m}$$
$$q_j^{\mathrm{BR}}(q_k) = \frac{\alpha - \beta m q_k}{\beta(n-m+1) + c_{\mathcal{M}^C}}$$

Accordingly, the emission-minimising club-quantities $q^{\min}$ solve the following problem:

$$\min_{q \geq 0} \eta_{\mathcal{M}} m q^2 + \eta_{\mathcal{M}^C}(n-m)\left(\frac{\alpha - \beta m q}{\beta(n-m+1) + c_{\mathcal{M}^C}}\right)^2$$

The first order conditions imply and some simple algebra imply: the emission-minimising $q^{\min}$ is equal to

$$q^{\min} = \frac{(n-m)\alpha\beta\eta_{\mathcal{M}^2}}{\eta_{\mathcal{M}}(\beta(n-m+1) + c_{\mathcal{M}^C})^2 + \eta_{\mathcal{M}^C}\beta^2 m(n-m)}.$$

Note that the objective function is convex (and continuous) in q; accordingly the above is a minimiser. Due to the convexity, the first derivative of the aggregate emission function above is negative for all $q < q^{\min}$. If now $q_k^0 < q^{\min}$, one can establish that any reduction in members' quantities to levels smaller than q_k^0 yield

higher emissions than in the equilibrium before the founding. Note also that a reduction in members' quantities to levels smaller than q_k^0 such that not all members produce the same quantity also increases emissions. All that matters is the aggregate club-quantity as it determines non-members' quantities. Given an aggregate club-quantity, emissions are increased even more if members produce non-uniform quantities (as all of them have the same convex emission function). Furthermore, if $q_k^0 < q^{\min}$, $\eta_{\mathcal{M}} m (q_k^0)^2 < \eta_{\mathcal{M}} m (q^{\min})^2$ and within-club emissions increase. By some simple reordering of the below condition, the following can be shown to be a necessary and sufficient condition for $q_k^0 < q^{\min}$:

$$\frac{\alpha(c_{\mathcal{M}^C}+\beta)}{(c_{\mathcal{M}}+\beta m+\beta)(\beta(n-m+1)+c_{\mathcal{M}^C})-\beta^2(n-m)m} < \frac{(n-m)\alpha\beta\eta_{\mathcal{M}^2}}{\eta_{\mathcal{M}}(\beta(n-m+1)+c_{\mathcal{M}^C})^2+\eta_{\mathcal{M}^C}\beta^2 m(n-m)} \iff \frac{\eta_{\mathcal{M}^C}}{\eta_{\mathcal{M}}}\frac{\beta+c_{\mathcal{M}}}{\beta+c_{\mathcal{M}^C}}\frac{\beta(n-m)}{\beta(n-m+1)+c_{\mathcal{M}^C}} > 1.$$

Note that by firms' first order conditions in the equilibrium before the founding, $\frac{q_j^0}{q_k^0} = \frac{\beta+c_{\mathcal{M}}}{\beta+c_{\mathcal{M}^C}}$. This establishes the result. □

Proof of Observation 3. The proof follows from that of Proposition 8. □

Proof of Lemma 15. Consider the Lagrangian and first order conditions implied by (∗): Let $\boldsymbol{v}$ be a vector of length N. θ_j are scalars for all $l \in \mathcal{M}^C$.

$$\min_{\boldsymbol{q}_{\mathcal{N}}} \quad (1-\omega)\Bigg(\sum_{k\in\mathcal{M}} e_k(q_k)+\sum_{j\in\mathcal{M}^C} e_j(q_j)\Bigg)-\omega\, 0.5\beta\Big(\textstyle\sum_{i\in\mathcal{N}} q_i\Big)^2 + \sum_{j\in\mathcal{M}^C}\theta_j\big(\alpha-\beta\textstyle\sum_{i\in\mathcal{N}} q_i - c_j'(q_j)-\beta q_j\big)-(q_1,...,q_N)\cdot\boldsymbol{v}$$

Then, for the solution $(q_1^*, ..., q_N^*)^T$:

$$(1-\omega)e_k'(q_k^*) = \omega\beta \sum_{i\in\mathcal{N}} q_i^* + \sum_{j\in\mathcal{M}^C} \theta_j\beta + \upsilon_k \ \forall k \in \mathcal{M}$$

$$(1-\omega)e_j'(q_j^*) = \omega\beta \sum_{i\in\mathcal{N}} q_i^* + \sum_{j\in\mathcal{M}^C} \theta_j\beta + c_j''(q_j^*) + \beta + \upsilon_j \ \forall j \in \mathcal{M}^C$$

$$(q_1^*, ..., q_N^*) \cdot \boldsymbol{\upsilon} = 0$$

$$\theta_j \left(\alpha - \beta \textstyle\sum_{i\in\mathcal{N}} q_i^* - c_j'(q_j^*) - \beta q_j^*\right) = 0 \ \ \forall j \in \mathcal{M}^C \Rightarrow$$

$$(1-\omega)e_k'(q_k^*) = (1-\omega)e_j'(q_j^*) - c_j''(q_j^*) - \beta \ \forall k \in \mathcal{M},\ j \in \mathcal{M}^C \text{ s.t. } q_k^*, q_j^* > 0$$

$$e_k'(q_k^*) = e_j'(q_j^*) - \frac{c_j''(q_j^*) + \beta}{1-\omega} \ \forall k \in \mathcal{M},\ j \in \mathcal{M}^C \text{ s.t. } q_k^*, q_j^* > 0$$

$$e_j'(q_j^*) = e_k'(q_k^*) + \frac{c_j''(q_j^*) + \beta}{1-\omega} \ \forall k \in \mathcal{M},\ j \in \mathcal{M}^C \text{ s.t. } q_k^*, q_j^* > 0$$

Note that the solution implied by the above constitutes a subgame perfect Nash equilibrium, as D optimises its objective subject to non-members' best response functions. Accordingly, the solution induces an equilibrium in the subgame in which non-members react to the club-output determined by D. □

Proof of Lemma 16. This result is a slightly adjusted version of the result on reaction functions[30] in Dixit (1986) (p.118 pp); the proof is almost equivalent. Non-members' reaction functions to changes in the club's output $Q_{\mathcal{M}}$ are implicitly defined by their respective first order conditions, treating $Q_{\mathcal{M}}$ as a parameter. To establish the Lemma, I show that the aggregate non-member-reaction function facing the club (hereinafter $R_{\mathcal{M}^C} := \frac{dQ_{\mathcal{M}^C}}{dQ_{\mathcal{M}}}$) has a slope strictly higher than -1 and smaller than 0 at all points. To do so, note that in equilibrium (internal solution exists due to the fundamental assumptions)

$$0 = -q_j\beta - c_j'(q_j) - \beta \sum_i q_i + \alpha =: \xi_j$$

$$-2\beta - c_j''(q_j) = \frac{\partial \xi_j}{\partial q_j} =: a_j$$

$$-\beta = \frac{\partial \xi_j}{\partial Q_{-j}} =: b_j.$$

[30] The definition of the term "reaction functions" in this context is due to Perry (1982).

Accordingly,

$$a_j dq_j + b_j dQ_{-j} = 0 \iff a_j dq_j - b_j dq_j + b_j dQ_{\mathcal{M}^C} + b_j dQ_{\mathcal{M}} = 0 \iff$$
$$dq_j + \frac{b_j}{a_j - b_j} dQ_{\mathcal{M}^C} + \frac{b_j}{a_j - b_j} dQ_{\mathcal{M}} = 0 \iff$$
$$dQ_{\mathcal{M}^C} + \sum_{j \in \mathcal{M}^C} \frac{b_j}{a_j - b_j} dQ_{\mathcal{M}^C} + \sum_{j \in \mathcal{M}^C} \frac{b_j}{a_j - b_j} dQ_{\mathcal{M}} = 0 \iff$$
$$dQ_{\mathcal{M}^C} \left(1 + \sum_{j \in \mathcal{M}^C} \frac{b_j}{a_j - b_j} \right) + \sum_{j \in \mathcal{M}^C} \frac{b_j}{a_j - b_j} dQ_{\mathcal{M}} = 0 \iff$$
$$-1 < R_{\mathcal{M}^C} = \frac{dQ_{\mathcal{M}^C}}{dQ_{\mathcal{M}}} = -\sum_{j \in \mathcal{M}^C} \frac{b_j}{a_j - b_j} \Big/ \left(1 + \sum_{j \in \mathcal{M}^C} \frac{b_j}{a_j - b_j} \right) =$$
$$-\sum_{j \in \mathcal{M}^C} \frac{\beta}{\beta + c''(q_j)} \Big/ \left(1 + \sum_{j \in \mathcal{M}^C} \frac{\beta}{\beta + c''(q_j)} \right) < 0.$$

As Dixit (1986) writes, the "equilibrium moves along their reaction functions". Hence, $dQ_{\mathcal{M}^C} = R_{\mathcal{M}^C} dQ_{\mathcal{M}}$ and an increase in the club's production goes along with an increase in the total quantity produced. □

Proof of Proposition 9. First, consider $E^{\text{minE}}_{\mathcal{N}}$. As members' production remains the same and the equilibrium pre founding is unique, non-members' production values and, hence, emissions remain constant. Members' emissions decrease if (1) emission functions are not strictly concave and there exist k and $l \in \mathcal{M}$ such that $e'_k(q^0_k) > e'_l(q^0_l)$: emissions can already be reduced by a small shift of production from k to l (keeping the aggregate club-production at $Q^0_{\mathcal{M}}$).

Emissions also decrease if (2) the e_k are strictly concave: in this case, even if all members have the same marginal emissions before the founding, it is optimal to have only one member produce.[31] Accordingly, one member, say l, produces $Q^0_{\mathcal{M}}$, where l is the[32] member that can do so at the lowest emission level.

Together, (1) and (2) establish the necessity of equal pre-club marginal emissions and no strict concavity for equality stated in the Proposition. Sufficiency follows from the fact that if all pre-club marginal emissions are equal and emission functions not strictly concave, the club cannot decrease emissions without changing its aggregate production as the pre-club quantities are distributed "emission-efficiently". Accordingly, $E^{\text{minE}}_{\mathcal{N}} \leq E^0_{\mathcal{N}}$, and equality holds if and only if all pre-club marginal emissions are equal and emission functions not strictly concave.

[31] Note that, as $c_i(0) = 0$ and cost functions are continuous and convex, all members supply strictly positive quantities in the equilibrium before the founding.

[32] Due to concavity, this holds true even if there are two equally "clean" members.

The arguments for E^{maxQ} are almost the same: if (3) emission functions are not strictly concave and marginal emissions are not all equal for all members, a small shift of production between a "high"- and a "low"-marginal emitter reduces the club's aggregate emissions and allows club members to increase their production to a level higher than $Q^0_{\mathcal{M}}$ without increasing its emissions to a level higher than $E^0_{\mathcal{M}}$. As a reaction, non-members decrease their production (and emissions) leading to an equilibrium with output $Q^1_{\mathcal{N}} > Q^0_{\mathcal{N}}$ (by Lemma 16). Recall: the optimisation problem constrains members to emit exactly as much carbon as they did before the founding. Taken together, non-members' emissions decrease, members' emissions remain constant and $E^{\text{maxQ}} < E^0$.

The same holds, if (4) members' emission functions are strictly concave: Let l be the member that can produce the highest quantity while emitting exactly $E^0_{\mathcal{M}}$. Then, l is the[33] only club-member allowed to produce. Accordingly, due to Lemma 16, non-members emit less than they did before the founding, members' emissions remain unchanged and $E^0_{\mathcal{N}} < E^{\text{maxQ}}_{\mathcal{N}}$.

Yet again, the case in which pre-club marginal emissions are equal and emission functions not strictly concave does not allow for a higher club-production, as pre-club quantities were distributed emission-efficiently and any increase in production by members implies increased club-emissions. Along with (3) and (4), this establishes the statement about necessity and sufficiency of equal marginal and no strict concavity of emissions for $E^0_{\mathcal{N}} = E^{\text{maxQ}}_{\mathcal{N}}$.

To summarise: If emission functions are not strictly concave and not all marginal emissions equal before the founding, emissions decrease for both (minE) and (maxQ). If they are strictly concave, they do as well. If and only if they are not strictly concave and all marginal emissions are equal before the founding, the statement holds with equality. This covers all possible cases.
Lastly, note that all above discussed equilibrium objects result from subgame perfect responses of non-members to the quantities D commits to: all objects consider equilibria in which non-members react according to their best response functions; accordingly, they induce Nash equilibria in the subgame in which non-members set their quantities. □

Definition 9. *For two equilibrium quantity vectors $\boldsymbol{q}^0_{\mathcal{N}}$ and $\boldsymbol{q}^1_{\mathcal{N}}$, define **carbon leakage** L as follows:*

$$L(\boldsymbol{q}^0_{\mathcal{N}}, \boldsymbol{q}^1_{\mathcal{N}}) := \frac{\sum_{j\in\mathcal{M}^C} e_j(q^1_j) - e_j(q^0_j)}{\sum_{k\in\mathcal{M}} e_k(q^0_k) - e_k(q^1_k)}$$

[33] Cf. footnote 32.

Observation 5. *Consider the implementation of* $\tilde{\tau} : \tilde{\tau}_k(q) = \tilde{t}e_k(q),\ \tilde{t} > 0\ \forall k \in \mathcal{M}$ *and let* $\boldsymbol{q}^0_{\mathcal{N}}$ *and* $\boldsymbol{q}^{\tau}_{\mathcal{N}}$ *be the equilibrium quantity vectors before and after the implementation of the tax, respectively; then,*

$$L(\boldsymbol{q}^0_{\mathcal{N}}, \boldsymbol{q}^{\tau}_{\mathcal{N}}) > 0.$$

Proof. The proof follows from that of Lemma 16: One can think of the adjustment process as an iterated introduction of taxes. First, member 1 gets introduced to the tax scheme. Now, its best response to Q_{-1} is smaller than it was before the tax. Accordingly, q_1 falls and Q_{-1} increases.[34] Note that, by the proof of Lemma 16, the total quantity decreases nevertheless as the reaction function of all other firms has a negative slope with absolute value strictly bounded above by 1. Accordingly, the total quantity after the first iteration $\tilde{Q}^1_{\mathcal{N}} < Q^0_{\mathcal{N}}$. Now repeat this process until all members have been introduced to the tax: iterating over all $k \in \mathcal{M}$, we get that $\tilde{Q}^m_{\mathcal{N}} < Q^0_{\mathcal{N}}$. $\tilde{Q}^m_{\mathcal{N}}$ is equal to the equilibrium quantity after the implementation of the tax scheme (even if all taxes are introduced simultaneously). Accordingly, $p^0 < p^{\tilde{\tau}}$. Carbon leakage is positive as all non-members' quantities are decreasing in members' quantities, which must have decreased (otherwise, $\tilde{Q}^m_{\mathcal{N}} < Q^0_{\mathcal{N}}$ would not be possible). □

Proof of Corollary 2. The corollary follows from the fact that the total quantity remains constant upon implementation of (minE)'s solution (hence $p^{\text{minE}} = p^0$). Lemma 16 implies $p^{\text{maxQ}} \leq p^0$. Furthermore, the implementation of a uniform tax scheme reduces members' production for all non-member-quantities. Accordingly, the reaction function derived in the proof of Lemma 16 implies the total quantity to decrease as a response to the tax. Therefore, the price increases.
It is readily apparent that changes in the price level are sufficient statistics for the changes in consumer surplus. □

Proof of Observation 4.
(1) The statement follows from the proof of Proposition 10.
(2) Note that as the emission function is strictly concave, it is always optimal to have only one firm produce (i.e. the cleanest member, given a certain level $Q^0_{\mathcal{M}}$). Suppose member 1 is the unique cleanest given $Q^0_{\mathcal{M}}$ and take any other member k. Then, by subadditivity[35] (which holds for all concave functions f with $f(0) = 0$) for $1 > \alpha > 0$,

$$e_1\big(\alpha Q^0_{\mathcal{M}}\big) + e_k\big((1-\alpha)Q^0_{\mathcal{M}}\big) > e_1\big(\alpha Q^0_{\mathcal{M}}\big) + e_1\big((1-\alpha)Q^0_{\mathcal{M}}\big) > e_1\big(Q^0_{\mathcal{M}}\big).$$

[34] This follows from Lemma 16 in the case in which $\mathcal{M} = \{1\}$.

[35] For $1 > \alpha > 0$ and a strictly concave emission function e, $\alpha e(q) + (1-\alpha)e(0) = \alpha e(q) < e(\alpha q + (1-\alpha)0) = e(\alpha q)$.

Even if there are multiple cleanest members given $Q^0_{\mathcal{M}}$, this statement holds true: suppose, there are two cleanest members (1 and 2), then by subadditivity

$$e_1\big(Q^0_{\mathcal{M}}/2\big) + e_2\big(Q^0_{\mathcal{M}}/2\big) > e_1\big(Q^0_{\mathcal{M}}\big)/2 + e_2\big(Q^0_{\mathcal{M}}\big)/2 = e_1\big(Q^0_{\mathcal{M}}\big).$$

Analogously, for (maxQ), it is optimal to have only the (one) cleanest member produce as it is able to produce the most given a certain emission level.
(3) is trivial: the cleanest member is able to produce the most given a certain level of emissions and emit the least, given a certain level of production. Accordingly, it is the only member allowed to produce. If there are multiple cleanest members, any subset of them may produce—provided that for (minE) (for (maxQ)), the total club-quantity is equal to the pre-club quantity (the club's emission level equal to that before the founding). □

Lemma 17. *Suppose (Conv). e_k is invertible on $\mathbb{R}_+$ for all $k \in \mathcal{M}$.*

Proof. e_k is injective on $\mathbb{R}_+$ as it is strictly increasing. It is also surjective on $\mathbb{R}_+$ as it is unbounded. To see why a continuous and convex increasing function cannot be bounded from above, consider the following derivations. If e_k is strictly increasing, there must exist x_1 and x_2 such that $e_k(x_1) < e_k(x_2)$. Note that by convexity $\frac{e_k(x_2)-e_k(x_1)}{x_2-x_1}$ is strictly increasing in x_2. Now take some $x_3 > x_2$:

$$\frac{e_k(x_3)-e_k(x_1)}{x_3-x_1} \geq \frac{e_k(x_2)-e_k(x_1)}{x_2-x_1} \iff e_k(x_3) \geq e_k(x_2) + (x_3-x_1)\frac{e_k(x_2)-e_k(x_1)}{x_2-x_1}$$

But, as the RHS diverges to infinity as $x_3 \to \infty$, the LHS must do so as well. Hence, e_k is invertible on $\mathbb{R}_+$. □

Lemma 18. *Suppose (Conv). $\hat{e}_{\mathcal{M}}(q) := \sum_{k\in\mathcal{M}} e_k(q)$ is invertible on $\mathbb{R}_+$.*

Proof.
(1) First, note that $\hat{e}_{\mathcal{M}}$ is injective on $\mathbb{R}_+$ as all e_k are positive-valued and strictly increasing; hence, their sum $\hat{e}_{\mathcal{M}}$ is as well.
(2) Furthermore, $\hat{e}_{\mathcal{M}}$ is continuous on $\mathbb{R}_+$ as the sum of a finite number of continuous functions is continuous:

$$\lim_{x\to a} e_k(x) = e_k(a) \qquad \lim_{x\to a} e_j(x) = e_j(a) \Rightarrow \lim_{x\to a} e_k(x) + e_j(x) = e_k(a) + e_j(a).$$

(3) By the intermediate value theorem, $\hat{e}_{\mathcal{M}}$ is surjective on $\mathbb{R}_+$, as: all $e_k(0) = 0$, $\hat{e}_{\mathcal{M}}(0) = 0$. Furthermore, $\lim_{x\to\infty} e_k(x) = \infty$; hence, $\lim_{x\to\infty} \hat{e}_{\mathcal{M}}(x) = \infty$. Therefore, $\hat{e}_{\mathcal{M}}$ is bijective on $\mathbb{R}_+$. Accordingly, it is invertible on $\mathbb{R}_+$. □

Lemma 19. *Suppose (Conv).* $g_k(q) := c_k'(q) + 2\beta q$ *is invertible on* $\mathbb{R}_+$ *for all* $k \in \mathcal{M}$.

Proof.
(1) First, note that $g_k(q)$ is injective on $\mathbb{R}_+$ as both c_k' and $2\beta q$ are positive-valued and strictly increasing; hence, their sum is as well.
(2) Furthermore, $g_k(q)$ is continuous on $\mathbb{R}_+$ as the sum of a finite number of continuous functions is continuous (cf. proof of Lemma 18).
(3) By the intermediate value theorem, $g_k(q)$ is surjective on $\mathbb{R}_+$, as: $g_k(0) = 0$ and $\lim_{q\to\infty} g_k(q) = \infty$. Hence, $g_k(q)$ is bijective on $\mathbb{R}_+$. Accordingly, it is invertible on $\mathbb{R}_+$. □

Lemma 20. *Suppose (Conv).* $Q_{\mathcal{M}}^{e'}(x) := \sum_{k\in\mathcal{M}} e'^{-1}_k(x)$ *is invertible on* $\mathbb{R}_+$.

Proof. As all e'^{-1}_k are invertible and positive valued on $\mathbb{R}_+$, their sum $Q_{\mathcal{M}}^{e'}$ is injective on $\mathbb{R}_+$. It is continuous on $\mathbb{R}_+$ by the same reasoning as applied in step (2) in the proof of Lemma 18. Step (3) in the proof of Lemma 18 can be applied to $Q_{\mathcal{M}}^{e'}$ as well. Accordingly, $Q_{\mathcal{M}}^{e'}$ is invertible on $\mathbb{R}_+$. □

Lemma 21. *Suppose (Conv).* $\sum_{k\in\mathcal{M}} e_k\big(e'^{-1}_k(x)\big)$ *is invertible on* $\mathbb{R}_+$.

Proof. All $e_k\big(e'^{-1}_k(x)\big)$ are invertible on $\mathbb{R}_+$, as by Lemma 17 all e_k and e'^{-1}_k are. Accordingly, the sum $\sum_{k\in\mathcal{M}} e_k\big(e'^{-1}_k(x)\big)$ is injective on $\mathbb{R}_+$, as all functions are increasing and positive valued on $\mathbb{R}_+$. Furthermore, the function is continuous on $\mathbb{R}_+$ as a composite of continuous functions is continuous and the sum of continuous functions is as well (step (2) in the proof of Lemma 18). The third step in the proof of Lemma 18 can also be applied to $\sum_{k\in\mathcal{M}} e_k\big(e'^{-1}_k(x)\big)$; accordingly, the function is invertible on $\mathbb{R}_+$. □

Proof of Proposition 10. Note that Assumption (Conv) implies the below derived candidate solutions to be optimal as the Karush-Kuhn-Tucker conditions are fulfilled, the feasible sets are convex and the objective function convex (minE) and concave (maxQ), respectively (cf. Hanson, 1981).

Consider the Lagrangian and first order conditions implied by a slightly adjusted version of problem (minE). Let λ and $\boldsymbol{\mu}$ be a non-negative valued scalar and row vector of length m, respectively; then:

$$\min_{\boldsymbol{q}_{\mathcal{M}}} \sum_{k\in\mathcal{M}} e_k\big(q_k\big) + \lambda\big(Q_{\mathcal{M}}^0 - \sum_{k\in\mathcal{M}} q_k\big) - \boldsymbol{\mu}\cdot\boldsymbol{q}_{\mathcal{M}} \text{ s.t. } \lambda, \mu_k \geq 0 \; \forall k \in \mathcal{M}$$

Then, for solution $\boldsymbol{q}_{\mathcal{M}}^{\text{minE}}$:

$$e_k'(q_k^{\text{minE}}) = \lambda + \mu_k \iff q_k^{\text{minE}} = e'^{-1}_k(\lambda) \; \forall k \in \mathcal{M} \text{ s.t. } q_k^{\text{minE}} > 0$$
$$Q_{\mathcal{M}}^0 = \sum_{k\in\mathcal{M}} q_k^{\text{minE}} = \sum_{k\in\mathcal{M}} e'^{-1}_k(\lambda)$$

Note that $\boldsymbol{\mu} = \mathbf{0}$ as by Assumption (Conv), $e_k'(q) = 0 \iff q = 0$, which would require $\lambda = 0$ and, hence, $q_k^{\text{minE}} = 0$ for all k. This is only possible for $Q_{\mathcal{M}}^0 = 0$, $Q_{\mathcal{M}^C}^0 = \alpha/\beta$ and, hence, $p^0 = 0$. This case does not need to be considered as there are finitely many non-members who would not produce if the price level were 0; hence, $p^0 > 0$.

Let $Q_{\mathcal{M}}^{e'}(x) := \sum_{k\in\mathcal{M}} e'^{-1}_k(x)$, where $x \in \mathbb{R}_+$. Note that the function is both continuous[36] and invertible (cf. Lemma 20) on $\mathbb{R}_+$. Accordingly,

$$\lambda = \left[Q_{\mathcal{M}}^{e'}\right]^{-1}(Q_{\mathcal{M}}^0) \Rightarrow q_k^{\text{minE}} = e'^{-1}_k\left(\left[Q_{\mathcal{M}}^{e'}\right]^{-1}(Q_{\mathcal{M}}^0)\right).$$

Then, the tax scheme consisting of tax functions defined as

$$\tau_k^{\text{minE}}(q_k, Q_{-k}) := q_k\left(\alpha - \beta Q_{-k} - c_k'(q_k^{\text{minE}}) - 2\beta q_k^{\text{minE}}\right) \ \forall k \in \mathcal{M}$$

implements q_k^{minE} for all k: The maximisation problem and first order condition of firm k, given some level Q_{-k} take the following form:

$$\begin{aligned}
&\pi_k(q_k, Q_{-k}) - \tau_k(q_k, Q_{-k}) = \\
&\left(\alpha - \beta Q_{-k} - \left(\alpha - \beta Q_{-k} - c_k'(q_k^{\text{minE}}) - 2\beta q_k^{\text{minE}}\right)\right)q_k - c_k(q_k) - \beta q_k^2 = \\
&\left(c_k'(q_k^{\text{minE}}) + 2\beta q_k^{\text{minE}}\right)q_k - c_k(q_k) - \beta q_k^2 \\
&\frac{\partial\left(\pi_k(q_k, Q_{-k}) - \tau_k(q_k, Q_{-k})\right)}{\partial q_k} = c_k'(q_k^{\text{minE}}) + 2\beta q_k^{\text{minE}} - c_k'(q_k) - 2\beta q_k \overset{!}{=} 0
\end{aligned}$$

Note that the function $g_k(q) = c_k'(q) + 2\beta q$ is bijective by Lemma 19 and therefore:

$$\begin{aligned}
q_k = q_k^{\text{minE}} \Rightarrow \sum_{k\in\mathcal{M}} q_k^{\text{minE}} &= \sum_{k\in\mathcal{M}} e'^{-1}_k\left(\left[Q_{\mathcal{M}}^{e'}\right]^{-1}(Q_{\mathcal{M}}^0)\right) \\
&= Q_{\mathcal{M}}^{e'}\left(\left[Q_{\mathcal{M}}^{e'}\right]^{-1}(Q_{\mathcal{M}}^0)\right) = Q_{\mathcal{M}}^0,
\end{aligned}$$

as was to be shown.

Now, consider the Lagrangian and first order conditions implied by (maxQ). Again, let $\hat{\lambda}$ and $\hat{\boldsymbol{\mu}}$ be a scalar and row vector of length m, respectively and define (if admissible) $\tilde{\lambda} = 1/\hat{\lambda}$; then:

$$\max_{\boldsymbol{q}_{\mathcal{M}}} \sum_{k\in\mathcal{M}} q_k + \hat{\lambda}\left(E_{\mathcal{M}}^0 - \sum_{k\in\mathcal{M}} e_k(q_k)\right) + \hat{\boldsymbol{\mu}}\cdot\boldsymbol{q}_{\mathcal{M}} \ \text{ s.t. } \hat{\lambda}, \hat{\mu}_k \geq 0 \ \forall k \in \mathcal{M}$$

[36] The sum of a finite number of continuous functions is continuous (see proof of Lemma 18). Hence, $Q_{\mathcal{M}}^{e'}(x)$ is continuous and so is its inverse.

Then, for solution $\boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}}$:

$$
\begin{aligned}
&e_k'(q_k^{\text{maxQ}})\hat{\lambda} = 1 + \hat{\mu}_k \iff e'^{-1}_k(\tilde{\lambda}) = q_k^{\text{maxQ}} \ \forall k \in \mathcal{M} \text{ s.t. } q_k^{\text{maxQ}} > 0 \\
&E^0_{\mathcal{M}} = \sum_{k\in\mathcal{M}} e_k\big(q_k^{\text{maxQ}}\big) = \sum_{k\in\mathcal{M}} e_k\big(e'^{-1}_k(\tilde{\lambda})\big) \\
&\hat{\boldsymbol{\mu}} \cdot \boldsymbol{q}_{\mathcal{M}}^{\text{maxQ}} = 0
\end{aligned}
$$

Note that $q_k^{\text{maxQ}} > 0$ for all k (hence $\hat{\boldsymbol{\mu}} = \mathbf{0}$ due to similar reasoning as outlined for (minE)). Furthermore, $\hat{\lambda} = 0$ is not possible due to Assumption (Conv).

Let $\big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big](x) := \sum_{k\in\mathcal{M}} e_k\big(e'^{-1}_k(x)\big)$, where $x \in \mathbb{R}^+$. Note that $\big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]$ is both continuous[37] and invertible (cf. Lemma 21) on $\mathbb{R}_+$. Accordingly, for q_k^{maxQ}:

$$
\tilde{\lambda} = \big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]^{-1}\big(E^0_{\mathcal{M}}\big) \Rightarrow e'^{-1}_k\Big(\big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]^{-1}\big(E^0_{\mathcal{M}}\big)\Big) = q_k^{\text{maxQ}}.
$$

Then, the tax scheme consisting of tax functions defined as

$$
\tau_k^{\text{maxQ}}(q_k, Q_{-k}) := q_k\big(\alpha - \beta Q_{-k} - c_k'(q_k^{\text{maxQ}}) - 2\beta q_k^{\text{maxQ}}\big) \ \forall k \in \mathcal{M}
$$

implements q_k^{maxQ} for all k: Given some level Q_{-k}, the first order condition of firm k takes the following form:

$$
\begin{aligned}
&\alpha - \beta Q_{-k} - \big(\alpha - \beta Q_{-k} - c_k'(q_k^{\text{maxQ}}) - 2\beta q_k^{\text{maxQ}}\big) = c_k'(q_k) + 2\beta q_k \iff \\
&c_k'(q_k^{\text{maxQ}}) + 2\beta q_k^{\text{maxQ}} = c_k'(q_k) + 2\beta q_k
\end{aligned}
$$

Note that the function $g_k(q) = c_k'(q) + 2\beta q$ is bijective by Lemma 19 and therefore:

$$
\begin{aligned}
q_k = q_k^{\text{maxQ}} \Rightarrow \sum_{k\in\mathcal{M}} q_k^{\text{maxQ}}(Q_{\mathcal{M}^C}) &= \sum_{k\in\mathcal{M}} e_k\Big(e'^{-1}_k\Big(\big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]^{-1}\big(E^0_{\mathcal{M}}\big)\Big)\Big) \\
&= \big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]\Big(\big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]^{-1}\big(E^0_{\mathcal{M}}\big)\Big) = E^0_{\mathcal{M}},
\end{aligned}
$$

as was to be shown.

Both types of tax functions are continuous on the relevant domain as they are composites of continuous functions. □

Proof of Theorem 5. First, consider scheme $\boldsymbol{\tau}^{\text{minE}}$. Note that, as all emission functions are equal, $\bar{q}_{\mathcal{M}} := Q^0_{\mathcal{M}}/m = q_k^{\text{minE}}$ for all $k \in \mathcal{M}$. Furthermore, by the firms'

[37] The function is continuous as both e_k and $e'^{-1}_k(x)$ are continuous on $\mathbb{R}_+$ (Assumption (Conv)) and the sum of a finite number of continuous functions is itself continuous (see proof of Lemma 18). Being the inverse of a continuous function, $\big[E_{\mathcal{M}} \circ Q^{e'}_{\mathcal{M}}\big]^{-1}$ is continuous as well.

first order conditions:

$$\alpha - \beta Q^0_{\mathcal{N}} = p^0 = q^0_k(\gamma_k + \beta) \;\; \forall k \in \mathcal{M}.$$

Accordingly, the sum of taxes satisfies the following[38]

$$\sum_{k\in\mathcal{M}} \tau_k^{\text{minE}}(q_k^{\text{minE}}, Q^0_{\mathcal{N}} - q_k^{\text{minE}}) = p^0 m\bar{q}_{\mathcal{M}} - \sum_{k\in\mathcal{M}} \bar{q}^2_{\mathcal{M}}(\gamma_k + \beta) \propto$$
$$p^0 m - \bar{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} (\gamma_k + \beta) = p^0 m - \bar{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} p^0/q^0_k \propto$$
$$m - \bar{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} 1/q^0_k < 0 \iff 1/\bar{q}_{\mathcal{M}} < 1/m \sum_{k\in\mathcal{M}} 1/q^0_k,$$

which is true by Jensen's inequality. Now, consider $\boldsymbol{\tau}^{\text{maxQ}}$. In this case, $\hat{q}_{\mathcal{M}} := \sqrt{\sum_{k\in\mathcal{M}}(q^0_k)^2/m} = q_k^{\text{maxQ}}$ for all $k \in \mathcal{M}$: as members' emission functions are convex and all equal, q_k^{maxQ} must be equal for all k. Accordingly,

$$\eta\, m(q^{\text{maxQ}})^2 = \eta \sum_{k\in\mathcal{M}} (q^0_k)^2 \iff q^{\text{maxQ}} = \hat{q}_{\mathcal{M}}.$$

By derivations similar to those for $\boldsymbol{\tau}^{\text{minE}}$, the sum of taxes paid equals the following expression:

$$p^{\text{maxQ}} m\hat{q}_{\mathcal{M}} - \hat{q}^2_{\mathcal{M}} \sum_{k\in\mathcal{M}} (\gamma_k + \beta) < p^0\, m\hat{q}_{\mathcal{M}} - \hat{q}^2_{\mathcal{M}} \sum_{k\in\mathcal{M}} (\gamma_k + \beta) =$$
$$p^0\, m\hat{q}_{\mathcal{M}} - \hat{q}^2_{\mathcal{M}} \sum_{k\in\mathcal{M}} p^0/q^0_k$$

I proceed by comparing the above to the expression for the balance of $\boldsymbol{\tau}^{\text{minE}}$: Note that, by the Cauchy-Schwarz inequality, $\bar{q}_{\mathcal{M}} < \hat{q}_{\mathcal{M}}$ (strict inequality due to Assumption (Quad)).

$$p^0 m\bar{q}_{\mathcal{M}} - \bar{q}^2_{\mathcal{M}} \sum_{k\in\mathcal{M}} p^0/q^0_k > p^0\, m\hat{q}_{\mathcal{M}} - \hat{q}^2_{\mathcal{M}} \sum_{k\in\mathcal{M}} p^0/q^0_k \iff$$
$$\bar{q}_{\mathcal{M}}\Big(m - \bar{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} 1/q^0_k\Big) > \hat{q}_{\mathcal{M}}\Big(m - \hat{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} 1/q^0_k\Big),$$

which is true as the terms within the brackets are negative-valued and $\bar{q}_{\mathcal{M}} < \hat{q}_{\mathcal{M}}$. □

Proof of Theorem 6. The first inequality is trivial: in the equilibrium without regulation, firms produce if and only if their profit is greater than zero. As firms have

[38] Note that $p^0 > 0$ as such price would not constitute an equilibrium. The pre-club equilibrium exists by Szidarovszky and Yakowitz (1982).

quadratic costs, all firms produce a positive quantity in the unique pre-founding equilibrium (cf. Szidarovszky and Yakowitz, 1982). Accordingly, the sum of profits is greater than zero. The second inequality follows from the following derivations, where the characterisation of q_k^{minE} follows from the proof of Theorem 5.

$$\sum_{k\in\mathcal{M}} \pi_k(q_k^{\text{minE}}, Q_{-k}^{\text{minE}}) - \sum_{k\in\mathcal{M}} \tau_k^{\text{minE}}(q_k^{\text{minE}}, Q_{-k}^{\text{minE}}) = p^0 m\bar{q}_{\mathcal{M}} - \sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}^2 0.5\gamma_k -$$
$$p^0 m\bar{q}_{\mathcal{M}} + \sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}^2(\gamma_k+\beta) > \sum_{k\in\mathcal{M}} \pi_k(q_k^0, Q_{-k}^0) = p^0 m\bar{q}_{\mathcal{M}} - \sum_{k\in\mathcal{M}} (q_k^0)^2 0.5\gamma_k =$$
$$\bar{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} q_k^0(\gamma_k+\beta) - \sum_{k\in\mathcal{M}} (q_k^0)^2 0.5\gamma_k \iff$$
$$\sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}^2(\gamma_k+\beta) - \bar{q}_{\mathcal{M}} \sum_{k\in\mathcal{M}} q_k^0(\gamma_k+\beta) > \sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}^2 0.5\gamma_k - \sum_{k\in\mathcal{M}} (q_k^0)^2 0.5\gamma_k \iff$$
$$\sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}(\bar{q}_{\mathcal{M}} - q_k^0)(\gamma_k+\beta) > \sum_{k\in\mathcal{M}} (\bar{q}_{\mathcal{M}}^2 - (q_k^0)^2) 0.5\gamma_k =$$
$$\sum_{k\in\mathcal{M}} (\bar{q}_{\mathcal{M}} - q_k^0)(\bar{q}_{\mathcal{M}} + q_k^0) 0.5\gamma_k \iff$$
$$\sum_{k\in\mathcal{M}} (\bar{q}_{\mathcal{M}} - q_k^0)\big(\bar{q}_{\mathcal{M}}(\gamma_k+\beta) - (\bar{q}_{\mathcal{M}} + q_k^0) 0.5\gamma_k\big) =$$
$$\sum_{k\in\mathcal{M}} (\bar{q}_{\mathcal{M}} - q_k^0)\big(\bar{q}_{\mathcal{M}}(0.5\gamma_k+\beta) - q_k^0 0.5\gamma_k\big) =$$
$$\sum_{k\in\mathcal{M}} (\bar{q}_{\mathcal{M}} - q_k^0)^2 0.5\gamma_k + \sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}(\bar{q}_{\mathcal{M}} - q_k^0)\beta > 0,$$

which is true by assumption. Now consider the third inequality:

$$\sum_{k\in\mathcal{M}} \pi_k(q_k^{\text{maxQ}}, Q_{-k}^{\text{maxQ}}) - \sum_{k\in\mathcal{M}} \tau_k^{\text{maxQ}}(q_k^{\text{maxQ}}, Q_{-k}^{\text{maxQ}}) =$$
$$p^{\text{maxQ}} m\hat{q}_{\mathcal{M}} - \sum_{k\in\mathcal{M}} \hat{q}_{\mathcal{M}}^2 0.5\gamma_k - p^{\text{maxQ}} m\hat{q}_{\mathcal{M}} +$$
$$\sum_{k\in\mathcal{M}} \hat{q}_{\mathcal{M}}^2(\gamma_k+\beta) = \sum_{k\in\mathcal{M}} \hat{q}_{\mathcal{M}}^2(0.5\gamma_k+\beta) >$$
$$\sum_{k\in\mathcal{M}} \pi_k(q_k^{\text{minE}}, Q_{-k}^{\text{minE}}) - \sum_{k\in\mathcal{M}} \tau_k^{\text{minE}}(q_k^{\text{minE}}, Q_{-k}^{\text{minE}}) = \sum_{k\in\mathcal{M}} \bar{q}_{\mathcal{M}}^2(0.5\gamma_k+\beta),$$

which is true as $\bar{q}_{\mathcal{M}} < \hat{q}_{\mathcal{M}}$ by the Cauchy-Schwarz inequality. □

Let V_{+l}^{minE}, V_{+l}^{maxQ} and V_{+l}^{*} denote the values of the objective function upon implementation of the quantities implied by (minE), (maxQ) and $(*)$ in a market in which non-member $l \in \mathcal{M}^C$ is a club member as well, then:

Observation 6. *Compare the values of the objective function upon implementation of* $\boldsymbol{q}_{\mathcal{M}}^{minE}$, $\boldsymbol{q}_{\mathcal{M}}^{maxQ}$ *and* $\boldsymbol{q}_{\mathcal{M}}^{*}$ *in the case in which* l *is a member to the case in which it is not,*

respectively. Then,

$$V_{+l}^{minE} \leq V^{minE}, \qquad V_{+l}^{maxQ} \leq V^{maxQ} \quad \textit{and} \quad V_{+l}^{*} \leq V^{*}.$$

Proof. The proof is trivial: For all optimisation problems (minE), (maxQ) and $(*)$, it is admissible to not change firm l's supply (it produces the same quantity as it did before joining the club). Accordingly, the club is able to achieve the same outcome in the case in which l is a member as the case in which it is not. The emissions may even be lower in the former if the optimal choices are different from l's individually optimal quantity before joining the club, respectively. As the club has more control when l is a member as well, the value of the objective function is always weakly smaller compared to the case in which it is not a member. □

References

Babiker, Mustafa H. 2005. "Climate change policy, market structure, and carbon leakage." *Journal of International Economics*, 65(2): 421–445.

Baccianti, Claudio, and Oliver Schenker. 2021. "Cournot, Pigou, and Ricardo Walk in a Bar – Unilateral Environmental Policy and Leakage With Market Power and Firm Heterogeneity." *Working Paper*.

Barrett, Scott. 1994. "Self-enforcing International Environmental Agreements." *Oxford Economic Papers*, 46: 878–894.

Bosetti, Valentina, Carlo Carraro, Enrica De Cian, Emanuele Massetti, and Massimo Tavoni. 2013. "Incentives and stability of international climate coalitions: An integrated assessment." *Energy Policy*, 55: 44–56.

David, Maia, and Bernard Sinclair-Desgagné. 2005. "Environmental regulation and the eco-industry." *Journal of Regulatory Economics*, 28(2): 141–155.

Dietz, Simon, and Frank Venmans. 2019. "Cumulative carbon emissions and economic policy: In search of general principles." *Journal of Environmental Economics and Management*, 96: 108–129.

Dixit, Avinash. 1986. "Comparative Statics for Oligopoly." *International Economic Review*, 27(1): 107–122.

Ebert, Udo, and Heinz Welsch. 2007. "Environmental emissions and production economics: Implications of the materials balance." *American Journal of Agricultural Economics*, 89(2): 287–293.

Finus, Michael, Juan-Carlos Altamirano-Cabrera, and Ekko C. Van Ierlan. 2005. "The Effect of Membership Rules and Voting Schemes on the Success of International Climate Agreements." *Public Choice*, 125(1/2): 99–127.

Fowlie, Meredith L. 2009. "Incomplete Environmental Regulation, Imperfect Competition, and Emissions Leakage." *American Economic Journal: Economic Policy*, 1(2): 72–112.

Ganapati, Sharat, Joseph S. Shapiro, and Reed Walker. 2020. "Energy cost pass-through in US manufacturing: Estimates and implications for carbon taxes." *American Economic Journal: Applied Economics*, 12(2): 303–342.

Hanson, Morgan A. 1981. "On sufficiency of the Kuhn-Tucker conditions." *Journal of Mathematical Analysis and Applications*, 80(2): 545–550.

Hoel, Michael. 1991. "Global environmental problems: The effects of unilateral actions taken by one country." *Journal of Environmental Economics and Management*, 20(1): 55–70.

Holland, Stephen, Matthew Kotchen, Erin Mansur, and Andrew Yates. 2022. "Why marginal CO2 emissions are not decreasing for US electricity: Estimates and implications for climate policy." *PNAS*, 119(8): 1–11.

Hovi, Jon, Detlef Sprinz, Håkon Sælen, and Arild Underdal. 2019. "The Club Approach: A Gateway to Effective Climate Co-operation?" *British Journal of Political Science*, 49(3): 1071–1096.

International Energy Association. 2020. "Iron and Steel Technology Roadmap." *https://www.iea.org/reports/iron-and-steel-technology-roadmap*.

Koenig, Evan F. 1985. "Indirect methods for regulating externalities under uncertainty." *Quarterly Journal of Economics*, 100(2): 479–493.

Lazkano, Itziar, Walid Marrouch, and Bruno Nkuiya. 2016. "Adaptation to climate change: How does heterogeneity in adaptation costs affect climate coalitions?" *Environment and Development Economics*, 21(6): 812–838.

Lee, Sang-ho, and Chul-hi Park. 2010. "Two-Part Tax for Polluting Oligopolists with Endogenous Entry." *Environmental and Resource Economics Review*, 19(3).

Mason, Charles F., Stephen Polasky, and Nori Tarui. 2017. "Cooperation on climate-change mitigation." *European Economic Review*, 99: 43–55.

McKitrick, Ross. 1999. "A Cournot mechanism for pollution control under asymmetric information." *Environmental and Resource Economics*, 14(3): 353–363.

Neuhoff, Karsten, and Robert A Ritz. 2019. "Carbon cost pass-through in industrial sectors." *Working Paper*.

Nordhaus, By William. 2015. "Climate Clubs: Overcoming Free-riding in International Climate Policy." *American Economic Review*, 105(4): 1339–1370.

Nordhaus, William D. 2018. "Prize Lecture: Climate change: The Ultimate Challenge for Economics."

Perry, Martin K. 1982. "Oligopoly and Consistent Conjectural Variations." *The Bell Journal of Economics*, 13(1): 197–205.

Ritz, Robert A. 2009. "Carbon leakage under incomplete environmental regulation: An industry-level approach." *Working Paper*.

Sælen, Håkon. 2016. "Side-payments: an effective instrument for building climate clubs?" *International Environmental Agreements: Politics, Law and Economics*, 16(6): 909–932.

Szidarovszky, F., and S. Yakowitz. 1982. "Contributions to Cournot oligopoly theory." *Journal of Economic Theory*, 28(1): 51–70.

Yomogida, Morihiro, and Nori Tarui. 2013. "Emission taxes and border tax adjustments for oligopolistic industries." *Pacific Economic Review*, 18(5): 644–673.