Unity in Diversity: How Norms of Compromise Enable Cooperation

Saumya Deojain *

W. Allen Wallis Institute of Political Economy, University of Rochester

January 2022

Abstract

This paper introduces the notion of a 'norm of compromise' and demonstrates its importance for cooperation among agents with diverse preferences over a public policy. Agents choose to cooperate when they join a coalition and agree to support a commonly proposed policy that could be far away from their preferred policy. A norm of compromise is an exogenous protocol used by a coalition to arrive at this commonly proposed or 'compromise' policy. I consider a parameterized class of norms in which the compromise policy's relative sensitivity to moderates and extremists in the coalition can be dialed. I study the effect of these norms on the stability of the grand coalition (or full cooperation) in a model where an agent faces a trade-off between compromise if she joins a coalition and increased risk if she does not. I find that polarization does not always reduce cooperation: it destabilizes the grand coalition under norms with low relative sensitivity to moderates but stabilizes it under norms with high relative sensitivity to moderates. This can lead to a situation where norms enabling cooperation in a polarized society do not enable cooperation in a homogeneous one. I also find the counterintuitive result that under some norms extremists are less willing to cooperate when moderates' preferences get closer to these extremists. This work sheds light on the emergence of cooperation within social movements like Black Lives Matter and the Arab Spring, where political actors compromise in the relative absence of formal institutional structures.

JEL: H41, D71, D72

Keywords: stability, coalitions, public policy, norms, diversity

*I would like to thank Marcus Berliant, Randall Calvert, Valerio Dotti, David Lindequist, John Nachbar and Brian Rogers for their invaluable inputs. I would like to acknowledge all the student seminars in the Economics Department and Political Science Department in Washington University in St. Louis which were integral to the evolution of this work and the (Virtual) 2020 Missouri Valley Economic Association Conference where I received important inputs.

1 Introduction

Any collective effort requires a compromise between a diverse set of actors. This is true for people who come together to write a constitution, form cohesive coalitions of protesters/lobbyists, or form international alliances. These compromises are often arrived at using pre-existing norms of compromise, such as majority voting, contest games, or bargaining. Sometimes these norms are successful in enabling cooperation, and sometimes they are not. Sometimes, the same set of people that can cooperate under one norm fragment in another. In this work, I analyze the conditions where a norm of compromise enables cooperation between a diverse set of agents and where it does not.

This work extends the literature concerning the effect of diversity and institutions on cooperation (Alesina et al., 2001; Easterly and Levine, 1997; Gorodnienko and Roland, 2015; Stichnoth and Van der Straeten, 2013) and, specifically, on the prevention of fragmentation in diverse communities by institutions (Alesina and Reich, 2015, Reynal-Querol, 2002, Spolaore and Wacziarg, 2017). Unlike most of the existing literature which has been concerned with the distribution of public goods and conflict, this paper is focused on the trade-offs between the cost of compromise and fragmentation for individuals within a society. I show how an individual's willingness to compromise and cooperate depends upon the distribution of preferences and norms of compromise within a society. This willingness depends on how sensitive a norm of compromise is to the preferences of members of a coalition. The main contribution of this paper is identifying this channel of sensitivity to individual preferences that explicitly links the emergence of cooperation to the underlying norms of compromise and the nature of diversity within a society.

In the model, policies are defined over the real line and each agent has a 'most preferred' policy. The closer a policy is implemented to an agent's preferred policy, the better off she is. All agents know each other's preferences. A group of agents whose preferences need not be the same may form a coalition and agree on a compromise policy that will be proposed by the coalition. A norm of compromise determines which policy is chosen by the coalition as the compromise policy, taking into account the preferences of each agent in the coalition[1]. Each norm is characterized by a real number between zero and unity. When this parameter is close to unity, the compromise policy within the coalition is more sensitive to moderate positions than it is to extreme positions. When the parameter is close to zero, the compromise policy is more sensitive to the extreme positions than it is to moderate positions. Therefore, the compromise policy of the coalition non-trivially depends on the interplay between the norm of compromise and the distribution of preferences of the agents in the coalition.

I apply these norms of compromise in a simple game in which an agent can either join a coalition or back her individual preferred policy. If the agent chooses to join the coalition, she must support its compromise policy (which depends on the norms of compromise as

[1] This means the compromise policy is endogenous to the membership of the coalition even though the norm of compromise, as will be defined, is exogenously specified.

discussed in the previous paragraph). The policy that is finally implemented has a probability distribution over the policies that are backed. The greater the number of people backing a policy, the higher the probability that that policy is implemented. An agent's utility depends on the distance between her preferred policy and the finally implemented policy. The risk aversion of an agent characterizes how much she prefers an uncertain implementation of policy to a certain one. Every agent chooses to join a coalition or back her individually preferred policy based on the assessment of which decision increases her expected utility. (Full) Cooperation emerges in society if all agents join a grand coalition and agree to compromise with each other.

Joining a coalition generates a cost of compromise for an agent since the compromise policy of the coalition need not coincide with the agent's own preferred policy. At the same time, breaking off the coalition creates a cost of fragmentation. When an agent leaves the coalition the compromise policy of the new fragmented coalition is determined by the same norms of compromise without consideration of her preferences. This creates the risk that a policy worse for her is implemented with a positive probability. This results in a trade-off between the cost of compromise and the cost of fragmentation for an agent. If the compromise policy of a coalition is not sensitive to an agent, then the agent faces no risk of leaving the coalition. Cooperation by joining a coalition only reduces the risk for an agent when the compromise policy is sensitive to her preferences. Norms of compromise used by the coalition determine this sensitivity to her preference, therefore also determining her willingness to cooperate.

One of the main results concerns the emergence of cooperation when policy preferences become polarized. Polarization occurs when agents' preferred policies start concentrating on the preferred policy of the extremists in a society. A polarized society is characterized by a bimodal distribution of preferences at the extreme political position. I find when norms of compromise are relatively more sensitive to extremists than moderates, polarization results in a smaller willingness to cooperate. The compromise policies determined by the set of norms with more sensitivity to the extremists are the same as the equilibrium policies chosen in the contest game described in Duggan and Gao (2019). Therefore, this relationship between cooperation and polarization is in line with Esteban and Ray (1999) who find increased polarization results in a greater conflict when agents are in a contest game. When norms of compromise are relatively more sensitive to moderates, polarization results in a greater willingness to cooperate. The more sensitive norms of compromise are to moderates, the closer the compromise policy gets to the outcome of majority voting: the median preferred policy. This relationship between cooperation, polarization, and sensitivity to moderates is mirrored in the result in Reynal-Querol (2002) who finds that conflict is reduced in a polarized society in the presence of democracy.

By applying these results about polarization to a simple example, I find that in a polarized society cooperation does not emerge under norms that are too sensitive to extremes because moderates leave. Conversely, cooperation does not emerge in a more homogeneous society under norms that are too sensitive to moderates because extremes leave. For a given level of polarization, there is a window of norms that depends on the level of polarization

under which cooperation will emerge. Therefore, societies with the same diversity of preferences may have vastly different capabilities to form large coalitions if their norms of compromise are different. This could explain why in the Arab Spring different countries had different levels of cooperation between protesters. Many of these countries had very polarized protesters, but some of these countries were more reliant than others on tribal ties to galvanize the opposition against the status quo. One can argue that when tribal groups cooperate the norm of compromise is to play a contest game. Compromise policies of contest games have high relative sensitivity to extremist positions. As tribes are more differentiated in a society, the compromise is more sensitive to extreme positions. This could explain why Libya had more fractured protests compared to Tunisia even though both were highly polarized. Libya used its more entrenched tribal ties to collectivize in the wake of the Arab Spring - a norm of compromise that could not enable cooperation between its protesters (Anderson, 2011, Tufecki, 2017).

The last main result in the model looks at how the willingness of an extreme to cooperate changes as a moderate comes closer to her preferences. This willingness depends on the norms of compromise too and leads to some counter-intuitive results. When norms of compromise are sensitive to moderate positions and the moderate moves towards an extremist the cost of fragmentation for that extremist increases. Therefore, an extremist is more willing to cooperate within the grand coalition when the moderate's preferred policy gets closer to the extremist's preferred policy. This result is what one would expect. I get counter-intuitive results when compromise policies are sensitive to the extremes. In this case, the cost of fragmentation for the extremist decreases as the moderate's preferred policy gets closer to the extremist's preferred policy. So, the extremist is less willing to cooperate when a moderate's preferences move closer to hers. In other words, this extremist could move out of the coalition and back her own policy as other agents' preferred policy move closer to hers. This could explain an interesting phenomenon in the Black Lives Matter movement. After many moderates within the left changed their preferences towards greater police regulation, sections of the movement switched to a more radical ask of defunding the police. This created a fissure within the movement.

The rest of the paper is divided as follows. The next section is a literature review of the theoretical apparatus used in this paper and other related works. Section 3 introduces the theoretical framework discussed in the paper. This is divided into subsection 3.1, where our set of norms of compromise is defined, and subsection 3.2, which describes the non-cooperative game and stability conditions that drive the results discussed in the body of the paper. The results are discussed in section 4. I discuss the effect of polarization and radicalization of moderates in section 4.1 and 4.2 respectively. I conclude with section 5.

2 Related Literature

This paper uses the framework provided by non-cooperative games with partition functions (Ray and Vohra, 2013, Yi, 1997, Diamantoudi and Xue, 2007). The tension between cooperation and competition has been studied in many papers like Finus and McGinty

(2019), Levy (2004) and Dotti (2020). Of these Finus and McGinty (2019) has focused on the specific relationship between diversity and cooperation the most. The authors show that diversity can sometimes be good for enabling cooperation by using a Cournot model of competition and cartel formation. They also discuss the multidimensionality of diversity and show that diversity in some dimensions may not be effective in enabling cooperation. My contribution is an added layer of 'norms of compromise' that interacts with diversity for cooperation to emerge. These norms of compromise provide a very specific structure on decision making within a coalition. This structure is different from that of Demange (2004) who models this as an exogenous hierarchy/network of decision making. In the present work, it is imposed by an exogenous function that determines how the cost of compromise within a coalition is distributed among its members through a single compromise policy. Specifically, this structure/function is parameterized by a variable that measures the relative sensitivity of a compromise policy to preferences of the moderate with respect to the extremist. To the best of my knowledge, this is a new way of teasing out concrete relationships between distribution of preferences and structural differences among societies.

One structure on cooperation and compromise that has been extensively studied is legislative bargaining, reviewed in Eraslan et al. (2020) and Eraslan and Evdokimov (2019). This literature investigates which policies are chosen by heterogeneous players that are playing a bargaining game (Baron, 1991; Cho and Duggan, 2009; Calvert and Dietz, 2005; Battaglini, 2020). In the context of the present work, the bargaining literature is important in providing micro-foundations for policy outcomes achieved by a norm of compromise. This paper assumes the existence of certain kinds of effective norms of compromise without going into their micro-foundations. This allows me to focus on explicitly studying how different norms of compromise interact with preferences to make cooperation successful.

Games that model information frictions (Barbera and Jackson, 2020; Dai and Yang, 2019, Battaglini, 2017) are generally very concerned with the coordination problems that arise because of diversity. Dai and Yang (2019) directly address the tension between joining a coalition and staying independent when preferences are heterogeneous. They model a coordination game in which an agent forgoes her independence when she joins an organization that aggregates information and chooses the strategy of the agent with median preferences within the organization. This fundamental trade-off between cost of compromise and cost of fragmentation is very similar to the one I use in this paper. The main difference is between their focus on information aggregation, and my focus on norms of compromise. This shift away from information aggregation allows me to talk of a system of compromise within organizations that could be separate from their system of information aggregation. This distinction effects the results of the model where I find that cooperation emerges only if the compromise policy is sufficiently sensitive to *both* extremes *and* moderates. Information aggregation models on the other hand are generally looking for solutions to cooperation by worrying solely about those who are at the extremes.

There are many ways to model social interaction or norms or culture [2]. Norms of compromise in the present work are modeled as exogenously determined protocols used by coalitions to arrive at compromise policy. They are slightly different from 'norms' understood as outcomes reached in a repeated game succinctly reviewed in Bisin and Verdier (2017). Instead, one can think of these norms as historically (or culturally) determined rules of an underlying strategic game agents play to coordinate which compromise policy is chosen by a coalition. This is akin to the concept of 'core equilibrium beliefs' discussed in Schofield (2006) in which it is hypothesized that 'core equilibrium beliefs' determine which game agents believe they are playing at a given point of time. A natural way to fit the norms of compromise in the present paper to this literature is to endogenize these norms by studying their evolution over time. This is a promising direction of further study that is beyond the scope of this paper.

Lastly, these norms of compromise can also be interpreted as a class of preference aggregators used in social choice theory. In the context of social choice theory, the class of preference aggregators considered in the paper violate the assumption of neutrality[3]. This is why the aggregators in question can be uniquely distinguished by a continuous parameter and violate May's theorem and Arrow's theorem. Despite being in a world of single peaked preferences these norms are not strategy-proof when there is imperfect information [4]. To simplify the analysis I assume perfect information. While this is a strong assumption it is not out of place given the motivation behind norms of compromise[5].

3 Setup

This section is split in two parts major parts. The first part, subsection 3.1, defines and investigates the properties of the class of norms of compromise considered. The second part, subsection 3.2, describes the model for which these norms of compromise are applied. I use the example of this model to show that norms of compromise and diversity work together for cooperation to emerge.

Preferences

The model is a one stage game of complete information with a finite set of agents, $\alpha \in \mathcal{N} = \{1, ..., N\}$. I consider cases where $N > 2$ to eliminate trivial cases. Agents have

[2] Alesina and Giuliano (2005), Young (2015), Schofield (2006), Bisin and Verdier (2017), Gorodnienko and Roland (2015): These papers conceive of norms, culture and social interaction very differently from each other

[3] Also known as the Independence of Irrelevant Alternatives Axiom. These norms however satisfy other axioms in Arrow's theorem such as Anonymity and Weak Pareto Optimality.

[4] See Barbera et al. (1993) for discussion on which types of social choice functions are strategy-proof.

[5] For example the equilibrium policy arrived at by playing contest games as modeled in Duggan and Gao (2019) can also be interpreted in terms of social choice functions. These social choice functions will also not be strategy-proof and require the assumption of perfect information for equilibrium to be reached. These contest games give us the same compromise policies for a subclass of the norms of compromise considered in the present work

Euclidean preferences over a single policy dimension with their ideal policy denoted by $y_\alpha \in \mathbb{R}$. In other words, if the distance between some policy y and y_α, $d_\alpha(y) \equiv |y - y_\alpha|$, is lesser than the distance between another policy y' and y_α agent α prefers y over y'.

Without loss of generality I assume $y_1 \leq ... \leq y_N$, and to avoid trivial solutions assume $y_1 < y_N$. For the purposes of this analysis I modify the ideal policy space to $x = \frac{y-y_1}{y_N-y_1}$. This will not change the qualitative results of the paper and will simplify analysis of the effect of the distribution of preferences. For the rest of the paper $x_\alpha = \frac{y_\alpha-y_1}{y_N-y_1}$ will be referred to as the preferred policy of α, thus $x_1 = 0$, $x_N = 1$ and $x_\alpha \in [0,1]$. The set of ideal policies of all agents, $\mathcal{A} = \{x_1,...,x_N\}$, fully describes the distribution of preferences within a society.

3.1 Norms

'Norms of compromise' determine the proposed or 'compromise' policy of coalitions. A norm of compromise is essentially a protocol used by a coalition to account for the preferences of all agents in it and minimize the collective cost of compromise. In the present work, a norm is a rule characterized by a number $\rho \in (0,1)$ which determines the proposed policy of a coalition given the ideal policies of all the agents belonging to the coalition. Consider a coalition $C \subset \mathcal{N}$ containing m agents. Let $A \in [0,1]^m$ be the list or distribution of ideal policies of all agents in C. Then the proposed policy or 'compromise' of the coalition C is given by:

$$f^\rho(A) = \arg\min_{x\in\mathbb{R}} \Big(\sum_{\alpha\in C} \big(d_\alpha(x)\big)^{\frac{1}{\rho}} \Big)^\rho \tag{1}$$

The distance from a proposed policy, $d_\alpha(f^\rho(A))$, describes the cost of compromise that the coalition C imposes on agent α. At $x = f^\rho(A)$ the socially determined cost of compromise the coalition imposes on all its members of is minimized. This socially determined cost is parametrized by ρ. Moderates of a coalition are defined as agents closest to the proposed policy, $f^\rho(A)$, and extremists of a coalition are agents that are furthest from the policy proposed[6]. We will see the exact relationship between ρ and the sensitivity to moderates when we investigate the properties of the norms of compromise. An alternative way of thinking about a norm of compromise, is that it is a measure of centrality given by ρ which a coalition uses to choose its proposed policy[7].

The solution of the optimization problem in (1) always exists and is unique [8], for $\rho \in (0,1)$.

[6] In the grand coalition, either 1 or N will be an extremist depending on which policy is chosen.

[7] Another interpretation of ρ comes from the use of a constant elasticity of substitution objective function in equation 1. One can think of this objective function as the cooperation compromise objective function, or the specific way this coalition minimizes the cost of compromise of each agent. With this interpretation, ρ determines of elasticity of substitution of costs of compromise of agents in the social compromise objective function. The exact measure of elasticity of substitution of a moderate's cost of compromise with respect to an extremist's cost of compromise is $\frac{\rho}{1-\rho}$. Note, when $\rho = 0$ the elasticity of substitution is 0 and when $\rho \to 1$ then elasticity of substitution tends to arbitrarily large values.

[8] See appendix for all proofs.

This means $\forall\ \rho \in (0,1)$, the function f^ρ is well-defined for any distribution of preferences $A \in [0,1]^m$. As this paper focuses only on policy proposals generated by the function f^ρ, $d_\alpha^\rho(A) \equiv d_\alpha(f^\rho(A))$ denotes distance of α's preferred policy from the coalition's proposed policy $f^\rho(A)$ where A is the distribution of preferences and ρ is the norm of compromise. To get a sense of how the proposed policy changes with ρ let us explore the following example that has a closed form solution of f^ρ.

Example 3.1. *Let C be a coalition that uses the norm of compromise ρ. Suppose $\forall \alpha \in C$, $x_\alpha \in \{x, y\}$ with $x < y$. Let l denote the total number of agents with $x_\alpha = x$, and $r \equiv m - l$ the number of agents with $x_\alpha = y$. Then,*

$$f^\rho(A) = \frac{x + y\left(\frac{r}{l}\right)^{\frac{\rho}{1-\rho}}}{1 + \left(\frac{r}{l}\right)^{\frac{\rho}{1-\rho}}}$$

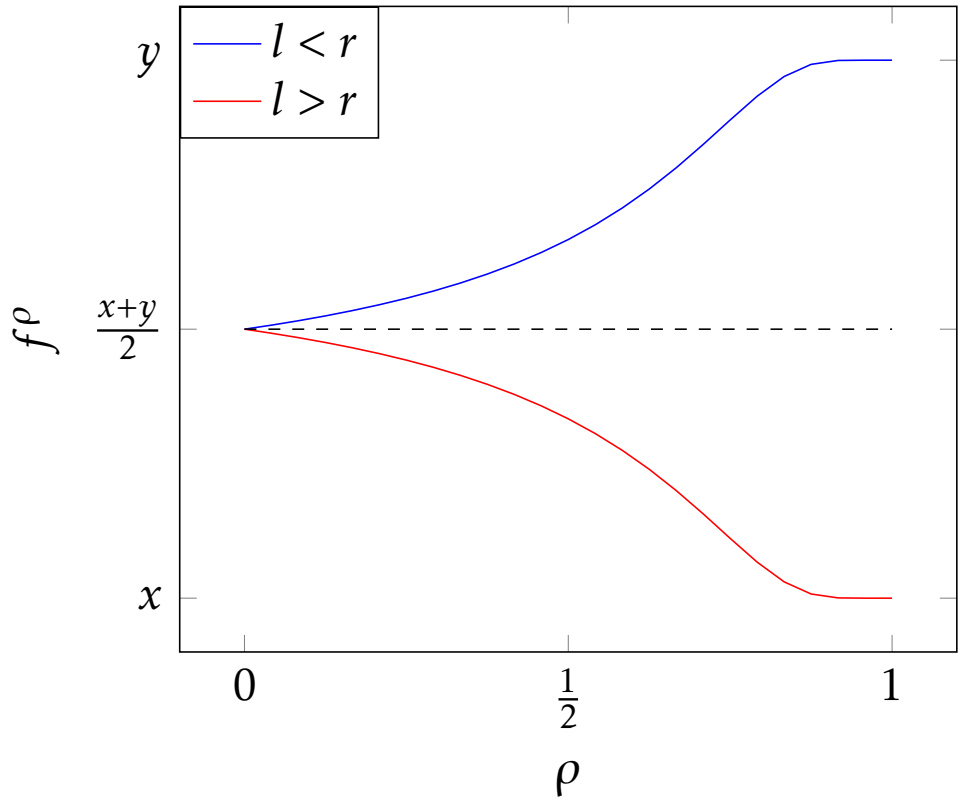

Figure 1: When $r > l$ then $f^\rho(A)$ is strictly increasing. When $r < l$ then $f^\rho(A)$ is strictly decreasing.

Figure 1 shows how f^ρ changes if there are only two types of agents within a coalition. We see that f^ρ converges to $(x+y)/2$ as $\rho \to 0$ whether $l > r$ or $l < r$. When $r > l$ then f^ρ increases as ρ increases converging to the median agent's preferred policy, y. When $r < l$ then f^ρ decreases as ρ increases, finally converging to the median agent's preferred

policy, x. This example illustrates how dependent policy proposals are on the norm of compromise *and* the distribution of preferences within a coalition. It follows that costs of compromise imposed by the coalition on each agent also depends on this interaction between norms of compromise and the distribution of preferences within a coalition.

Norms of compromise at $\rho \to 0$

In the example above, at the limit values of ρ, one can see how the sensitivity of the policy proposed changes from the median to the extremists. Even for a general distribution of preference this change in sensitivity is most stark at the limiting values of ρ.

$$f^0(A) = \lim_{\rho \to 0} f^\rho(A) = \arg\min_x \max_{\alpha \in C} d_\alpha(x) = \frac{\max_{\alpha \in C} x_\alpha + \min_{\alpha \in S} x_\alpha}{2} \tag{2}$$

Expression (2) states that when sensitivity to the cost of risk imposed by the coalition becomes arbitrarily large (or $\rho \to 0$) the norms of compromise will only consider minimizing the cost of risk of the extremes within a coalition. This is because these agents bear the highest cost of compromise within the coalition from the compromise policy.

Norms of compromise at $\rho \to 1$

When $\rho = 1$ then the solution to (1) is no longer a function for all possible coalitions in $\mathcal{N}$. This is because the solution to (1) at $\rho = 1$ is the set of values in $[0,1]$ that are a median to the distribution of ideal points A. From here on out assume for the sake of simplicity at $\rho = 1$,

$$f^1(A) = \begin{cases} x_\alpha & m \text{ is odd; } x_\alpha = med\,\langle A \rangle \\ \frac{x_\alpha + x_\beta}{2} & m \text{ is even; } x_\alpha, x_\beta \in med\langle A \rangle;\ \alpha \neq \beta \end{cases} \tag{3}$$

where $med\langle A \rangle$ is a set of all median values in A. This means that at $\rho = 1$, either the unique median ideal policy of agents in C is chosen or the midpoint of two distinct median ideal policies in A is chosen by the coalition. These functional forms of f^ρ at the limits give us an idea of which norms are more sensitive to extremists and which to moderates. When ρ is close to 0 then the compromise policy is more sensitive to the preferences of the extremists relative to the moderates. At $\rho = 0$, the compromise policy ignores the moderates' preferences; it depends only on the extremists. When ρ is close to 1, the policy proposed is sensitive to preferences of the moderates relative to the extremes. At $\rho = 1$, it ignores the extremists' preferences; it only depends on the median(s).

Norms of compromise at $\rho = 1/2$

At $\rho = 1/2$, the compromise policy is simply mean of the ideal points of all the agents in a

coalition. In other words,

$$f^{\frac{1}{2}}(A) = \frac{\sum_{\alpha \in C} x_\alpha}{m}, \tag{4}$$

where m is the number of agents within coalition C. The norm of compromise represented by $\rho = 1/2$ is very special. The compromise policy induced by the norm $\rho = 1/2$ is equally sensitive to all members of a coalition. Additionally, under $\rho > 1/2$, the compromise policy is sensitive to moderates and under $\rho < 1/2$ the compromise policy is sensitive to extremists. This discussed in greater detail when we explore the properties of the norm.

Before we continue to a more detailed discussion on the sensitivity of a compromise policy to extremists or moderates in the coalition, it is important to note that the policy proposed by coalition need not be monotonic in ρ. Consider the following example:

Example 3.2. *$C \subset \mathcal{N}$ has a distribution of preferences given by $A = (0, 0, 2/5, 2/5, 1)$.*

Here $f^1(A) = 2/5$, $f^{\frac{1}{2}}(A) = 9/25 < 2/5$ and $f^0(A) = 1/2 > 2/5$. The non-monotonicity is shown on a schematic diagram in figure 2.

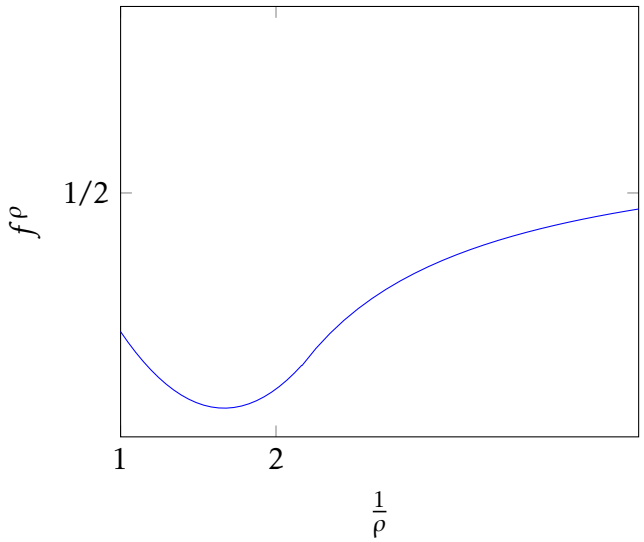

Figure 2: A schematic diagram of $f^\rho(A)$ as a function of $\frac{1}{\rho}$ when $A = (0, 0, 2/5, 2/5, 1)$

In this example, the minimum value of f^ρ, the proposed policy, is outside the range $[med\langle A\rangle, 1/2]$. It is possible to numerically see that the minimum value of f^ρ is reached for this distribution when $\rho \in (1/2, 1)$. If one were to slightly change the median value in A, we would find that when $\rho \in (0, 1/2)$ as ρ increases, the compromise policy does not change very much with the median value i.e. the compromise policy is not sensitive to preferences close the median. On the other hand changing $med\langle A\rangle$ when $\rho \in (1/2, 1)$, the compromise policy is much more sensitive to preferences of the median. This already alludes to an interesting property of the norm: increasing sensitivity of the policy proposed to the moderate by increasing ρ is not equivalent to saying there is a monotonic change in the policy proposed as ρ increases.

3.1.1 Properties of the Norm

To better understand the effect of norms and the distribution of preferences on cooperation, let us see how perturbations of preferences within a coalition affect policy proposals under a norm. The following expression gives us the relationship between a social norm and the preferences of an agent α.

Lemma 3.1. *Let $\rho \in (0,1)$, and $\alpha \in A \subseteq \mathcal{N}$ with $x_\alpha = x$ then,*

$$\left.\frac{\partial f^\rho(A)}{\partial x_\alpha}\right|_{x_\alpha = x} = \left.\frac{\left(d_\alpha^\rho(A)\right)^{\frac{1-2\rho}{\rho}}}{\sum_{\beta \in C}\left(d_\beta^\rho(A)\right)^{\frac{1-2\rho}{\rho}}}\right|_{x_\alpha = x} \tag{5}$$

The right hand side of expression (3.1) is an increasing function of $d_\alpha^\rho(A)$ for $\rho < 1/2$ and an increasing function of $d_\alpha^\rho(A)$ for $\rho > 1/2$. Consequently for $\rho < 1/2$ is most sensitive to the preference of the extremist (whose $d_\alpha^\rho(A)$ is largest) while for $\rho > 1/2$ it is most sensitive to the preferences of the moderate (whose $d_\alpha^\rho(A)$ is the smallest). In other words, when $\rho < 1/2$, the extremes will affect the compromise policy the most. When $\rho > 1/2$, the extremes will affect the compromise policy the least. Further, when $\rho = 1/2$, the right hand side of expression (4) is independent of $d_\alpha^\rho(A)$ and hence of α. This means $f^\rho(A)$ is equally sensitive to the preferences of all agents.

This sensitivity of a compromise or proposed policy to preferences is important in understanding the trade off for α. Sensitivity towards an agent's preferences is defined as follows:

Definition 3.1. *Sensitivity of a proposed policy to an agent α's preference at a given distribution of preferences, A, within a coalition, is defined as $\left|\frac{\partial f^\rho(A)}{\partial x_\alpha}\right|$ and denoted by $S_\alpha(\rho, A)$.*

The following figure plots the sensitivity of policy proposed to α's preference ρ, against the distance of x_α from f^ρ to see how this sensitivity qualitatively changes for moderates and extremes for different ranges of ρ.

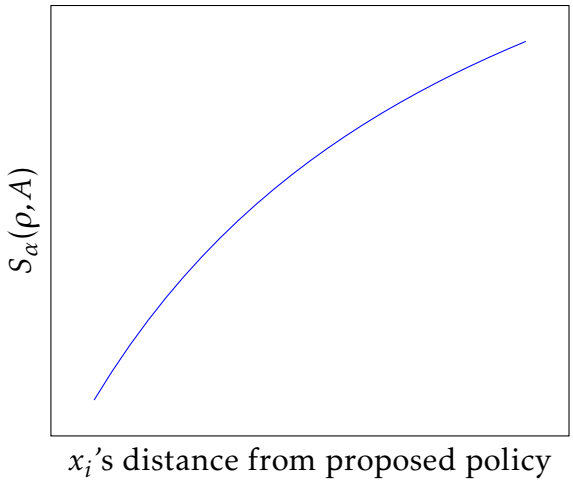

Figure 3: Sensitivity of f^ρ to α's preference increases as x_α is more distant from policy proposed when $\rho \in (0, 1/2)$

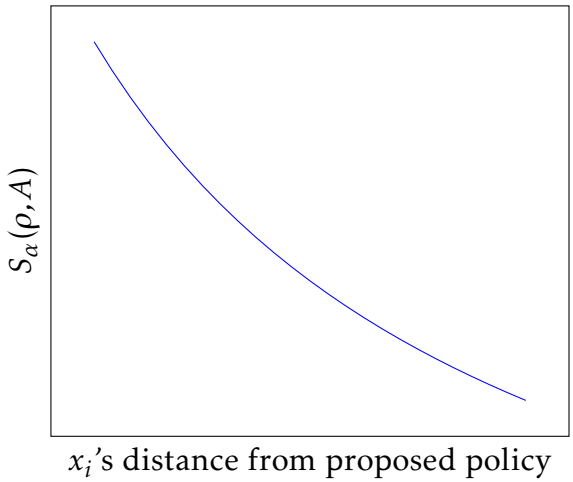

Figure 4: Sensitivity of f^ρ to α's preferences decreases as x_α is more distant from policy proposed ($\rho \in (1/2, 1)$)

To get into a better sense of how sensitivity changes for moderates and extremists, recall moderates within a coalition are agents with ideal policies closest to the policy proposed, and extremists within a coalition are agents with the ideal policy furthest away from the policy proposed. Notationally, $mod \in C$ minimizes $d_\alpha(f^\rho(A))\ \forall \alpha \in C$ and $ext \in C$ maximizes $d_\alpha(f^\rho(A))\ \forall \alpha \in C$. Therefore, the relative sensitivity, given by

$$\frac{S_{mod}(\rho, A)}{S_{ext}(\rho, A)},$$

increases as ρ increases. The elasticity of the ratio of the relative distances of the moderates and the extreme with respect to the relative sensitivity of the policy proposed under a

given norm is given by

$$\epsilon_{d,S} \equiv \frac{\%\text{ change in } \frac{d_{mod}(f^{\rho}(A))}{d_{ext}(f^{\rho}(A))}}{\%\text{ change in } \frac{S_{mod}(\rho,A)}{S_{ext}(\rho,A)}} = \frac{\rho}{1-2\rho}$$

Note, the absolute value of this elasticity is the lowest when $\rho = 0$. In the interval $(1/2, 1]$, $\rho = 1$ also has the lowest elasticity. This elasticity measures how responsive relative distances are to changes in relative sensitivity. This change in relative distances is precisely what agents factor in when considering whether or not it is worthwhile to stay within the grand coalition. In the next section a game of coalition formation is discussed which fully describes the cost of an agent fragmenting from C. This will allow us to determine the stability of a coalition once the notion of equilibrium is also defined.

One can think of norms of compromise, as defined above, as the pre-existing rules that agents tacitly agree on when they join a coalition.This makes the proposed policy endogenous to membership but exogenously constrained by the process of arriving at a proposed policy encapsulated in the norm ρ. For example, if lobbyists come together to form a coalition they may all choose a policy between them that depends on their financial contributions. In this casethe compromise policy depends on the distribution of preferences within the coalition but it also depends on the method of agreement: financial contributions. This rule of agreement that uses financial contributions to determine policy of a coalition would mean ρ is somewhere between 0 and 1/2, as shown in Duggan and Gao (2019). Alternatively, the rules of agreement within the coalition may involve majority voting; in this case $\rho \to 1$; where the policy proposed tends to the median preferred policy. This way, ρ acts as a representation of the underlying rules of compromise within a coalition. This paper is silent about the *strategic process* through which the compromise policy is arrived at. While a discussion of these micro-foundations of norms of compromise is out of the scope of this paper, it is important to underscore that each norm of compromise encapsulated by ρ can capture a special underlying exogenous social process through which agreement is reached within a coalition.

3.2 Coalition Structure

An agent, α, has two choices: she can propose and back her most preferred policy, x_α, or she can back the policy of a coalition, $C \subset \mathcal{N}$, that proposes z. This is essentially an announcement game with two possible strategies: all players that announce 1 are part of the coalition C and back its policy and all those that announce 0 are not part of C and back their own policy. This is similar to the non-cooperative coalition game structure used in Finus and McGinty (2019) and Yi (1997) where agents can join a coalition as long as they adhere to the rules within the coalition. It is important to note that this paper focuses solely on the stability of the grand coalition[9].

[9]One can generalize the concept of equilibrium that emerges from this type of game structure to one in which any coalition can block the grand coalition. For the purposes of exposition the discussion in the main body of the paper restricts the equilibrium discussion to individual deviation from the grand coalition. The proofs in the appendix solve for more general definitions of the stability of the grand coalition

The probability a public policy, $x \in \mathbb{R}$, is implemented depends on the number of agents backing x. This probability is denoted by $p(b_x, \mathcal{P})$ where b_x is the number agents backing policy x and $\mathcal{P}$ is the partition of the set of agents $\mathcal{N}$. To determine the expected utility of an agent the functional form of the utility of agent α when policy x is implemented is given by:

$$u(x) = -|x_\alpha - x|^\theta \equiv -d_\alpha(x)^\theta. \tag{6}$$

Here $\theta \geq 1$ captures the level of relative risk aversion of an agent [10]. This means that an increase in θ will decrease the attractiveness of a gamble of policy positions with respect to a safe policy position. One way to interpret θ is the dislike agents have for uncertainty of public policy implementation. This distaste for uncertainty about which public policy is implemented would likely increase when there are sudden exogenous shocks like a pandemic or war[11].

Finally, let us denote $\mathcal{A} \in [0,1]^N$ as the full distribution of ideal points of the agents in $\mathcal{N}$. The expected utility of an agent α for a given partition $\mathcal{P}$ that describes whether or not agents choose to be in a coalition C under norm ρ is given by

$$\mathbb{E}_\alpha^\theta(\mathcal{A}, z, \mathcal{P}) = \int_x p(b_x, \mathcal{P}) d_\alpha(x)^\theta = -p(m, \mathcal{P})(d_\alpha(z))^\theta - \sum_{\beta \notin S} p(1, \mathcal{P})(d_\alpha(x_\beta))^\theta \tag{7}$$

where the expected utility of α depends on the distribution of preferences in society, $\mathcal{A}$, the policy proposed within the coalition, z and the partition $\mathcal{P}$ which generates the probability distribution over the backed policies given by $p(b_x, \mathcal{P})$. Note that agents can join only one coalition for more than two people. This assumption will allow us to clearly identify the trade-off agents face when deciding to join or leave a coalition.

Given that $z = f^\rho(A)$ in this model, both the probability distribution and the proposed policy is determined by which agents choose to join the coalition. In fact, agent α's expected utility from a partition can be re-written as:

$$\mathbb{E}_\alpha^{\rho,\theta}(\mathcal{A}, C) = -p(m, \mathcal{P})(d_\alpha^\rho(A))^\theta - \sum_{\beta \notin S} p(1, \mathcal{P})(d_\alpha(x_\beta))^\theta \tag{8}$$

where the expected utility of an agent from a partition depends on the norm of compromise[12], ρ, the members of the coalition, C, and the complete distribution of preferences in the society $\mathcal{A}$. An agent chooses to join C if and only if $\mathbb{E}_\alpha^{\rho,\theta}(\mathcal{A}, C \cup \{\alpha\}) > \mathbb{E}_\alpha^{\rho,\theta}(\mathcal{A}, C)$ and

[10] As distance of a policy from the implemented policy increases the risky behavior decreases (increasing absolute risk aversion of policy distance). The risky behavior with respect to the current distance of policy remains constant (constant relative risk aversion of policy distance given by θ).

[11] Consider again the related policy spaces $y \in \mathbb{R}$ and $x = \frac{y - y_1}{y_N - y_1}$. The absolute risk aversion for an agent $\alpha \in \mathcal{N}$ in both policy spaces is the same, and the relative risk aversion of an agent in the policy space determined by y is given by $\theta(y_N - y_1) + 1$.

[12] Remember $d_\alpha^\rho(A) \equiv d_\alpha(f^\rho(A))$.

$C \cap \{\alpha\} = \emptyset$[13].

The coalition structure C is said to be stable under ρ if and only if all agents in C prefer to remain in C and all agents not in C prefer to stay outside of C given the norm of compromise ρ. Formally, S is stable under $\mathcal{A}$ if and only if internal and external stability are satisfied, where

Internal stability: $\mathbb{E}_\alpha^{\rho,\theta}(C,\mathcal{A}) \geq \mathbb{E}_\alpha^{\rho,\theta}(C_{-\alpha},\mathcal{A})$, $\forall \alpha \in C$
External stability: $\mathbb{E}_\alpha^{\rho,\theta}(C,\mathcal{A}) > \mathbb{E}_\alpha^{\rho,\theta}(C \cup \{\alpha\},\mathcal{A})$, $\forall \alpha \notin C$

where $C_{-\alpha} \equiv \{\beta\}_{\beta \in C:\beta \neq \alpha}$. Full cooperation is considered stable under ρ when the grand coalition, $C = \mathcal{N}$, is stable under ρ.

The formulation of this model may elicit more than one equilibrium but for the purposes of this paper we will focus our attention to the case when full cooperation is stable. The next section discusses the cost of fragmenting from this grand coalition.

3.3 Cost of fragmentation

Stability of the grand coalition depends on agents' trade-off between compromising by staying in the grand coalition and increasing their risk by leaving it. This trade-off is encapsulated by the cost of fragmentation given by the expression

$$R_\alpha^{\rho,\theta}(\mathcal{A}) = p^{\frac{1}{\theta}} d_\alpha^\rho(\mathcal{A}_{-\alpha}) - d_\alpha^\rho(\mathcal{A}) \tag{9}$$

where $p = p(N-1, \langle \mathcal{N}_{-\alpha}, \{\alpha\} \rangle)$ and $\mathcal{A}_{-\alpha}$ denotes the distribution of preferences of agents in $\mathcal{N}_{-\alpha}$. The cost of leaving the grand coalition is given by $p^{\frac{1}{\theta}} d_\alpha^\rho(\mathcal{A}_{-\alpha})$ which depends on how far away the fragmented coalition's proposed policy will be from α's preferred policy, and on the probability that that policy is chosen. The cost of staying in it is given by the cost of compromise $d_\alpha^\rho(\mathcal{A})$. The cost of fragmentation is essentially the cost of leaving normalized by the cost of compromise under ρ. When the cost of fragmentation $R_\alpha^{\rho,\theta}(\mathcal{A})$, is negative agent α will fragment the grand coalition by leaving it, and when cost of fragmentation is positive α will not. Therefore, when the cost of fragmentation is non-negative $\forall \alpha \in \mathcal{N}$ then full cooperation is stable, otherwise there exists at least one agent that prefers to fragment the grand coalition by leaving it.

The riskiness of leaving the grand coalition increases with $p^{\frac{1}{\theta}} d_\alpha^\rho(\mathcal{A}_{-\alpha})$ and decreases with $d_\alpha^\rho(\mathcal{A})$ for α. Conversely, the willingness to leave the grand coalition increases in the cost of compromise, $d_\alpha^\rho(\mathcal{A})$, and decreases with $p^{\frac{1}{\theta}} d_\alpha^\rho(\mathcal{A}_{-\alpha})$ for α. If $d_\alpha^\rho(\mathcal{A}_{-\alpha}) = d_\alpha^\rho(\mathcal{A})$ then agent α has no incentive to stay in the coalition. On the other hand, if $p = 1$ then full-cooperation is always stable since leaving the coalition always creates more risk for any agent α since $d_\alpha^\rho(\mathcal{A}_{-\alpha}) > d_\alpha^\rho(\mathcal{A})$ for $\rho \in (0,1)$.

[13] If an agent is indifferent between joining a coalition and being independent, it is assumed she joins the coalition.

4 Results

The first relationship established is between risk aversion and the stability of the grand coalition. Stability of the grand coalition is achieved when creating divisions becomes very costly for the agents, or cost of fragmentation is very high. The most straight forward way that these costs rise is when risk aversion, θ, increases. Risk aversion captures an agent's aversion to political uncertainty created by divisions along a specific public policy dimension. For example, educated people who may be risk averse to income losses (Jung, 2015) would also be more risk averse about public infrastructure being allocated away from their preferences. This could be an alternate explanation as to why there is cooperation between diverse social groups within college campuses (Dahlum and Wig, 2017).

The first lemma states the relationship between risk aversion and cost of risk for an agent α. The larger the risk aversion to divisions, the more attractive a safe option offered by the grand coalition becomes for an agent.

Lemma 4.1. *Fix $\rho \in (0,1)$ and distribution of preferences $\mathcal{A}$, then $\exists \bar{\theta}^{\rho}_{\alpha}(\mathcal{A}) \geq 1$ such that $\forall \theta > \bar{\theta}^{\rho}_{\alpha}(\mathcal{A})$ and $\forall\ i \in \mathcal{N}$*

$$R^{\rho,\theta}_{\alpha}(\mathcal{A}) > 0 \tag{10}$$

Lemma 4.1 states that, under a fixed norm, for every α there exists a threshold of risk aversion that makes the grand coalition more attractive than any other partition. As there are a finite number of agents, there must exist a maximum threshold level of risk aversion. Any risk aversion larger than this maximum threshold of risk aversion makes the grand coalition attractive for *all* agents in the population. This leads us to the first proposition.

Proposition 4.1. *Fix $\rho \in (0,1)$, then $\exists \bar{\theta}^{\rho}(\mathcal{A}) \geq 1$ such that $\forall \theta > \bar{\theta}^{\rho}(\mathcal{A})$ the grand coalition is stable.*

For large enough risk aversion, under a norm ρ, the grand coalition is the most preferred partition for any agent in $\mathcal{N}$. It is important to note, that for low θ the grand coalition need *not* be either stable nor efficient under a specific norm. Proposition 4.1 merely states that high enough risk aversion overrides the instability of grand coalition for a given norm. This relationship between θ and ρ is illustrated in figure **??**. The shaded region represents the θ, ρ combinations where full cooperation is stable. Outside the shaded region risk aversion is not high enough to induce cooperation for a given ρ and θ.

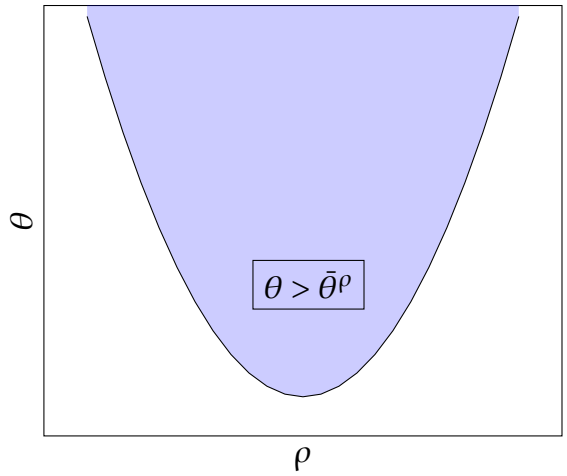

Figure 5: A schematic diagram for when full-cooperation emerges for a given level of risk-aversion and norms of compromise. Full cooperation is possible in shaded region.

Proposition 4.1 alludes to the fact that when risk-aversion is high enough, the distribution of preferences do not matter. Therefore, when risk aversion is high enough there is no relationship of diversity of preferences and norms of compromise. This result is here so that when we begin exploring the relationship between norms of compromise, distribution of preferences and cooperation it is clear that the societies investigated has low enough risk aversion, θ so that variation is possible. The next two sections that explore this relationship with this clear assumption that risk aversion is low.

4.1 Polarization and Norms of Compromise

In this section, how effectively norms of compromise enable cooperation when agents' policy preferences polarize. For simplicity the analysis is restricted to symmetric distribution of preferences. The results show that norms of compromise that enable cooperation within a polarized society may not enable compromise within an unpolarized society and vice versa. Before defining what polarization is let us consider an example by comparing two societies with different levels of polarization. One society is fully polarized, i.e. agents only have preferences for extreme policies. The other society is 'minimally' polarized, i.e. all agents are concentrated in the middle of the spectrum except agents 1 and N who have preferred policies at 0 and 1 respectively. For the same norm, agents in a polarized society have different costs of fragmentation than agents in an unpolarized society. This is true even for extremists whose positions do not change. When norms are too sensitive to moderates, ρ close to 1, then extremes have low costs of fragmentation in an unpolarized society. This is because these norms substitute the extreme agents' preferences with the moderate agents' preferences when reaching a norm of compromise, making the costs of fragmentation low for the extremes.

Example 4.1. *Let $\theta \approx 1$, consider $\mathcal{A}$ which is symmetric such that $\forall \alpha \in \mathcal{N} \setminus \{1, N\}$ $x_\alpha = 1/2$ and $N \geq 4$.*

This example describes a minimally polarized or maximally homogeneous society where all agents except 1 and N are have their preferred policies at the mid-point of the policy spectrum,$x = 1/2$. In this example, clearly, all moderates, agents with preferred policy, 1/2, will be indifferent between staying and leaving the coalition. The threat of fragmentation comes from the extremists. Extremist agents face very small costs of fragmentation when ρ is close to 1 because policy proposed by the fragmented coalition, $f^{\rho}(\mathcal{A}_{-1})$, is very close to policy proposed by the grand coalition, $f^{\rho}(\mathcal{A})$. The reason policy proposed by coalition does not change much when it is fragmented by the extremists is because for high ρ the proposed policy is most sensitive to moderates all of whom are concentrated at 1/2. Alternatively, when ρ is low $f^{\rho}(\mathcal{A})$ is very sensitive to the extreme positions. If an extremist leaves the coalition, this makes the policy proposed by the fragmented coalition risky as the new policy proposed will get closer to the other extremist. This new risky proposition will raise the cost of fragmentation high enough that neither extreme will be willing to leave the grand coalition.

Example 4.2. *Fix θ, consider $\mathcal{A}$ which is symmetric such that $\forall \alpha \in \mathcal{N}$, $x_\alpha \in \{0,1\}$. Let l be the number of agents with ideal point x_α. If $l = N - l$, $\exists \hat{\rho}$ such that $\forall \rho > \hat{\rho}$ the grand coalition is stable and $\forall \rho < \hat{\rho}$ the grand coalition is not stable.*

This example describes a society in which is 'maximally' or fully polarized i.e. agents either have their preferred policy at 0 or 1. When $l = N - l$, the grand coalition has a symmetric distribution of preferences where all agents have median preferences. If N fragments from the grand coalition then agents with preferences at 0 will become the new medians of the fragmented coalition and if 1 fragments from the grand coalition then agents with preferences at 1 become the median. Unlike the previous example, if an extremist fragments from the grand coalition the agents at the other end of the political spectrum become moderates in the new fragmented coalition. Therefore, the risk of fragmentation for an extremist is highest when norms of compromise are sensitive to moderates (ρ close to 1) as policy proposed by the fragmented group will be much further away from the extremist than if she stays in the grand coalition.

This incentive to stay in the grand coalition gets reversed for very low ρ when the proposed policy is more sensitive to extreme preferences relative to the moderates. The proposed policy of the fragmented coalition will not change much from the policy proposed in the grand coalition because everyone in the society is an extremist. As cost of fragmentation is be small in a maximally polarized society full cooperation breaks down. Already with these examples we can see how the same norms of compromise have opposing effects on cooperation for starkly different levels of polarization in a society. Let us investigate a more general result about this relationship between polarization and norms by giving a formal definition of polarization.

Given a symmetric distribution of preferences, polarization is defined as a symmetric change in the preferred policy of agents towards extreme positions.

Definition 4.1. *Consider a symmetric distribution $\mathcal{A} = (x_\alpha)_{\alpha \in \mathcal{N}} \in [0,1]^N$. $\mathcal{A}$ is symmetrically polarized to another symmetric distribution of preferences $\mathcal{A}' \in \mathbb{R}^N$ if and only if $v = \mathcal{A}' - \mathcal{A}$ is also symmetric and $v_\alpha < 0$ if $x_\alpha < 1/2$ and $v_\alpha > 0$ if $x_\alpha > 1/2$.*

Here $\|v\|$ tells us the magnitude of polarization.The following proposition illustrates the effect of polarization on cooperation depends on the norms of compromise. When norms are sensitive to the moderates, $\rho > 1/2$, then the willingness to be part of a coalition increases with polarization for agents who are not polarized. However, when norms are sensitive to extremes, $\rho < 1/2$ then the willingness to be part of a coalition decreases with polarization for agents who are not polarized.

Proposition 4.2. *Consider a symmetric distribution of preferences given by $\mathcal{A}$ that is symmetrically polarized to $\mathcal{A}'$ with $v = \mathcal{A}' - \mathcal{A}$. Let $\beta \in \mathcal{N}$ be such that $v_\beta = 0$.Then,*

- $\rho < 1/2$,

$$v.\nabla R_\beta^{\rho,\theta} < 0$$

  In words, if full cooperation is not stable to deviations from β at $\mathcal{A}$ under ρ then full cooperation is not stable to deviations from β at $\mathcal{A}'$ under ρ. increases.

- $\rho > 1/2$,

$$v.\nabla R_\beta^{\rho,\theta} > 0$$

  In words, if full cooperation is not stable to deviations from β at $\mathcal{A}'$ under ρ then full cooperation is not stable to deviations from β at $\mathcal{A}$ under ρ.

where ∇R_β is the partial of the cost of fragmentation of β with respect to the distribution of preferences $\mathcal{A}$.

Proposition (4.2) states that for agents that are not getting polarized, polarization of preferences makes the grand coalition less attractive under norms that are more sensitive to the extremists. Therefore, polarization is 'destabilizing' for full cooperation when $\rho \in (0, 1/2)$. Conversely, under norms that are more sensitive to moderates, polarization makes full cooperation more attractive to those who are not getting polarized. In other words, polarization is more 'stabilizing' to full cooperation when $\rho \in (1/2, 1)$.
The intuition lies in how sensitivity towards extremes and moderates play into the cost of fragmentation. A fragmented coalition's proposed policy is always further away from the agent who fragments away from the grand coalition. In a symmetric distribution, fragmentation of the grand coalition by some $\alpha \in \mathcal{N}$ results in agents at the opposite side of the political spectrum (other side of $x = 1/2$) to become moderates in the fragmented coalition and agents on the same side of the political spectrum of α closer to the extremist. This means, when $\rho \in (0, 1/2)$, extremists of the fragmented coalition pull policies closer to them. If an agent who is not polarized leaves the grand coalition, polarization brings the proposed policy closer to them than it did before. This results in a lowering of a cost of fragmentation and a decreasing willingness to cooperate. Conversely, when $\rho \in (1/2, 1)$ policy proposed is sensitive to moderates. This means the policy proposed is pulled in the direction of the new moderates in the fragmented coalition. With this, cost of fragmentation goes up and agents have a greater willingness to stay in the grand coalition.

The following example illustrates the power of this result. It shows when the grand coalition can be vulnerable to fragmentation from extremists, and when it can be vulnerable to fragmentation from moderates.

Example 4.3. *Consider a symmetric distribution of preferences given by $\mathcal{A}(y)$ where $x_\alpha \in 0,(1-y),(y),1$ for an even number of agents in a population. Let the number of agents with $x_\alpha = \{(1-y)$ or $x_\alpha = y$ be the same and equal to $m = (N/2-2)$. Further, assume that $p \geq \frac{N-1}{N}$.*

- *For extremist agents, $\alpha \in \{0, N\}$,*
  - *fix $\rho \in (0, 1/2)$, $\exists y^*_{ext}(\rho) \in (0,1)$ such that,*

    $\forall y < y^*_{ext}(\rho)$, full-cooperation is stable to deviations from extremists.
    $\forall y > y^*_{ext}(\rho)$, full-cooperation is not stable to deviations from extremists.

  - *fix $\rho \in (1/2, 1)$, $\exists y^{**}_{ext}(\rho) \in (0,1)$ such that,*

    $\forall y > y^{**}_{ext}(\rho)$, full-cooperation is stable to deviations from extremists.
    $\forall y < y^{**}_{ext}(\rho)$, full-cooperation is not stable to deviations from extremists.

- *For moderate agents $\alpha \in \mathcal{N} \backslash \{1, N\}$,*
  - *$\exists \bar{\rho} \in (0, 1/2)$ such that $\forall \rho < \bar{\rho}$ $\exists y^*_{mod}(\rho) \in (0,1)$ such that*

    $\forall y > y^*_{mod}(\rho)$, full-cooperation is stable to deviations from moderates
    $\forall y > y^*_{mod}(\rho)$, full-cooperation is not stable to deviations from moderates

  - *$\exists$ $\underline{\rho} < 1/2$ such that if $\rho > \underline{\rho}$, $\forall y \in (0,1)$*

    full-cooperation is stable to deviations from moderates.

This result is illustrated in figure **??**. When polarization is high and ρ is low moderates will leave the grand coalition. At high polarization, moderates do not have much to lose from fragmentation as norms of compromise are very sensitive to extremes that remain in the coalition. As moderates are themselves close to extremes fragmentation is no longer as costly. When norms of compromise are sensitive to moderates, $\rho > 1/2$, polarized moderates will always find fragmentation costly. As long as moderates stay in the grand coalition agents of both sides of the political spectrum are moderates within the grand coalition. Any fragmentation of the grand coalition by moderates of one side of the political spectrum will result in agents on the other side of the political spectrum become the sole moderates in the fragmented coalition. When $\rho > 1/2$, this creates bigger risks of fragmentation for moderates. Therefore, the willingness of moderates to cooperate always increases as polarization increases. when $\rho > 1/2$.

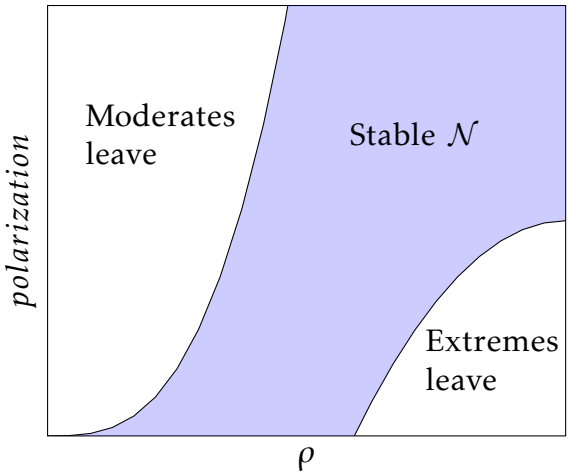

Figure 6: Schematic diagram of the relationship between polarization and norms of compromise

For extremists the intuition for staying or leaving a coalition is very similar. When polarization increases and norms of compromise become more sensitive to the extremes, extremists find it costlier to fragment. This is why when polarization is high enough extremists do not fragment from the grand coalition. In contrast, when polarization increases and norms of compromise are more sensitive to extremes ($\rho < 1/2$), the incentive to stay in the grand coalition decreases. When $\rho < 1/2$, the extremists within the fragmented coalition are closer to the extremes. When an extremist of the grand coalition fragments it, the fragmented coalition's proposed policy is most sensitive to its extremists which are close to the extremists of the grand coalition. Therefore, fragmentation for the extremists becomes less costly as polarization of the moderates increases. This is why for sufficiently large polarzation extremists are no longer willing to be part of the grand coalition.

The examples in this section show that the same distribution of preferences may not support full cooperation because norms of compromise may decrease. This may explain why countries in the Arab spring saw such many types of protests. Some protests were fragmented like Libya and Yemen where moderates within the protesting group were fragmented into extremists and moderates. Some were more cohesive protests like Tunisa and Egypt which had a a single coalition of protesters that were able to get to the point of drafting constitutions after overthrowing the dictatorship government. These differences in the character of protests may have come from the differences in the way norms of compromise work in these societies. People in Yemen and Libya are have stronger tribal ties (Khan and Mezran (2013), Holm (2013)) and so compromise between these tribal groups may involve a contest game with lower ρ.

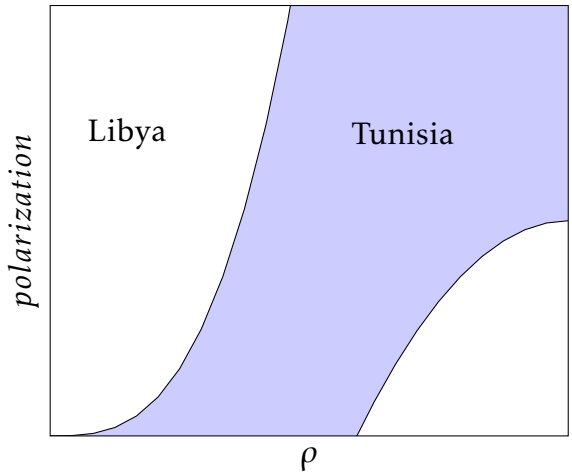

Figure 7: Schematic diagram of where countries in the Arab Spring were placed in the polarization/norms axis

A natural question to ask is which norms enable cooperation independent of polarization. This is discussed in the next section.

Proposition 4.3. *Let* $p \geq \frac{N-1}{N}$ *then,* $\forall \theta \geq 1$ *full cooperation is always stable under* $\rho = 1/2$.

For $\theta \neq 2$, $\rho = 1/2$ does not propose the most efficient policy in the garnd coalition, but under the norm $\rho = 1/2$ induces the most efficient partition it can, which is the grand coalition. Thus, $\rho = 1/2$ is able to make agents sufficiently internalize the costs of risk and compromise so that the most efficient partition under the norm is stable to deviations.

The relationship of norms and diversity of preferences studied through polarization has been by comparing symmetric distributions. However, it is worth considering what happens when symmetric distributions skew towards one extreme position because the change of preferences of the moderate. In the next section, we will see how the willingness to cooperate of the extremist changes for different types of norms when moderates change their preferences towards them.

4.2 Effect of radicalization of moderates on extremes

Even small changes to the moderates preferred policy position has significant implications on the cost of fragmentation for extremes. Consider a distribution given by $\mathcal{A}$ which is symmetric and $1/2 = x_\alpha \in \mathcal{A}$. Then, for $\rho < 1/2$, the effect of a distortion of α's ideal point makes the grand coalition more attractive to the agent who is furthest away from the distortion. Where as, for $\rho > 1/2$, the agent who is in the same direction of the distortion is finds the grand coalition more attractive.

Proposition 4.4. *Let* $\rho \in (1/2, 1)$ *and* $\mathcal{A}$ *be the distribution of preferences such that* $\exists \alpha \in \mathcal{N}$

where $x_\alpha = f^\rho(\mathcal{A})$. *Then*

$$\left.\frac{\partial R_N^{\rho\theta}(\mathcal{P})}{\partial x_\alpha}\right|_{x_\alpha = f^\rho} > 0 \tag{11}$$

$$\tag{12}$$

In words, if at $\mathcal{A}$ *the grand coalition is stable to deviations from* N *under* N*, then when a moderate shifts her preferences towards* N *the grand coalition will still be stable to deviations from* N *under* ρ.

The intuition behind proposition 4.4 comes from the sensitivity of norms of compromise towards moderates. When $\rho > 1/2$ the proposed policy is sensitive to the moderate's position. This means when a distribution is skewed towards $x_N = 1$, agent N's influence within the grand coalition will increase relative to its influence on a fragmented partition. This makes the grand coalition more attractive to this extreme, in other words the cost of fragmentation increases with this perturbation. A similar argument can be made to agent 1's willingness to stay in the grand coalition when the moderate's preferred policy comes closer to hers.

Proposition 4.5. *Let* $\rho \in (0, 1/2)$ *and* $\mathcal{A}$ *be the distribution of preferences. If* $x_\alpha = f^\rho(\mathcal{A})$ *then*

$$\left.\frac{\partial R_N^{\rho\theta}(\mathcal{P})}{\partial x_\alpha}\right|_{x_\alpha = f^\rho} < 0 \tag{13}$$

In words, if at $\mathcal{A}$ *the grand coalition is not stable to deviatitions from* N *under* ρ*, then when moderate shifts her preferences towards* N *the grand coalition will still not be stable to deviations from* N *under* ρ.

The intuition behind proposition 4.5 given by the sensitivity of norms towards the extremes. When $\rho < 1/2$, policy proposals are most sensitive to the extremes. Therefore, when distribution is skewed toward an extreme the influence of the extreme is diminished within the grand coalition compared to a fragmented partition. This is why, even when the distribution of preferences is skewed toward an extreme, the grand coalition becomes less attractive for that extreme when $\rho < 1/2$. This could explain when the Black Lives Matter movement gained momentum and moderates were changing their positions towards more police regulation, many in the black lives movement took the conversation further extreme to demand for 'defunding the police' creating fissures within the movement.

Proposition 4.5 and 4.4 illustrate that skewness of a distribution of preferences may affect the cost of fragmentation very differently for different social norms of decision making. Thus, the effect of moderates moving closer to the extremes on the stability of cooperation may entirely depend on what norms of compromise are employed within the coalition. We are agnostic as to which norms are better or worse.

5 Conclusion

This paper provides a theoretical framework to explore the combined effect of norms of compromise and distribution of preferences on the emergence of cooperation. Full cooperation under a norm for a given distribution emerges when the grand coalition is stable to individual deviations under a norm. I find that high risk aversion end up stabilizing the grand coalition for a given norm and preference distribution. Importantly, when risk aversion is low enough, the norms of compromise play a crucial role in explaining whether or not polarization will enable cooperation. It is also important whether the radicalization of the moderate makes the extremist more willing to stay in the coalition. The key driving force of these results is the sensitivity a norm of compromise has towards a moderate with respect to the extreme. When the norm is more sensitive to moderates then polarization is stabilizes the grand coalition whereas when it is more sensitive to extremists polarization is 'destabilizes' the grand coalition.

Further work would include extending the model to multiple dimensions, which would strengthen the generality of these results. Another important aspect to explore would be the effect of the position of the status quo on cooperation. This model has abstracted away from the status quo in order to focus solely on the relationship between cooperation and social norms. Once status quo is included to the model, the political process that chooses policy proposals will become important in determining cooperation. Introducing a status quo will allow us to look at cooperation across an entire society as opposed to a group of individuals already opposed to the status quo. This might allow for analysis on the more general question of when revolution and counter revolutions emerge.

In the real world, change of risk aversion to public good provision probably comes hand in hand with changes in social norms. This would make studying the partial effect of norms on cooperation difficult. For example, the BLM movement happened after a large change in life style but also saw a large shock on income for many citizens. The change in lifestyle of staying home and shock in income were probably highly correlated. However, it may be possible to study one hypothesis of the model in the data: the positive effect of risk aversion to public good provision on collective action. In the case of the US, the two of the largest movements in the past two centuries (the Civil Rights movement and the BLM movement) have occurred either during war or a pandemic. It would be worthwhile investigating time-series data for changing risk aversion and its affect on collective action for which this model has clear predictions. To the best of my knowledge, the literature on link between risk aversion the fragmentation of protests is scarce.

References

Alesina, A. and P. Giuliano (2005). Culture and institutions. *Journal of Economic Literature*.

Alesina, A., E. Glaeser, and B. Sacerdote (2001, October). Why doesn't the us have a european-style welfare system? Working Paper 8524, National Bureau of Economic Research.

Alesina, A. and B. Reich (2015). Nation building. Working Paper 18839, National Bureau of Economic Research.

Anderson, L. (2011). Demystifying the arab spring: Parsing the differences between tunisia, egypt, and libya. *Foreign Affairs 90*(3), 2–7.

Barbera, S., F. Gul, and E. Stacchetti (1993). Generalized median voter schemes and committees. *Journal of Economic Theory 61*(2), 262 – 289.

Barbera, S. and M. Jackson (2020, 01). A model of protests, revolution, and information. *Quarterly Journal of Political Science 15*, 297–335.

Baron, D. P. (1991). A spatial bargaining theory of government formation in parliamentary systems. *American Political Science Review 85*(1), 137–164.

Battaglini, M. (2017). Public protests and policy making. *Quarterly Journal of Economics 132*(1), 485–549.

Battaglini, M. (2020, April). Coalition formation in legislative bargaining. *NBER Working Paper No. 25664*.

Bisin, A. and T. Verdier (2017). On the joint evolition of culture and institutions. *NBER working paper series*.

Calvert, R. L. and N. Dietz (2005). Legislative coalitions in a bargaining model with externalities. In D. Austen-Smith and J. Duggan (Eds.), *Social Choice and Strategic Decisions*. Studies in Choice and Welfare: Springer.

Cho, S.-j. and J. Duggan (2009). Bargaining foundations of the median voter theorem. *Journal of Economic Theory 144*(2), 851 – 868.

Dahlum, S. and T. Wig (2017, 13 September). Educating demonstrators: Education and mass protest in africa. *Journal of Conflict Resolution 63*.

Dai, L. and M. Yang (2019, 8 August). Organizations and coordination in a diverse population. *SSRN: https://ssrn.com/abstract=3051192*.

Demange, G. (2004). On group stability in hierarchies and networks. *Journal of Political Economy 112*(4), 754–778.

Diamantoudi, E. and L. Xue (2007). Coalitions, agreements and efficiency. *Journal of Economic Theory 136*, 105–125.

Dotti, V. (August 2020). Reaching across the aisle to block reform. *The Economic Journal*.

Duggan, J. and J. Gao (2019, 14 September). Lobbying as a multidimensional tug of war. *Social Choice and Welfare*.

Easterly, W. and R. E. Levine (1997, November). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics 112*.

Eraslan, H. and K. S. Evdokimov (2019). Legislative and multilateral bargaining. *Annual Review of Economics 11*(1), 443–472.

Eraslan, H., K. S. Evdokimov, and J. Zapal (2020, June 10). Dynamic legislative bargaining. ISER DP No. 1090, Available at SSRN: https://ssrn.com/abstract=3634333 or http://dx.doi.org/10.2139/ssrn.3634333.

Esteban, J. and D. Ray (1999). Conflict and distribution. *Journal of Economic Theory 87*(2), 379 – 415.

Finus, M. and M. McGinty (2019). The anti-paradox of cooperation: Diversity may pay! *Journal of Economic Behavior and Organization 157*, 541–559.

Gorodnienko, Y. and G. Roland (2015, April). Culture, institutions and democratization. Working Paper 21117, National Bureau of Economic Research.

Holm, U. (2013). Libya in transition: the fragile and insecure relation between the local, the national and the regional. In L. R. Andersen (Ed.), *How the local matter: Democratization in Libya, Pakistan, Yemen and Palistine*. Danish Institute for International Studies Report.

Jung, S. (2015). Does education affect risk aversion? evidence from the british education reform. *Applied Economics 47*(28), 2924–2938.

Khan, M. and K. Mezran (2013). *The Libyan Economy After the Revolution: Still No Clear Vision*. Atlantic Council, Rafik Hariri Center for the Middle East.

Levy, G. (2004). A model of political parties. *Journal of Economic Theory*, 250–277.

Ray, D. and R. Vohra (2013). Coalition formation. *Handbook of Game Theory*.

Reynal-Querol, M. (2002). Ethnicity, political systems, and civil wars. *Journal of Conflict Resolution 46*(1), 29–54.

Schofield, N. (2006). *Architects of Political Change*. Cambridge University Press.

Spolaore, E. and R. Wacziarg (2017, March). The political economy of heterogeneity and conflict. Working Paper 23278, National Bureau of Economic Research.

Stichnoth, H. and K. Van der Straeten (2013). Ethnic diversity, public spending, and individual support for the welfare state: A review of the empirical literature. *Journal of Economic Surveys 27*(2), 364–389.

Tufecki, Z. (2017). *Twitter and Tear Gas*. Yale University Press.

Yi, S.-S. (1997). Stable coalition structures with externalities. *Games and Economic Behavior 20*(2), 201 – 237.

Young, P. H. (2015). The evolution of social norms. *Annual Review of Economics*, 359–87.

Appendix

A Norms of Compromise

A.1 Existence and uniqueness of the compromise policy

Denote $C \in \mathcal{N}$ and m is the number of agents in C. Let A be the list of ideal policy positions of agents in C.

Lemma A.1. *Let $\rho \in (0,1)$. Define $G^\rho : \mathbb{R}^m \to \mathbb{R}$,*

$$G^\rho(A,x) = \sum_{\alpha \in C} \Big(d_\alpha(x)\Big)^{\frac{1}{\rho}} \tag{14}$$

then $G^\rho(A,.)$ is strictly convex in x.

Proof. For $\rho \in (0,1)$, $G_\rho(A,.)$ is differentiable function of x.

$$D^1 G^\rho(A,x) = \frac{1}{\rho} \sum_{\alpha \in C} \Big(d_\alpha(x)\Big)^{\frac{1}{\rho}-1} \frac{\partial d_\alpha(x)}{\partial x} \tag{15}$$

Since $\frac{\partial^2 d_\alpha(x)}{\partial x^2}[d_\alpha^{1/\rho-1}] = 0$

$$D^2 G^\rho(A,x) = \left(\frac{1-\rho}{\rho^2}\right) \sum_{\alpha \in C_i} \Big(d_\alpha(x)\Big)^{\frac{1}{\rho}-2} \left(\frac{\partial d_\alpha(x)}{\partial x}\right)^2 > 0 \tag{16}$$

■

Lemma A.2. *Fix $A \in [0,1]^{|C|}$ and let $\rho \in (0,1)$. Consider the optimization problem*

$$\min_x G^\rho(A,x) \tag{17}$$

Then policy proposal of $C \subset \mathcal{N}$, generated by $f^\rho : [0,1]^{|C|} \to \mathbb{R}$ as defined in (1) is uniquely defined, always exists and solves (17). Specifically,

$$D^1 G^\rho(A, f^\rho(A)) = 0 \tag{18}$$

Proof. Since objective function in (1) is a strictly increasing function in the value of G^ρ, if solution to (17) exists then solution to optimizing problem in (1) exists. Furthermore, any solution of (17) is a solution to (1).
Since $G^\rho(A^i,.)$ is a strictly convex differentiable function, if $x^* \in \mathcal{R}$ solves (17) then $D^1G^\rho(A,x^*) = 0$.
To show x^* exists: Let $x < x_1$, then $D^1G^\rho(A,x) < 0$ since $\frac{\partial d_\alpha(x)}{\partial x} < 0 \ \forall \ \alpha \in C$. Let $x > x_N$, then $D^1G^\rho(A,x) > 0$ since $\frac{\partial d_\alpha(x)}{\partial x} > 0 \ \forall \ \alpha \in C$. Since at $\rho \in (0,1)$, $D^1G^\rho(A,x)$ is continuously differentiable on $\mathbb{R}$, by intermediate value theorem $\exists\ x^* \in [x_1, x_N]$ such that $D^1G^\rho(A,x^*) = 0$. As $G^\rho(A,x)$ is strictly convex in x, x^* is a unique solution to (17) and (1). This implies, f^ρ is well defined function and $D^1G^\rho(A, f^\rho(A^i)) = 0$. ■

A.2 Examples of norms in section 3.1

Solving for f^ρ when $\rho \to 0$ Consider $d_\alpha(x) = \max_\beta d_\beta(x)$.

$$\Big(\sum_{\beta\in C} d_\beta(x)^{\frac{1}{\rho}}\Big)^\rho = d_\alpha(x)\Big(\sum_{\beta\in C}\frac{d_\beta(x)}{d_\alpha(x)}^{\frac{1}{\rho}}\Big)^\rho \tag{19}$$

$$\Big(\lim_{\rho\to 0}\sum_{\beta\in C} d_\beta(x)^{\frac{1}{\rho}}\Big)^\rho = d_\alpha(x)\Big(\sum_{\{\beta: d_\beta(x)=d_\alpha\}} 1\Big)^\rho \tag{20}$$

$$\implies \lim_{\rho\to 0} f^\rho(A) = \arg\min_{x\in\mathbb{R}} \max_{\beta\in C} d_\beta(x) \tag{21}$$

$$= \frac{\min\limits_{\beta\in C} x_\beta + \max\limits_{\beta\in C} x_\beta}{2} \tag{22}$$

Solving for f^ρ when $\rho \to 1$

$$G^1(A,x) = \sum_{\alpha\in C} d_\alpha(x)$$

G^1 is convex but may Guess that solution to 17 is the median.
Case 1: Suppose N is even and we have two distinct medians, $x^* < x^{**}$. In this case, then $G^\rho(A,x)$ is differentiable at $x \in (x^*, x^**)$, then from our first order condition we get

$$\sum_{\alpha\in C: x_\alpha < x^*} 1 - \sum_{\alpha\in C: x_\alpha > x^{**}} 1 = 0 \tag{23}$$

Thus, any $x \in (x^*, x^{**})$ a solution to minimizing $G^1(A,x) = \sum\limits_{\alpha\in C} d_\alpha(x)$.
Case 2: Suppose the median, x^*, is unique.

Let $I_> = \{\beta \in C : x_\beta > x^*\}$, $I_< = \{\beta \in C : x_\beta < x^*\}$
and $I_= = \{\beta \in C : x_\beta = x^*\}$.
For $x < x^*$, our first order condition is :

$$\frac{\partial G^1(A,x)}{\partial x} = |I_<| - |I_=| - |I_>| < 0 \tag{24}$$

For $x < x^*$, our first order condition is:

$$\frac{\partial G^1(A,x)}{\partial x} = |I_<| + |I_=| - |I_>| > 0 \tag{25}$$

Thus, x^* must be the minimizer of G^1.

Solving for f^ρ when $\rho = 1/2$
At $\rho = 1/2$ we have

$$D^1 G^\rho(A, f^\rho(A)) = -\sum_{\{\alpha\in C: x_\alpha > f^\rho(A)\}} (x_\alpha - f^\rho(A))^{2-1} + \sum_{\{\alpha\in C: x_\alpha < f^\rho(A)\}} (f^\rho(A) \tag{26}$$

$$\tag{27}$$

$$-\sum_{\{\alpha\in C: x_\alpha > f^\rho(A)\}} (x_\alpha - f^\rho(A)) + \sum_{\{\alpha\in C: x_\alpha < f^\rho(A)\}} (f^\rho(A) - x_\alpha) = 0 \tag{28}$$

$$m f^\rho(A) = \sum_{\alpha\in C} x_\alpha \tag{29}$$

$$f^\rho(A) = \frac{\sum_{\alpha\in C} x_\alpha}{m} \tag{30}$$

A.3 Properties of norms of compromise (3.1.1)

A.3.1 Proof of Lemma 3.1

Lemma A.3. *Let $\rho \in (0,1)$, and $\alpha \in C \subseteq \mathcal{N}$ with $x_\alpha = x$ then,*

$$\left.\frac{\partial f_\rho(A)}{\partial x_\alpha}\right|_{x_\alpha = x} = \left.\frac{\left(d_\alpha^\rho(A)\right)^{\frac{1-2\rho}{\rho}}}{\sum_{\beta\in C}\left(d_\beta^\rho(A)\right)^{\frac{1-2\rho}{\rho}}}\right|_{x_\alpha = x} \tag{31}$$

Proof. Let $x_\alpha > f^\rho(A)$
Differentiating both sides w.r.t x_α we get:

$$\left(\frac{1-\rho}{\rho}\right)\frac{\partial f^\rho(A)}{\partial x_\alpha}\sum_{\{x_\beta < f^\rho(A):\ \beta\in C_i\}}\left(d_\beta^\rho(A)\right)^{\frac{1}{\rho}-2} = \left(\frac{1-\rho}{\rho}\right)\left[\left(d_\alpha^\rho(A)\right)^{\frac{1}{\rho}-2} - \frac{\partial f^\rho(A)}{\partial x_\alpha}\sum_{\{x_\beta > f^\rho(A):\ \beta\in C_i\}}\left(d_\beta^\rho(A)\right)^{\frac{1}{\rho}-2}\right] \tag{32}$$

$$\implies \frac{\partial f^\rho(A)}{\partial x_\alpha} = \frac{\left(d_\alpha^\rho(A)\right)^{\frac{1}{\rho}-2}}{\sum_{\{\beta\in C_i\}}\left(d_\beta^\rho(A)\right)^{\frac{1}{\rho}-2}} \tag{33}$$

Similar proof for $x_\alpha < f^\rho(A)$. ∎

A.3.2 Proof related to example 3.1

Proof. Given m number of ideal points with x and $m-l$ number of of ideal points y in any coalition $C_i \subset \mathcal{N}$ from (A.2) we get:

$$l.(f^{\rho}(A)-x)^{\frac{1-\rho}{\rho}} = (m-l)(y-f^{\rho}(A))^{\frac{1-\rho}{\rho}} \tag{34}$$

$$f^{\rho}(A) = \frac{y\left(\frac{m-l}{l}\right)^{\frac{\rho}{1-\rho}} + x}{1+\left(\frac{m-l}{l}\right)^{\frac{\rho}{1-\rho}}} \tag{35}$$

So, for $l > m-l$, f^{ρ} is decreasing in ρ and for $l < m-l$, f^{ρ} is increasing ρ. Also, for $l > m-l$, $\lim_{\rho\to 1} f^{\rho}(A) = x$; and for $l < m-l$, $\lim_{\rho\to 1} f^{\rho}(A) = y$. $\lim_{\rho\to 0} f^{\rho}(A) = 1/2$. ∎

B Stability of the Grand Coalition

For a more general proof of stability we allow for a coalition $C_j \in \mathcal{N}$ to block the policy proposed by the grand coalition under norm of compromise ρ. We say C_j is a blocking coalition of $\mathcal{N}$ under ρ if and only if

$$p(m_i,\mathcal{P})u_{\alpha}(f^{\rho}(A_i)) + p(m_j,\mathcal{P})u_{\alpha}(f^{\rho}(A_j)) > u_{\alpha}(f^{\rho}(\mathcal{A})) \qquad \forall \alpha \in C_j$$

where $m_k = |C_k|$ and A_k is the vector of ideal point in coalition C_k for $k \in \{i,j\}$, and $C_i = \mathcal{N}\backslash C_j$.

Stricter Stability condition A grand coalition is strictly stable under ρ if there does not exists a blocking coalition $C \subset \mathcal{N}$ of $\mathcal{N}$ under ρ.

Note, if the grand coalition is strictly stable under ρ then it is also internally stable as defined in the main body of the paper.

For the rest of the proofs in the appendix we use this more general definition of stability to prove our results.

We denote $CE_{\alpha}^{\rho\theta}(\mathcal{P}) = (p(m_i,\mathcal{P})d_{\alpha}(f^{\rho}(A_i))^{\theta} + p(m_j,\mathcal{P})d_{\alpha}(f^{\rho}(A_j))^{\theta})^{\frac{1}{\theta}}$, where $CE_{\alpha}^{\rho\theta}(\mathcal{P})$ is the certainty equilavent distance of α from the gamble of policy proposed generated by the partition $\mathcal{P}$.

For the purposes of these proofs, cost of fragmentation is redefined for coalitional separation.

$$R_{\alpha}^{\rho,\theta}(\mathcal{P}) = CE_{\alpha}^{\rho\theta}(\mathcal{P}) - d_{\alpha}^{\rho}(\mathcal{N}) \tag{36}$$

The grand coalition is stable if and only if $R_{\alpha}^{\rho,\theta}(\mathcal{P}) > 0$.

B.1 Risk Aversion

Lemma B.1. *Fix $\rho \in (0,1)$ and N, then $R_\alpha^{\rho\theta}$ increases with θ.*

Proof. Let p_i be the probability that $g = f^\rho(A_i)$ under norm ρ.
Since $(x \ln x)$ is a convex function in x and $\left(CE_\alpha^{\rho\theta}\right)^\theta$ is a convex combination of $\left\{\left(d_\alpha^\rho(A_i)\right)^\theta\right\}_{C_i \in \mathcal{P}}$, we have:

$$\frac{\partial CE_\alpha^{\rho\theta}(\mathcal{P})}{\partial \theta} = \frac{\sum_{C_i \in \mathcal{P}} p_i \left(d_\alpha^\rho(A_i)\right)^\theta \ln\left(d_\alpha^\rho(A_i)\right)^\theta - \left(CE_\alpha^{\rho\theta}(\mathcal{P})\right)^\theta \ln\left(CE_\alpha^{\rho\theta}(\mathcal{P})\right)^\theta}{\theta (CE_\alpha^{\rho\theta}(\mathcal{P}))^\theta} > 0 \tag{37}$$

By definition

$$R_\alpha^{\rho\theta}(\mathcal{P}) = CE_\alpha^{\rho\theta}(\mathcal{P}) - d_\alpha^\rho(A_\mathcal{N}) \tag{38}$$

$$\implies \frac{\partial R_\alpha^{\rho\theta}(\mathcal{P})}{\partial \theta} = \frac{\partial CE_\alpha^{\rho\theta}(\mathcal{P})}{\partial \theta} > 0 \tag{39}$$

■

B.1.1 Proof of Lemma 4.1

Lemma B.2. *Given $\mathcal{A}$, let $\mathcal{P} \in \Pi(\mathcal{N})$ such that $\mathbb{E}_\beta^{\rho\theta}(\mathcal{P}) \neq \mathbb{E}_\beta(\mathcal{N}, \rho\theta)$ for $\beta \in \mathcal{A}$. Then $\exists C_i \in \mathcal{P}$ such that $d_\beta^\rho(A_i) > d_\beta^\rho(\mathcal{N})$.*

Proof. We know from (A.2)

$$\sum_{\{x_\alpha < f^\rho(\mathcal{A}):\ \alpha \in \mathcal{N}\}} \left(d_\alpha^\rho(\mathcal{A})\right)^{\frac{1}{\rho}-1} = \sum_{\{x_\alpha > f^\rho(\mathcal{A}):\ \alpha \in \mathcal{N}\}} \left(d_\alpha^\rho(\mathcal{A})\right)^{\frac{1}{\rho}-1} \tag{40}$$

Case 1: $x_\alpha < f^\rho(X_\mathcal{N})$
If $\mathbb{E}_\alpha^{\rho\theta}(\mathcal{P}) \neq \mathbb{E}_\alpha^{\rho\theta}(\mathcal{N})$ then $\exists C_i \in \mathcal{P}$

$$\sum_{\{x_\alpha < f^\rho(\mathcal{A}):\ \alpha \in C_i\}} \left(d_\alpha^\rho(\mathcal{A})\right)^{\frac{1}{\rho}-1} > \sum_{\{x_\alpha > f^\rho(\mathcal{A}):\ \alpha \in C_i\}} \left(d_\alpha^\rho(\mathcal{A})\right)^{\frac{1}{\rho}-1} \tag{41}$$

$$\implies f^\rho(A_i) > f^\rho(\mathcal{A}) \tag{42}$$

$$\implies d_\beta^\rho(A_i) > d_\beta^\rho(\mathcal{A}) \tag{43}$$

Case 2: $x_\beta > f^\rho(X_\mathcal{N})$
If $\mathbb{E}_\alpha^{\rho\theta}(\mathcal{P}) \neq \mathbb{E}_\alpha^{\rho\theta}(\mathcal{N})$ then $\exists C_i \in \mathcal{P}$

$$\sum_{\{x_\alpha < f^\rho(\mathcal{A}):\ \alpha \in C_i\}} \left(d_\alpha^\rho(\mathcal{A})\right)^{\frac{1}{\rho}-1} < \sum_{\{x_\alpha > f^\rho(\mathcal{A}):\ \alpha \in C_i\}} \left(d_\alpha^\rho(\mathcal{A})\right)^{\frac{1}{\rho}-1} \tag{44}$$

$$\implies f^\rho(A_i) < f^\rho(\mathcal{A}) \tag{45}$$

$$\implies d_\beta^\rho(A_i) > d_\beta^\rho(\mathcal{A}) \tag{46}$$

■

Lemma 4.1 Fix $\rho \in (0,1)$, then $\exists \bar{\theta}^{\rho}_{\alpha} \geq 1$ such that $\forall \theta > \bar{\theta}^{\rho}_{\alpha}$ and $\forall\ i \in \mathcal{N}$

$$R^{i}_{\rho\theta}(\mathcal{P}) > 0 \tag{47}$$

Proof. Let $C_j = \arg\max_{C_i \in \mathcal{P}} d^{\rho}_{\alpha}(A_i)$ and p_i denote the probability that $g = f^{\rho}(A_i)$.

$$CE^{\rho\theta}_{\alpha}(\mathcal{P}) = \left(\sum_{\alpha \in C_i} p_i \left(d^{\rho}_{\alpha}(A_i) \right)^{\theta} \right)^{\frac{1}{\theta}} \tag{48}$$

$$= d^{\rho}_{j}(A_j)(p_j)^{\frac{1}{\theta}} \left(1 + \sum_{\alpha \in C_i} \frac{p_i}{p_j} \left(\frac{d^{\rho}_{\alpha}(A_i)}{d^{\rho}_{\alpha}(A_j)} \right)^{\theta} \right)^{\frac{1}{\theta}} \tag{49}$$

Thus, $\lim_{\theta \to \infty} CE^{\rho\theta}_{\alpha}(\mathcal{P}) = d^{\rho}_{j}(A_j)$.
Since $CE^{\rho\theta}_{\alpha}(\mathcal{P})$ is strictly increasing in θ (from inequality (37)) and $d^{\rho}_{j}(A_j) > d^{\rho}_{j}(\mathcal{A})$ (from (B.2)), by the intermediate value theorem we know $\exists \bar{\theta}^{\rho}_{\alpha}(\mathcal{P}) \geq 1$ such that $CE^{\rho\theta}_{\alpha}(\mathcal{P}) = d^{\rho}_{j}(A_j) > d^{\rho}_{\alpha}(\mathcal{A})$. Since $\mathcal{N}$ is finite, $\theta^{\rho}_{\alpha} = \max_{\mathcal{P} \in \Pi(\mathcal{N})} \theta^{\rho}_{\alpha}$ is well defined. ■

B.1.2 Proof of Proposition 4.1

Proof. Using lemma 4.1 we know for some $\forall \alpha \in \mathcal{N}\ \exists \theta^{\rho}_{\alpha} \geq 1$ such that $\forall \theta > \theta^{\rho}_{\alpha}$, $R^{\rho\theta}_{\alpha} > 0$. Since $\mathcal{N}$ is finite, $\theta^{\rho} = \max_{\alpha \in X_{\mathcal{N}}} \theta^{\rho}_{\alpha}$ is well defined. ■

B.2 Polarization and Norms

B.2.1 Proof related to example 4.2

Proof. Consider cost of fragmentation for the members of the possible blocking coalitions that induces partition $\mathcal{P}$:
Case 1: $C_i \in \mathcal{P}$ such that $\forall \alpha \in C_i$, $x_\alpha = 1$:

$$R^{\rho\theta}_{N}(\mathcal{P}) = CE^{\rho\theta}_{N}(\mathcal{P}) - d^{\rho\theta}_{N}(\mathcal{A}) \tag{50}$$

$$= \frac{N - m_i}{N} d_N \left(\left[\frac{1}{1 + \left(\frac{N/2}{N/2 - m_i} \right)^{\frac{\rho}{1-\rho}}} \right] \right)^{\theta} - \left(\frac{1}{2} \right)^{\theta} \tag{51}$$

We know from proof (A.3.2) as $\exists\ \underline{\rho} > 1/2$ such that $\forall \rho > \underline{\rho}$, $R^{\rho\theta}_{N}(\mathcal{P}) < 0$;
and $\exists\ \bar{\rho} < 1/2$ such that $\forall \rho < \bar{\rho}$, $R^{\rho\theta}_{N}(\mathcal{P}) < 0$.

Case 2: $C_i \in \mathcal{P}$ such that $\forall \alpha \in C_i$, $x_\alpha = 0$:

$$R_1^{\rho\theta}(\mathcal{P}) = CE_1^{\rho\theta}(\mathcal{P}) - d_1^{\rho\theta}(\mathcal{A}) \tag{52}$$

$$= \frac{N - m_i}{N} d_1 \left(\left[\frac{1}{1 + \left(\frac{N/2 - m_i}{N/2} \right)^{\frac{\rho}{1-\rho}}} \right] \right)^\theta - \left(\frac{1}{2} \right)^\theta \tag{53}$$

We know from proof (A.3.2) as $\exists\ \bar{\rho} > 1/2$ such that $\forall \rho > \underline{\rho}$, $R_1^{\rho\theta}(\mathcal{P}) < 0$;
and $\exists\ \underline{\rho} < 1/2$ such that $\forall \rho < \bar{\rho}$, $R_1^{\rho\theta}(\mathcal{P}) > 0$.

■

B.2.2 Proof related to example 4.1

Consider cost of fragmentation for the members of the possible blocking coalitions that induces partition $\mathcal{P}$:
Case 1: $C_i \in \mathcal{P}$ such that $\forall \alpha \in C_i$, $x_\alpha = 1$:

$$R_1^{\rho\theta}(\mathcal{P}) = CE_N^{\rho\theta}(\mathcal{P}) - d_1^{\rho\theta}(\mathcal{A}) \tag{54}$$

$$= \frac{N - 1}{N} d_1 \left(\left[\frac{1/2 + \left(\frac{1}{N-2} \right)^{\frac{\rho}{1-\rho}}}{1 + \left(\frac{1}{N-2} \right)^{\frac{\rho}{1-\rho}}} \right] \right)^\theta - \left(\frac{1}{2} \right)^\theta \tag{55}$$

We know from proof (A.3.2) as $\exists\ \underline{\rho} > 1/2$ such that $\forall \rho > \underline{\rho}$, $R_N^{\rho\theta}(\mathcal{P}) < 0$;
and $\exists\ \bar{\rho} < 1/2$ such that $\forall \rho < \bar{\rho}$, $R_N^{\rho\theta}(\mathcal{P}) > 0$.
Case 2: $C_i \in \mathcal{P}$ such that $\forall \alpha \in C_i$, $x_\alpha = 1$:

$$R_N^{\rho\theta}(\mathcal{P}) = CE_N^{\rho\theta}(\mathcal{P}) - d_N^{\rho\theta}(\mathcal{A}) \tag{56}$$

$$= \frac{N - 1}{N} d_N \left(\left[\frac{0.5}{1 + \left(\frac{1}{N-2} \right)^{\frac{\rho}{1-\rho}}} \right] \right)^\theta - \left(\frac{1}{2} \right)^\theta \tag{57}$$

We know from proof (A.3.2) as $\exists\ \underline{\rho} < 1/2$ such that $\forall \rho > \underline{\rho}$, $R_N^{\rho\theta}(\mathcal{P}) < 0$;
and $\exists\ \bar{\rho} > 1/2$ such that $\forall \rho < \bar{\rho}$, $R_N^{\rho\theta}(\mathcal{P}) > 0$.

B.2.3 Proof of Proposition 4.2

This proof uses the weaker definition of stability where we consider deviations by single agents. That is cost of fragmentation considered is

$$R_\beta^{\rho\theta}(\mathcal{A}) = p^{1/\theta} d_\beta^\rho(\mathcal{A}_{-\alpha}) - d_\beta^\rho(\mathcal{A})$$

Proof.

$$R_\beta^{\rho\theta}(\mathcal{A}) = p^{1/\theta} d_\beta^\rho(\mathcal{A}_{-\alpha}) - d_\beta^\rho(\mathcal{A}) \tag{58}$$

Since β is not getting polarized itself (59)

$$\implies \frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial x_\alpha} = p^{1/\theta} \frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial f^\rho(\mathcal{A}_{-\beta})} \frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial x_\alpha} - \frac{\partial d_\beta^\rho(\mathcal{A})}{\partial f^\rho(\mathcal{A})} \frac{\partial f^\rho(\mathcal{A})}{\partial x_\alpha} \tag{60}$$

$$v.\nabla R_\beta^{\rho\theta}(\mathcal{A}) = \sum_{\gamma\in\mathcal{N}} v_\gamma \left(p^{1/\theta} \frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial f^\rho(\mathcal{A}_{-\beta})} \frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial x_\gamma} - \frac{\partial d_\beta^\rho(\mathcal{A})}{\partial f^\rho(\mathcal{A})} \frac{\partial f^\rho(\mathcal{A})}{\partial x_\gamma} \right) \tag{61}$$

$$= p^{1/\theta} \frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial f^\rho(\mathcal{A}_{-\beta})} \sum_{\gamma\in\mathcal{N}} v_\gamma \frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial x_\gamma} - \frac{\partial d_\beta^\rho(\mathcal{A})}{\partial f^\rho(\mathcal{A})} \sum_{\gamma\in\mathcal{N}} v_\gamma \frac{\partial f^\rho(\mathcal{A})}{\partial x_\gamma} \tag{62}$$

Since v is symmetric and $\mathcal{A}$ is symmetric (63)

$$\sum_{\gamma\in\mathcal{N}} v_\gamma \frac{\partial f^\rho(\mathcal{A})}{\partial x_\gamma} = 0 \tag{64}$$

Since v is symmetric, denote the set of polarized individuals as π and denote $pol-, pol+ \in \pi$, where $v_{pol+} = -v_{pol-} > 0$

$$\sum_{\gamma\in\mathcal{N}} v_\gamma \frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial x_\gamma} = \sum_{pol+\in\pi} v_\gamma \left(\frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial x_{pol+}} - \frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial x_{pol-}} \right) \tag{65}$$

$$= \sum_{pol+\in\pi} v_\gamma \left(\frac{(d_{pol+}^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} - (d_{pol-}^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum_{\gamma\in\mathcal{N}_{-\alpha}} (d_\gamma(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}} \right) \tag{66}$$

Case 1: $x_\beta > f^\rho(\mathcal{A}) = 1/2$

$$\frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial x_{pol+}} = \frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial x_{pol-}} < 0$$

and

$$d_{pol+}^\rho(\mathcal{A}_{-\beta}) > d_{pol-}^\rho(\mathcal{A}) \qquad \forall\ pol+ \in \pi \text{ and } \forall\ \rho \in (0,1) \tag{67}$$

Then using expression 66, If $\rho > 1/2$ we get $v.\nabla R_\beta^{\rho,\theta} > 0$

Then using expression 66, if $\rho < 1/2$ we get $v.\nabla R_\beta^{\rho,\theta} < 0$

Case 2: $x_\beta < f^\rho(\mathcal{A}) = 1/2$

$$\frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial x_{pol+}} = \frac{\partial d_\beta^\rho(\mathcal{A}_{-\beta})}{\partial x_{pol-}} > 0$$

$$d^{\rho}_{pol+}(\mathcal{A}_{-\beta}) < d^{\rho}_{pol-}(\mathcal{A}) \qquad \forall\ pol+ \in \pi \text{ and } \forall\ \rho \in (0,1) \tag{68}$$

Then using expression 66, If $\rho > 1/2$ we get $v.\nabla R^{\rho,\theta}_{\beta} > 0$

Then using expression 66, if $\rho < 1/2$ we get $v.\nabla R^{\rho,\theta}_{\beta} < 0$ ■

B.2.4 Proof relating to example 4.3

Proof. We know from lemma 4.2 for **extremists**:

If $\rho < 1/2$ $\frac{\partial R^{\rho\theta}_{\alpha}}{\partial y}(\mathcal{A}) < 0$ and

If $\rho > 1/2$ $\frac{\partial R^{\rho\theta}_{\alpha}}{\partial y}(\mathcal{A}) > 0$. From example(4.2) we know that for maximum polarization ($y = 1$) $\exists \bar{\rho}$ such that $\forall \rho < \bar{\rho}$ $R^{\rho\theta}_{\alpha} < 0$.

Also from example (4.1) we know that for minimum polarization ($y = 1/2$) $\exists \bar{\bar{\rho}}$ such that $\forall \rho < \bar{\bar{\rho}}$, $R^{\rho\theta}_{\alpha} > 0$.

Since R is a continuously decreasing function in y when $\rho < 1/2$, $\exists\ y^{**}$ such that $\forall y < y^{**}$ extremist would prefer to stay in grand coalition.

Similarly,

From example(4.2) we know that for maximum polarization ($y = 1$) $\exists \bar{\rho}$ such that $\forall \rho > \bar{\rho}$ $R^{\rho\theta}_{\alpha} > 0$.

Also from example (**??**) we know that for minimum polarization ($y = 1/2$), $\exists \bar{\bar{\rho}}$ such that $\forall \rho > \bar{\bar{\rho}}$, $R^{\rho\theta}_{\alpha} < 0$.

Since R is a continuously increasing function in y when $\rho > 1/2$, $\exists\ y^{*}$ such that $\forall y > y^{*}$ would prefer to stay in the grand coalition.

For **moderates**:

Let r be the number of polarized individuals, then using result in lemma 3.1

Case 1: For $x_{\beta} = 1 - y$ we have,

$$\frac{\partial f^{\rho}(\mathcal{A}_{-\beta})}{\partial y} = \frac{r(d^{\rho}_{y}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} - (r-1)(d^{\rho}_{1-y}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum\limits_{\gamma \in \mathcal{N}_{-\beta}} (d^{\rho}_{\gamma}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}} \tag{69}$$

Case 1.1 For $\rho > 1/2$ we have,

$$R^{\rho\theta}_{\beta}(\mathcal{A}) = p^{1/\theta} d(\mathcal{A}_{-\beta}) - d(\mathcal{A}) \tag{70}$$

$$\frac{\partial R^{\rho\theta}_{\beta}(\mathcal{A})}{\partial y} = p^{1/\theta}\left(\frac{\partial f^{\rho}}{\partial y} + 1\right) - (0+1) \tag{71}$$

$$> \frac{N-1}{N}\left(\frac{\partial f^{\rho}}{\partial y}\right) - \frac{1}{N} > 0 \qquad \text{if } \rho > 1/2 \tag{72}$$

So β would not want to leave grand coalition for $\rho > 1/2$ as polarization would increase costs.

Case 1.2 For $\rho < 1/2$:

$$\frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial y} = p^{1/\theta}\left(\frac{r(d_y^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} - (r-1)(d_{1-y}^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum_{\gamma\in\mathcal{N}_{-\beta}}(d_\gamma^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}} + 1\right) - 1 \tag{73}$$

$$= p^{1/\theta}\left(\frac{2r(d_y^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d_1^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d_N^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum_{\gamma\in\mathcal{N}_{-\beta}}(d_\gamma^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}\right) - 1 \tag{74}$$

$$\text{Since } \lim_{\rho\to 0}\left(\frac{2r(d_y^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d_1^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d_N^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum_{\gamma\in\mathcal{N}_{-\beta}}(d_\gamma^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}\right) \to 1 \quad \text{monotonically} \tag{75}$$

$$\exists\ \bar{\rho} \text{ such that } \forall \rho < r\bar{ho} \tag{76}$$

$$\frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial y} < 0 \tag{77}$$

Therefore, fixing some $\rho < \bar{\rho}$, given examples (4.1) and (4.2) there must be some y^*_{mod} such that $\forall y < y^*_{mod}$ right leaning moderates will want to leave.

Case 2: For $x_\beta = y$ we have,

$$\frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial y} = \frac{(r-1)(d_y^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} - r(d_{1-y}^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum_{\gamma\in\mathcal{N}_{-\beta}}(d_\gamma^\rho(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}} \tag{78}$$

Case 2.1 For $\rho > 1/2$ we have:

$$R_\beta^{\rho\theta}(\mathcal{A}) = p^{1/\theta} d(\mathcal{A}_{-\beta}) - d(\mathcal{A}) \tag{79}$$

$$\frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial y} = p^{1/\theta}\left(1 - \frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial y}\right) - (1+0) \tag{80}$$

$$> \frac{N-1}{N}\left(-\frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial y}\right) - \frac{1}{N} \tag{81}$$

$$\text{if } \rho > 1/2, \text{ then} \quad \left(-\frac{\partial f^\rho(\mathcal{A}_{-\beta})}{\partial y}\right) > 1/(N-1) \tag{82}$$

$$\implies \frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial y} > 0 \tag{83}$$

So β would not want to leave grand coalition for $\rho > 1/2$ as polarization would increase costs.
Case 2.2 For $\rho < 1/2$:

$$\frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial y} = p^{1/\theta}\left(\frac{r(d^\rho_{1-y}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} - (r-1)(d^\rho_{y}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum\limits_{\gamma\in\mathcal{N}_{-\beta}}(d^\rho_\gamma(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}} + 1\right) - 1 \tag{84}$$

$$= p^{1/\theta}\left(\frac{2r(d^\rho_{1-y}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d^\rho_{1}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d^\rho_{N}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum\limits_{\gamma\in\mathcal{N}_{-\beta}}(d^\rho_\gamma(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}\right) - 1 \tag{85}$$

$$\text{Since } \lim_{\rho\to 0}\left(\frac{2r(d^\rho_{1-y}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d^\rho_{1}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}} + (d^\rho_{N}(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}{\sum\limits_{\gamma\in\mathcal{N}_{-\beta}}(d^\rho_\gamma(\mathcal{A}_{-\beta}))^{\frac{1-2\rho}{\rho}}}\right) \to 1 \quad \text{monotonically} \tag{86}$$

$$\exists\ \bar\rho \text{ such that } \forall \rho < \bar\rho \tag{87}$$

$$\frac{\partial R_\beta^{\rho\theta}(\mathcal{A})}{\partial y} < 0 \tag{88}$$

Therefore, fixing some $\rho < \bar\rho$, given examples (4.1) and (4.2) there must be some y^*_{mod} such that $\forall y < y^*_{mod}$ right leaning moderates will want to leave. It would in fact be the same cut off point as the left leaning moderates as the distribution of preferences is symmetric. ■

B.2.5 Proof of Proposition 4.3

This proof is split in two parts. The first proof only considers internal stability. The second proof considers blocking coalitions

Proof for internal stability of grand coalition. Consider a vector of preferences given by $\mathcal{A}$ of the set of $\mathcal{N}$ agents. Let $\rho = 1/2$. Then by (17) we know for any $C \subset \mathcal{N}$, $f^\rho(\mathcal{A}) = \frac{1}{N}\sum\limits_{\beta\in C} x_\beta$.

We also know that $\mathcal{N}$ will be stable if and only if:

$$r_\alpha^{\rho\theta}(\mathcal{A}) = p\left(\frac{d^\rho_\alpha(\mathcal{A}_{-\alpha})}{d^\rho_\alpha(\mathcal{A})}\right)^\theta \geq 1 \tag{89}$$

$$\implies r_\alpha^{\rho,\theta}(\mathcal{A}) = p\left(\frac{x_\alpha - \frac{1}{N-1}\sum\limits_{\beta\in C\setminus\alpha} x_\beta}{x_\alpha - \frac{1}{N}\sum\limits_{\beta\in C} x_\beta}\right)^\theta \tag{90}$$

$$r_\alpha^{\rho,\theta}(\mathcal{A}) = \frac{p}{((N-1)/N)^\theta} \geq 1 \qquad \text{if } p \geq (N-1)/N \text{ and } \theta \geq 1 \tag{91}$$

■

Proof for stability of grand coalition from blocking. Consider any partition, $\mathcal{P}$, of $\mathcal{N}$ with $p(m_i, \mathcal{P}) = \frac{m_i}{N}$, then:

$$\sum_{C_i \in \mathcal{P}} \frac{m_i}{N} f^{\frac{1}{2}}(A_i) = \sum_{C_i \in \mathcal{P}} \frac{m_i}{N} \frac{\sum\limits_{\alpha \in C_i} x_\alpha}{m_i} = f^{\frac{1}{2}}(\mathcal{N})$$

By concavity of u_α,

$$u_\alpha(f^{\frac{1}{2}}(\mathcal{N})) = u_\alpha\left(\sum_{C_i \in \mathcal{P}} (m_i/N) f^{\frac{1}{2}}(C_i)\right) \geq \sum_{C_i \in \mathcal{P}} \frac{m_i}{N} u_\alpha(f^{\frac{1}{2}}(C_i)) = \mathbb{E}_\alpha^{1/2\theta}$$

■

B.3 Effect of radicalization of Moderates

B.3.1 Proof of Proposition 4.4

Proof. If $\rho > 1/2$ then $\frac{1}{\rho} - 2 < 0$. Therefore, if $\alpha \in A_i$ such that $\alpha = f^\rho(X_\mathcal{N})$, then for $\beta \neq \alpha$:

$$\frac{\partial f^\rho(A_i)}{\partial x_\alpha} = 1 \qquad \text{(Using lemma ??)} \tag{92}$$

$$\implies \frac{\partial R_\beta^{\rho\theta}}{\partial x_\alpha} = \frac{\partial CE_\beta^{\rho\theta}}{\partial x_\alpha} - \frac{\partial d_\alpha^\rho(\mathcal{A})}{\partial f^\rho(\mathcal{A})} \tag{93}$$

$$= p_i \left[\frac{d_\beta^\rho(A_i)}{CE_\beta^{\rho\theta}(\mathcal{P})}\right]^{\theta-1} \left[\frac{\partial d_\beta^\rho(A_i)}{\partial f^\rho(A_i)}\right] \left(\frac{\partial f^\rho(A_i)}{\partial x_\alpha}\right) - \frac{\partial d_\alpha^\rho(A_i)}{\partial f^\rho(\mathcal{A})} \tag{94}$$

$$= \tag{95}$$

If $d_\beta^\rho(A_i) \leq CE_\beta^{\rho\theta}$ then $\left[\frac{d_\beta^\rho(A_i)}{CE_\beta^{\rho\theta}(\mathcal{P})}\right]^{\theta-1} \in [0,1] \implies$

For $\beta = 0$: $\frac{\partial d_\beta^\rho(A_i)}{\partial f^\rho(A_i)}, \frac{\partial d_\alpha^\rho(A_i)}{\partial f^\rho(A_\mathcal{N})} > 0, \implies \frac{\partial R_\beta^{\rho\theta}(\mathcal{P})}{\partial x_\alpha} < 0.$

For $\beta = 1$: $\frac{\partial d_\beta^\rho(A_i)}{\partial f^\rho(A_i)}, \frac{\partial d_\alpha^\rho(A_i)}{\partial f^\rho(A_\mathcal{N})} > 0 \implies \frac{\partial R_\beta^{\rho\theta}(\mathcal{P})}{\partial x_\alpha} > 0.$

If $d_\beta^\rho(A_i) \geq CE_\beta^{\rho\theta}$ then

$$p_i\left[\frac{d_\beta^\rho(A_i)}{CE_\beta^{\rho\theta}}\right]^{\theta-1} = p_i\left[\frac{d_\beta^\rho(A_i)}{d_\beta^\rho(A_i)\left(p_i + \sum_{C_j\in\mathcal{P}} p_j\left(\frac{d_\beta^\rho(A_j)}{d_\beta^\rho(A_i)}\right)^\theta\right)^{\frac{1}{\theta}}}\right]^{\theta-1} \tag{96}$$

$$= \frac{p_i\left(p_i + \sum_{C_j\in\mathcal{P}} p_j\left(\frac{d_\beta^\rho(A_j)}{d_\beta^\rho(A_i)}\right)\right)}{p_i + \left(\sum_{C_j\in\mathcal{P}} p_j\left(\frac{d_\beta^\rho(A_j)}{d_\beta^\rho(A_i)}\right)\right)^\theta} \tag{97}$$

$$= \frac{p_i}{p_i + \left(\sum_{C_j\in\mathcal{P}} p_j\left(\frac{d_\beta^\rho(A_j)}{d_\beta^\rho(A_i)}\right)\right)^\theta}\left(\frac{\left(\sum_{C_j\in\mathcal{P}} p_j\left(d_\beta^\rho(A_j)\right)^\theta\right)^{\frac{1}{\theta}}}{d_\alpha^\rho(A_i)}\right) \tag{98}$$

$$= \frac{p_i}{p_i + \left(\sum_{C_j\in\mathcal{P}} p_j\left(\frac{d_\beta^\rho(A_j)}{d_\beta^\rho(A_i)}\right)\right)^\theta}\left(\frac{CE_\alpha^{\rho\theta}(\mathcal{P})}{d_\alpha^\rho(A_i)}\right) < 1 \tag{99}$$

From lemma (3.1) we have $\frac{\partial f^\rho(A_i)}{\partial x_\alpha} < 1 \implies p_i\left[\frac{d_\beta^\rho(A_i)}{CE_\beta^{\rho\theta}(\mathcal{P})}\right]^{\theta-1}\left(\frac{\partial f^\rho(X_\mathcal{N})}{\partial\alpha}\right) \in [0,1)$. Thus,

For $\beta = 0$: $\frac{\partial d_\beta^\rho(A_i)}{\partial f^\rho(A_i)}, \frac{\partial d_\alpha^\rho(A_i)}{\partial f^\rho(A_\mathcal{N})} > 0, \implies \frac{\partial R_\beta^{\rho\theta}(\mathcal{P})}{\partial x_\alpha} < 0.$

For $\beta = 1$: $\frac{\partial d_\beta^\rho(A_i)}{\partial f^\rho(A_i)}, \frac{\partial d_\alpha^\rho(A_i)}{\partial f^\rho(A_\mathcal{N})} > 0 \implies \frac{\partial R_\beta^{\rho\theta}(\mathcal{P})}{\partial x_\alpha} > 0.$ ■

B.3.2 Proof Proposition (4.5)

Proof. If $\rho < 1/2$ then $\frac{1}{\rho} - 2 > 0$. Therefore, if $\alpha \in C_i$ such that $x_\alpha = f^\rho(X_\mathcal{N})$, then for $\beta \neq \alpha$:

$$\frac{\partial f^\rho(A_i)}{\partial x_\alpha} = 0 \qquad \text{(Using lemma 3.1)} \tag{100}$$

$$\implies \frac{\partial R_\beta^{\rho\theta}}{\partial x_\alpha} = \frac{\partial CE_\beta^{\rho\theta}}{\partial x_\alpha} \tag{101}$$

$$= p_i\left[\frac{d_\beta^\rho(A_i)}{CE_\beta^{\rho\theta}}\right]^{\theta-1}\left[\frac{\partial d_\beta^\rho(A_i)}{\partial f^\rho(A_i)}\right]\left(\frac{\partial f^\rho(A_i)}{\partial x_\alpha}\right) \tag{102}$$

If $\beta > f^{\rho}(A_i)$ then $\left[\frac{\partial d_{\beta}^{\rho}(A_i)}{\partial f^{\rho}(A_i)}\right] = -1$ which implies $\frac{\partial R_{\beta}^{\rho\theta}}{\partial \alpha} < 0$.

If $\beta < f^{\rho}(A_i)$ then $\left[\frac{\partial d_{\beta}^{\rho}(A_i)}{\partial f^{\rho}(A_i)}\right] = 1$ which implies $\frac{\partial R_{\beta}^{\rho\theta}}{\partial \alpha} > 0$. ■

B.3.3 Expression of Influence function in terms of norm of compromise and distribution of preference

Consider a coalition $C_i \subset \mathcal{A}$. Define the Influence function: $I^{\rho}(\mathcal{A}, \mathcal{A}_{-i}) = |f^{\rho}(\mathcal{A}) - f^{\rho}(\mathcal{A}_{-i})|$ where $\mathcal{A}_{-i}$ is the vector of ideal points of $\mathcal{N}\backslash C_i$. This section shows how the influence function, a function that captures the risk of leaving the grand coalition, depends on ρ.

Consider now an arbitrary coalition $C_j\mathcal{N}$ and the distribution of preferences, A_j. From the optimization problem that defines norms in 1.

$$\sum_{\{x_\alpha < f^{\rho}(A_j):\ \alpha \in C_j\}} \left(d_{\alpha}^{\rho}(A_j)\right)^{\frac{1}{\rho}-1} - \sum_{\{x_\alpha > f^{\rho}(A_j):\ \alpha \in C_j\}} \left(d_{\alpha}^{\rho}(A_j)\right)^{\frac{1}{\rho}-1} = 0 \tag{103}$$

Define

$$H(A_j, z) \equiv \sum_{\{x_\alpha < z:\ \alpha \in C_i\}} \left(d_{\alpha}(z)\right)^{\frac{1}{\rho}-1} - \sum_{\{x_\alpha > z:\ \alpha \in C_i\}} \left(d_{\alpha}(z)\right)^{\frac{1}{\rho}-1} \tag{104}$$

Thus,

$$H_z(A_j, z) = \left(\frac{1-\rho}{\rho}\right) \sum_{\{\alpha \in C_i\}} \left(d_{\alpha}(z)\right)^{\frac{1}{\rho}-2} \tag{105}$$

Since $H(A_j, .)$ is differentiable in z we can use Taylor's approximation.

$$H(A_j, z_0 + h) = H(A_j, z_0) + hH_z(A_j, z_0 + \lambda h) \tag{106}$$

$$\implies h = \frac{H(A_j, z_0 + h) - H(A_j, z_0)}{H_z(A_j, z_0 + \lambda h)} \tag{107}$$

$$\implies |h| \approx \left|\frac{H(A_j, z_0 + h) - H(A_j, z_0)}{H_z(A_j, z_0)}\right| \tag{108}$$

Let $z_0 = f(\mathcal{A}_{-i})$ with $h = f^{\rho}(\mathcal{A}) - f^{\rho}(\mathcal{A}_{-i})$ (109)

and $A_j = \mathcal{A}_{-i}$ where $\mathcal{A}_{-i}$ is the distribution of ideal points of $\mathcal{N}\backslash C_i$ (110)

$$|f(\mathcal{A}) - f(\mathcal{A}_{-i})| = \left|\frac{H(\mathcal{A}_{-i}, f^{\rho}(\mathcal{N})}{H_z(\mathcal{A}_{-i}, f^{\rho}(\mathcal{N}\backslash C_i))}\right| \tag{111}$$

(112)

$$\implies I^{\rho}(\mathcal{A}, \mathcal{A}_{-i}) \approx \frac{\rho}{1-\rho} \left| \frac{\sum_{\{x_\alpha < f^{\rho}(\mathcal{A}):\ \alpha \in \mathcal{N}\backslash C_i\}} \left(d_{\alpha}^{\rho}(\mathcal{A})\right)^{\frac{1}{\rho}-1} - \sum_{\{x_\alpha > f^{\rho}(\mathcal{A}):\ \alpha \in \mathcal{N}\backslash C_i\}} \left(d_{\alpha}^{\rho}(\mathcal{A})\right)^{\frac{1}{\rho}-1}}{\sum_{\{\alpha \in \mathcal{N}\backslash C_i\}} \left(d_{\alpha}^{\rho}(\mathcal{A}_{-i})\right)^{\frac{1}{\rho}-2}} \right| \tag{113}$$

where $d_\alpha^\rho(A) \equiv d_\alpha(f^\rho(A)) \equiv |x_\alpha - f^\rho(A)|$ for an arbitrary distribution of preferences of agents within a coalition given by A.