UNIVERSITY OF CALIFORNIA SAN DIEGO

Power, Sabotage, and Misdirection: Three Essays on Political Economy

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Frederick Aram Papazyan

Committee in charge:

Professor T. Renee Bowen, Chair
Professor Daron Acemoglu
Professor J. Lawrence Broz
Professor Alexis Akira Toda
Professor Joel Watson
Professor Johannes Wieland

2023

The dissertation of Frederick Aram Papazyan is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my parents, Denise and Arno,

my brothers, Nicky and Dennis,

and my cat, Jiji.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Words fail to express how grateful I am to my advisor, Renee Bowen, for her unshakeable belief in me and for always encouraging me to strive for excellence. I admire the dedication and standards she has for her students. Her guidance has made me into a better researcher and communicator and I am honored to call myself her student.

I am also deeply grateful to Daron Acemoglu for taking a chance on me. Giving me the opportunity to come to MIT to assist with *The Narrow Corridor* was a life-changing experience that helped me grow as a thinker. This experience and his work inspired my job market paper and the future work I hope to accomplish. He sparked my interest in political economy — an interest that Renee indelibly solidified when I returned to UC San Diego.

I could not ask for a better committee. I am thankful to Alexis Akira Toda and Joel Watson for their keen mathematical eyes and their high standards for rigor. I am also grateful to Johannes Wieland and Lawrence Broz for always keeping my work grounded and helping me see the bigger picture.

I would also like to thank Denis Shishkin, Aram Grigoryan, Songzi Du, Simone Galperti, Joel Sobel, David Lagakos, Emanuel Vespa, and Federica Izzo for all of the kindness they have shown me over the years and for the many insightful conversations I have had with each of them. I am deeply proud to have been a part of the Department of Economics at UC San Diego. This department is living proof that ambition and kindness are not opposing characteristics, and it should serve as an example for other departments to emulate.

I cherish the many friendships I have made over the years at this department. I am deeply thankful to Ellen Liaw, Adrian Wolanski, and Danil Dmitriev. I do not know what I would do without you three, and I do not know what Regents Pizzeria will do

without us four. I would like to also thank Stefan Faridani, Haitian Xie, Jin Xi, Wei-Lin Chen, Jin Han, Yuli Xu, Giampaolo Bonomi, Erica Chuang, Kelvin Leong, Jay Cizeski, Alex Garland, Connor Redpath, Rohini Ray, Xintong Li, and many, many others for their friendship, kindness, and the laughs we have shared. Finally, I would like to give a special thanks to Danil for all of the support and the countless hours of his time he has given me these past six years. I am so excited to see the amazing research projects we will pursue in the future!

I would also like to express my gratitude to the friends who came into my life prior to my doctoral studies. I would like to thank Michael Baltz, Chris Bohdjelian, Aram Zohrabian, Kaja Jankowska, Anna Anisimova, Anna-Marie Lichtenbergova, Georgia Soares, Sarah and Roxane Peloux, and Claire Hyunjee Sung for their unwavering friendship, even during the times when I'm lost down a research rabbit hole.

I am thankful to Ken Alexander, Sergey Lototsky, Neel Tiruviluamala, and David Crombecque at the University of Southern California's Department of Mathematics. I am grateful for their kindness, support, and for the mathematical toolset they helped me hone during my Applied Mathematics Master's and as an undergraduate. I would also like to express my gratitude to Yilmaz Koçer at USC's Department of Economics for sparking my interest in economics at the beginning of my academic journey.

Finally, I would like to express my deepest gratitude to my family. I would not have gotten through this journey without the love and support of my parents, Arno Papazyan and Denise Cizmeciyan, my brothers, Nicky and Dennis, and of course my cat, Jiji.

Chapter 1, "Power Consolidation in Groups," is currently being prepared for submission for publication of the full material therein. The dissertation author, Frederick Aram Papazyan, is the sole author of this chapter.

Chapter 2, "Sabotage-Proof Mechanism Design," is currently planned for submis-

sion for publication of the material therein. Danil Dmitriev and the dissertation author, Frederick Aram Papazyan, are co-authors of this chapter.

Chapter 3, "Strategic Misdirection," is currently planned for submission for publication of the material therein. The dissertation author, Frederick Aram Papazyan, is the sole author of this chapter.

VITA

2012	Bachelor of Science, University of Southern California
2017	Master of Arts, University of Southern California
2018	Research Assistant, Massachusetts Institute of Technology
2018-2023	Teaching Assistant, University of California San Diego
2023	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Power, Sabotage, and Misdirection: Three Essays on Political Economy

by

Frederick Aram Papazyan

Doctor of Philosophy in Economics

University of California San Diego, 2023

Professor T. Renee Bowen, Chair

This dissertation is a collection of three essays on political economy. In the first chapter, I develop an economic theory of how a society's distribution of power and resources evolves over time. Multiple lineages of players compete by accumulating *power*, which is modeled as an asset that increases the probability of winning conflicts over resources. Given any initial distribution of power, this model provides a unique equilibrium prediction of how it will evolve over time. Three types of stable distributions are approached in the long run: inclusive, oligarchic, and dictatorial, where power is uniformly distributed among all players, a few players, or held by just one player, respectively. I show that power and resources inevitably fall into the hands of a few when

political competition is left unchecked in large societies. This addresses a longstanding empirical puzzle, and I also provide policy implications for keeping inclusivity stable in large societies.

In the second chapter, my co-author, Danil Dmitriev, and I consider the problem of designing a voting mechanism that is robust to derailment by external groups. We show that plurality voting and other standard mechanisms are often not robust to sabotage; in fact it is sometimes preferable to not run any poll at all. The optimal voting mechanism is found to make saboteurs indifferent between each alternative they can vote for, since this undermines their ability to adversely affect the designer's predictions of other voters' preferences.

In the third chapter, I study how a sender can use verifiable binary evidence to influence a receiver about a binary state when the relevance of information is ex ante uncertain and asymmetrically known by the sender. The sender has access to two pieces of evidence: one they know to be perfectly informative of the state and one that is completely uninformative. I show that while full disclosure of evidence is possible in equilibrium, the receiver to fully unravel which piece of evidence is relevant. Consequently, the Receiver may gain little to no information about the state even when all evidence is disclosed.

Chapter 1

Power Consolidation in Groups

1.1 Introduction

As inequality continues to rise in the United States, so have concerns that it may be drifting towards *oligarchy*.[1] This trend is not exceptionally American: persistently rising political and economic inequality has been observed in several other nations in the OECD (2008, 2011, 2012, 2015, 2021) alongside worldwide trends of spreading authoritarian rule (Freedom House, 2022) and democratic backsliding (Repucci (2020), Hyde (2020)). Understanding how the distributions of political *power* and economic *resources* in a society evolve over time has become increasingly important, and since these distributions are fundamentally linked, their dynamics can only be understood when studied in tandem.[2]

What allows – or indeed prevents – power and resources from falling into the hands of a few in a society? Prevailing explanations overwhelmingly rely on *structural factors*, qualitative features of societies such as culture, geography, economic condi-

[1] Krugman (2014a), Piketty in Krugman (2014b), Saez and Zucman (2019), and Gilens and Page (2014).

[2] Piketty (1995, 2013, 2015, 2018, 2019), Stiglitz (2011, 2016), Rausser et al. (2011), Krugman (2020), World Bank (2005, 2017), United Nations (2020), and Callander et al. (2021), discussed in the literature review.

tions, exposure to external threats, etc. As Acemoglu and Robinson (2022b) point out, such explanations necessarily cannot account for how otherwise similar societies can arrive at vastly different power structures, which is a widespread occurrence (Acemoglu and Robinson, 2019).

To illustrate, consider China, Taiwan, and Pre-2020 Hong Kong: many argue that China has an intrinsic tendency towards authoritarian regimes – as opposed to more egalitarian ones – that stems from Chinese culture, especially its Confucian heritage.[3] However, these explanations are starkly at odds with the fact that "Hong Kong and Taiwan are cut from the same cultural cloth as mainland China, yet they rest on very different political systems" (Acemoglu and Robinson, 2022a).

This chapter constructs a theory of how a society's distribution of power evolves due to intergenerational competition over resources. I consider a society that is populated by (non-overlapping generations of) players from multiple lineages. Each lineage is initially endowed with a stock of power, which is modeled as an asset that increases one's probability of winning conflicts over consumable resources. Every period, players inherit and accumulate power, and then engage in conflicts over resources.[4]

In this model, society's distribution of power endogenously evolves due to the individual, strategic power accumulation decisions players make in the course of this intergenerational power accumulation contest. In the absence of shocks, players' initial distribution of power *uniquely* determines its equilibrium trajectory and long-run behavior. Three types of stable distributions emerge in the long run: *inclusive* (where power is uniformly distributed), *dictatorial* (where only one player holds a strictly positive amount of power), and *oligarchic* (where power is uniformly distributed among

[3] Huntington (1991, 1996), Dalio (2021), Qing (2013), and several others discussed in Spina et al. (2011).

[4] This standard contest theoretic approach reflects Max Weber's (1925) widely-adopted definition of power as "the probability that one actor within a social relationship will be in a position to carry out his own will despite resistance, regardless of the basis on which this probability rests," and the fact that it must be *accumulated* by, for instance, "expenditures of time and money on campaign contributions, political advertising, and other ways that exert political pressure" (Becker, 1983).

more than two – but fewer than all – players, with the rest being powerless).

In addition to providing sharp, rich equilibrium predictions, this chapter also generates novel insights and testable implications. Its main result (Proposition 4) is that inclusive power structures are *never* stable in sufficiently large societies when political competition is left unchecked.[5] In contrast, dictatorships and sufficiently concentrated oligarchies can remain stable in *arbitrarily large* groups (Proposition 5). This provides a novel theoretical explanation for why large groups appear to be more vulnerable to power consolidation, which is currently considered a long-standing empirical puzzle. Michels (1915) notably stated this as the *Iron Law of Oligarchy*, and over a century later, the tendency of power consolidation to take place in large groups of people appears to be widely accepted as a stylized fact, but there appears to be little agreement regarding its explanation (Leach (2005, 2015); Diefenbach (2019)). This model delivers insights not only on the competitive forces that underlie this tendency, but also on the policies that can counteract it. Finally, this chapter provides an additional result (Proposition 6) which characterizes how larger populations induce stronger dictatorships.

The framework I construct in this chapter builds on Acemoglu and Robinson (2022b), who model how the balance-of-power between *two* players – one representing elites, the other representing non-elites – evolves over time.[6] This chapter primarily generalizes and re-frames the analysis to societies made up of any finite number of players, who may be viewed as *individual* agents or as *representative* agents of a socioeconomic sub-group.[7] While this substantially expanded scope is valuable, it more importantly gives this framework the ability to generate previously unattainable insights

[5] I.e. when one does not intervene in the aforementioned power accumulation contest by reallocating power or policies that affect players' incentives in this contest (discussed in subsection 1.4.1).

[6] This in turn is related to Acemoglu (2005) – which provides much of its microfoundation – and in Acemoglu and Robinson (2019) the authors provide an extensive view of history through the lens their model.

[7] Becker (1983) notes how "groups – defined by occupation, industry, income, geography, age, and other characteristics – that ... use political influence to enhance the well-being of their members."

into why power consolidation appears to become increasingly likely as societies grow *large*. The rest of the (extensive) related literature is now reviewed.

Literature Review

Foundational writings on the emergence and persistence of elites and oligarchies include Michels (1915), Mosca (1939), Mills (1956), Pareto (1935, 1991), and Bottomore (1964), which are critiqued in Dahl (1958), Rustow (1966), Cammack (1990), and Ober (2008). The emergence of oligarchies in large groups of people was viewed as an inevitability by Michels (1915), stating it as the Iron Law of Oligarchy. He primarily focuses on the role played by bureaucratization, starting from the premise that large groups require bureaucracy in order to effectively organize and coordinate actions, and then arguing that bureaucracies naturally lead to hierarchies. Leach (2005) notes that "[d]espite almost a century of scholarly debate ... there is still no consensus about whether and under what conditions Michels's claim holds true." This assessment appears to be supported by the thorough reviews of the modern literature provided by Leach (2005, 2015) and Diefenbach (2019).

In a recent seminal work, Winters (2011) provides an extensive study of how oligarchies emerge and persist in a variety of societies around the world, where he also notes that a "consistent pattern in human history is for very small minorities to amass great wealth and power." Rather than focusing on how particular institutional structures allow or preclude the formation of oligarchies, he instead argues that wealth defence and the accumulation of *material power* – a notion of power that is conceptually similar to what I model here – are far more important factors in the formation of oligarchies. Moreover, he emphasizes how oligarchies can emerge even in the presence of democratic norms and institutions, and the possibility of having democracies only in name. This is also emphasized in Winters and Page (2009), an empirical study on the

distribution of material power in the United States.

This framework draws on the long-standing *contest theory* approach to modeling power and conflict. As Hirshleifer (1989, 1991a, 1991b) discusses, both *military* conflicts (e.g. Lanchester (1916), Ewerhart (2021)) and *political* conflicts[8] can be modeled as a contest, which "is a game in which players compete for a prize by exerting effort so as to increase their probability of winning" Skaperdas (1996). The "prize" in this model is control over resources (a consumable good such as natural resources, public funds, etc.), which is seized by the victor of conflicts. Players accumulate power ("effort") to increase their probability of winning conflicts.

A Contest Success Function (CSF) defines the conditional probability that a player wins the conflict given the amount of power they hold and the amount held by each of their opponents. This chapter focuses on the commonly used difference-form CSF which, as its name suggests, only depends on power differences. The properties of this form were notably discussed in Hirshleifer (1989), who discusses how it relaxes certain overly-idealized aspects of its counterpart, the ratio-form/Tullock CSF (Tullock, 1980); both forms are axiomized in Skaperdas (1996).

In the context of this framework, the most important property of the difference-form CSF is that each player's marginal benefit of accumulating power is increasing in how closely matched they are with their opponents. As mentioned above, power is modeled as an *asset* that is accumulated at a *cost*. I assume that the marginal cost of accumulating power (at any fixed rate) is diminishing in the amount of power one currently holds. Intuitively, this captures the notion that the more powerful one already is, the less costly it is to obtain more of it.[9] Since this chapter focuses on the competitive forces that underlie how players' distribution of power evolves, its results largely boil

[8] For example, Becker's (1983, 1985) models of interest group politics and Tullock's (1967, 1980) seminal works on rent seeking

[9] This assumption appears to be supported by the observations in Pierson (2000), Maxwell and Oliver (1993), Francois (2002), Desai and Olofsgård (2011), and Drutman (2015).

down to these two natural features of power accumulation incentives.

As will be seen below, the equilibrium dynamics of this model feature what is known as the *discouragement effect*. This phenomenon was notably observed in the Harris and Vickers (1987) model of patent races – and in a variety of other dynamic contests (Konrad, 2012) – as well as in Aghion et al. (2005) and Aghion (2005) which study innovation investments. Experimental evidence of the discouragement effect is reviewed in Dechenaux et al. (2015). Indirectly related to this work are Berry (1993), Clark and Riis (1996), and Chowdhury and Kim (2014) who study multi-winner contests. While the model herein explicitly uses a *single*-winner contest mechanism, one may interpret there being multiple "effective" winners at the oligarchic and inclusive power structures in this model.

The form of power considered here has parallels with the notion of *personal power*, which was first systematically studied by Bowen et al. (2022). Their notion is similar in that it increases the probability of actualizing one's ideal outcome by asserting one's will. However, it is qualitatively different in that personal power derives from one's personal characteristics, and its effectiveness must be learned by others. This chapter instead focuses on the forms of power that are accumulated and inherited.

Jeon and Hwang (2020) resembles the present chapter in motivation but not in approach. In contrast to the dynamic *contest* setup considered here, they work in a dynamic *bargaining* framework. Their model admits two classes of stable power structures that resemble the dictatorial and oligarchic power structures seen here. Another key difference is that Jeon and Hwang (2020) consider infinitely forward-looking agents while the agents considered here are short-lived, being replaced each period. In their model, dictatorial power structures are unstable given that agents are sufficiently forward looking. Interestingly, while my model does admit a dictatorial class of stable power structures, it also admits inclusive and oligarchic classes despite agents' short

lifespan. Hence, there also exist qualitative differences in terms of results. Similarly, Acemoglu and Robinson (2006) also provide a theoretical explanation for the Iron Law of Oligarchy (Michels, 1915), but through a different mechanism than the one seen here. In their model, elites entrench themselves by exploiting the structure of institutions in which they operate, while the results in this model do not rely on the explicit structure of institutions.

This chapter is also related to Bowen and Zahran (2012). In the present model, dictatorial and oligarchic power structures are reached by trajectories that originate sufficiently nearby. These classes respectively have qualitative similarities to the compromise and no-compromise classes in Bowen and Zahran (2012), which are reached in an analogous fashion.

As mentioned above, several economists and political scientists have stressed the intimate link between the studies of political inequality and economic inequality, and of their respective dynamics. Since power and wealth are intimately linked (Winters and Page (2009), Winters (2011), Page et al. (2013) Gilens and Page (2014)) so are the *dynamics* of their respective distributions, a phenomenon (Stiglitz, 2011) aptly summarized as "[w]ealth begets power, which begets more wealth."

The main conclusion of Piketty's groundbreaking 2013 work, *Capital in the Twenty-First Century*, was that "[t]he history of the distribution of wealth has always been deeply political, and it cannot be reduced to purely economic mechanisms... [i]t is shaped by the way economic, social, and political actors view what is just and what is not, as well as by the relative power of those actors and the collective choices that result." In Piketty's 2015 discussion of this work, he stresses the importance of "putting the distribution back at the center of economics" and "the role of political conflict in relation to inequality," noting how "[i]nstitutional changes and political shocks ... can be viewed as largely endogenous to the inequality and development process itself."

In parallel fashion, Stiglitz (2016) discusses why standard economic theory cannot currently explain the observed divergence in income inequality among countries with similar production and technological capacities, where he urges that more focus must be placed on the role played by rent seeking, political institutions, and power relations. Krugman (2020), Callander et al. (2021), the United Nations (2020), and the World Bank (2005, 2017) have expressed similar sentiments as the above.

The rest of this chapter is organized as follows. Section 1.2 constructs the model, states its assumptions, and defines equilibrium. Section 1.3 then characterizes the equilibrium dynamics of power structures and the stable power structures that arise in the long run. Properties of these stable power structures are characterized in section 1.4, which is followed by the conclusion in section 1.5. All proofs of the results in the main text of this chapter are found in Appendix A.1.1. Appendix A.1.2 contains auxiliary results (with their respective proofs). Appendix A.1.3 contains supplementary figures, and Appendix A.1.4 contains information on Contest Success Functions.

1.2 Model

Time has an infinite horizon and is initially[10] taken to be discrete with period-length, $\Delta > 0$ ($t \in \{0, \Delta, 2\Delta, ...\}$). There are $N \geq 2$ lineages of risk neutral, short-lived players that are replaced each period; lineage $i \in \{1, ..., N\}$ is formally defined as

$$\mathbb{i} \equiv \{i_0, i_\Delta, i_{2\Delta}, ...\},$$

where i_t denotes the generation-t player from lineage i.

Players compete by accumulating and passing along stocks of *power*, an asset

[10] Period length is later made arbitrarily small when attention is brought to model dynamics, which are more tractably characterized in continuous time.

that increases one's probability of winning conflicts over resources (in a way made precise below). The amount of power held by the lineage-i player at time t is given by $x_{it} \in [0, \chi]$, where $\chi > 0$ is arbitrarily fixed.[11] The publicly-observable state vector $\boldsymbol{x}_t := (x_{1t}, ..., x_{Nt})$ corresponds to the group's *power structure* at time t and will be the central object of analysis in this chapter.

Lineage i is initially endowed with x_{i0} units of power; this is held by player i_0, who is assumed to remain inactive for the entirety of period 0 and simply serves to initialize the game. Play then proceeds as follows: at the beginning of period $t \in \{\Delta, 2\Delta, ...\}$, player i_t inherits their predecessor's power $x_{i,t-\Delta}$ which linearly depreciates at *rate*, $\delta > 0$ (hence by *amount* $\delta\Delta$ each period). Players then simultaneously choose how much to invest in their own power. Formally, player i_t commits to accumulating power at *rate*, $I_{it} \geq 0$ throughout the period, which adds $I_{it}\Delta$ units of power to i_t's stock by the end of the period. The instantaneous flow cost of investing at rate I_{it} when starting at $x_{i,t-\Delta}$ is given by $C(I_{it}, x_{i,t-\Delta})$.

Afterward, society endows a lump-sum unit of resources (a consumable good). Players compete over these resources through a *winner-takes-all* conflict whose victor is randomly chosen according to the (conditional) probability distribution,[12]

$$H(x_{it}, \boldsymbol{x}_{-i,t}; N) \equiv \mathbb{P}\Big\{\text{Player } i_t \text{ wins the conflict} \,\Big|\, \text{Power structure is } \boldsymbol{x}_t\Big\}. \tag{1.1}$$

That is, each players' probability of victory depends not only on how much power they hold (x_{it}), but also that held by others ($\boldsymbol{x}_{-i,t}$). At the end of period t, player i_t earns an

[11] Assuming that power takes values in $[0, \chi]$ simplifies exposition, but does not qualitatively affect the results of this model. χ is made arbitrarily large in Section 1.4.1. This operation only plays a significant role in the latter two parts of Proposition 5. All other results – including the main result (Proposition 4) – remain qualitatively unchanged. It also allows this model to comment on the implications of decisions such as *Citizens United v. FEC (2010)*, as I discuss in Section 1.4.

[12] Assuming that players *necessarily* engage in conflict – and that the conflict is *winner-takes-all* – do not affect the results of this chapter.

expected "lifetime" net payoff of

$$\pi_i\left(x_{it}, I_{it}, \boldsymbol{x}_{-i,t}, x_{i,t-\Delta}\right) \equiv H(x_{it}, \boldsymbol{x}_{-i,t}; N) - \Delta \cdot C(I_{it}, x_{i,t-\Delta}), \tag{1.2}$$

where $C(\cdot,\cdot)$ is weighted by period length – while $H(\cdot,\cdot;\cdot)$ is not – since the former is a *flow* cost while the latter is a *lump sum* benefit.

The following assumption is made on the *cost* and *marginal cost* $D_1C(I_{it}, x_{i,t-\Delta}) \equiv \frac{\partial}{\partial I_{it}}C(I_{it}, x_{i,t-\Delta})$ of power accumulation.

Assumption 1. $C : [0,\infty)^2 \to [0,\infty)$ satisfies

1. Cost $C(\cdot, x_{i,t-\Delta})$ and marginal cost $D_1C(\cdot, x_{i,t-\Delta})$ are strictly increasing $\forall x_{i,t-\Delta}$.
2. $C(I_{it},\cdot)$ and $D_1C(I_{it},\cdot)$ are decreasing $\forall I_{it} \geq 0$.
3. D_1C is continuously differentiable in its first argument and continuous in its second argument.

Assumption 1.1 (increasing, convex power accumulation costs) is standard. More important is Assumption 1.2, which states that the cost – and marginal cost – of power accumulation diminish with how much power one currently holds. This captures the idea that the more powerful one already is, the less costly it is to further accumulate power, both in absolute terms and on the margin. The notion that there are upfront costs to power accumulation is intuitive, and also appears to be supported by real world observations (see Footnote 9 in the Literature Review). Finally, Assumption 1.3 is a mild smoothness assumption.[13]

[13] Assumption 1.3 is essentially a relaxation of the assumption that C is twice continuously differentiable: it permits $D_1C(I,\cdot)$ to have "kinks" ($\forall I \geq 0$).

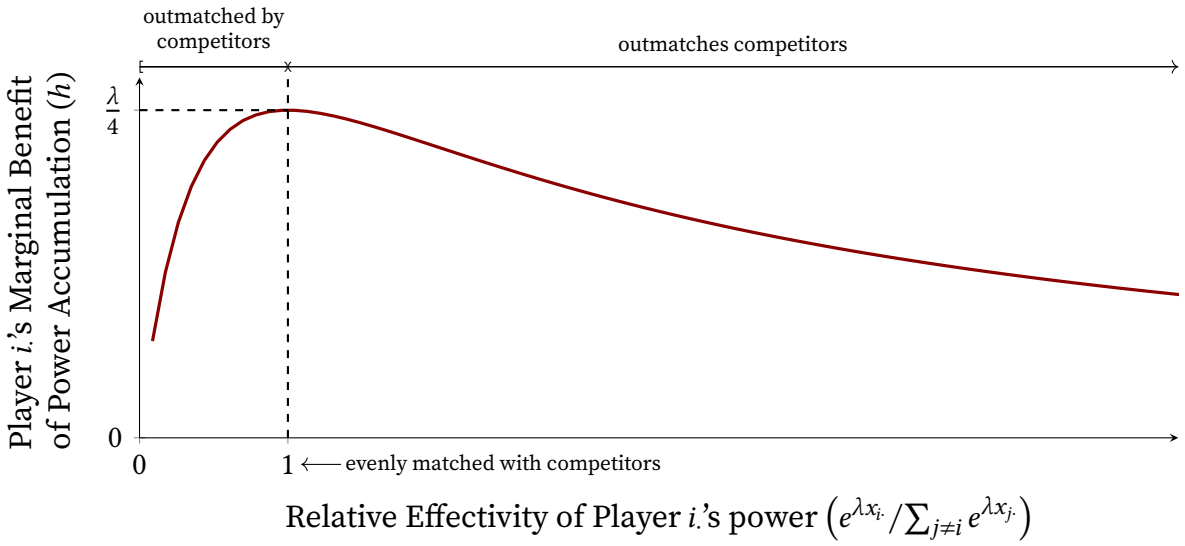


Figure 1.1: Player i.'s marginal benefit of power accumulation h as a function the relative effectivity $e^{\lambda x_{i\cdot}}/\sum_{j\neq i} e^{\lambda x_{j\cdot}}$ of their power.

The second assumption made in this chapter concerns the *benefit* H of holding power. I assume it takes a standard form that is commonly referred to as the difference-form or logistic-form contest success function, whose properties were notably studied by Hirshleifer (1989).

Assumption 2. Given $\boldsymbol{x}$, the lineage-i player wins the conflict with probability

$$H(x_i, \boldsymbol{x}_{-i}; N) \equiv \frac{e^{\lambda x_i}}{\sum_{j=1}^{N} e^{\lambda x_j}} = \frac{1}{1+\sum_{j\neq i} e^{-\lambda(x_i - x_j)}}, \quad (\lambda > 0). \tag{1.3}$$

Beyond assuming that H is continuous and only directly depends on power *differences*, assuming the above functional form is equivalent to assuming that it satisfies a collection of innocuous to mild axioms (Skaperdas, 1996, Theorem 3).[14] As Corchón and Dahm (2010) note, it is standard to interpret $e^{\lambda x_i}$ as the *effectivity* of player i's power,

[14]For readers' convenience, these axioms and Skaperdas's Theorem are summarized in Appendix A.1.4.

which corresponds to how *effectively* player i's power influences their victory probability.

The main implication of this assumption is that the marginal benefit of power accumulation ("contest incentives")

$$h(x_i, \boldsymbol{x}_{-i}; N) \equiv \frac{\partial}{\partial x_i} H(x_i, \boldsymbol{x}_{-i}) = \frac{\lambda \sum_{j \neq i} e^{-\lambda(x_i - x_j)}}{\left[1 + \sum_{j \neq i} e^{-\lambda(x_i - x_j)}\right]^2} = \frac{\lambda e^{\lambda x_i} \Big/ \sum_{j \neq i} e^{\lambda x_j}}{\left(1 + e^{\lambda x_i} \Big/ \sum_{j \neq i} e^{\lambda x_j}\right)^2} \tag{1.4}$$

is *increasing* in how *closely-matched* one is with other players in terms of power. This captures the idea that gains over a closely-matched opponent are more valuable than those made against a much weaker (or much stronger) one. This property is formalized in the final equality of (1.4): the closer the *relative effectivity* $e^{\lambda x_{i\cdot}} \big/ \sum_{j \neq i} e^{\lambda x_{j\cdot}}$ of player i·'s power is to 1, the larger their marginal benefit of power accumulation, as shown in Figure 1.1. Note that the dependence of H and h on N will henceforth be suppressed when there is little risk of confusion.

Remark 1. Parameter λ provides a tractable, systematic way to analyze the role played by the institutional constraints on the effectivity of power in reduced form. Larger λ increase the effectivity $e^{\lambda x_{\cdot}}$ of any given level of power $x_{\cdot}$. This is because larger λ correspond to conflicts that are less noisy in that their outcome depends more heavily on players' relative powers (Hirshleifer, 1989). To illustrate, note that as $\lambda \to 0$, the victor is essentially decided by a fair N-sided dice roll. As $\lambda \to \infty$, (one of) the strongest player(s) win with probability 1, like in an all-pay auction.[15]

I focus on Markov perfect equilibrium (Maskin and Tirole, 2001). The state variable in period t is $\boldsymbol{x}_{t-\Delta} \in [0, \chi]^N$, the previous period's power structure; the initial power

[15] Lanchester (1916) and Hillman and Riley (1989) consider the latter limiting case.

structure $\boldsymbol{x}_0 \in [0, \chi]^N$ is exogenously fixed. Given $\boldsymbol{x}_{t-\Delta}$, player i_t's action set is [16]

$$X_{it}(\boldsymbol{x}_{t-\Delta}) \equiv [\max\{x_{i,t-\Delta} - \delta\Delta, 0\}, \chi] \tag{1.5}$$

A *strategy* $x_{it} : [0, \chi]^N \to X_{it}$ for player i_t maps each state $\boldsymbol{x}_{t-\Delta}$ to an action x_{it} in $X_{it}(\boldsymbol{x}_{t-\Delta})$. The sequence $\{(x^*_{1t}, ..., x^*_{Nt})\}_{t\in\{\Delta,2\Delta,...\}}$ is a (*Markov perfect*) *equilibrium* – henceforth simply referred to as "equilibrium" – if at each t, $x^*_{it}(\boldsymbol{x}^*_{t-\Delta})$ is a best response to $\boldsymbol{x}^*_{-i,t}(\boldsymbol{x}^*_{t-\Delta})$ $\forall i \in \{1, ..., N\}$.

1.3 Equilibrium Power Structures

The problem faced by the lineage-i player in period $t \in \{0, \Delta, 2\Delta, ...\}$ is given by

$$\begin{cases} \max\limits_{x_{it}, I_{it}} & H(x_{it}, \boldsymbol{x}_{-i,t}) - \Delta \cdot C(I_{it}, x_{i,t-\Delta}) \\ \text{s.t.} & x_{it} = I_{it}\Delta + x_{i,t-\Delta} - \delta\Delta \\ & 0 \le x_{it} \le \chi \\ & I_{it} \ge 0 \end{cases} \tag{1.6}$$

The equilibrium of the game described above can be characterized using (1.6). Before doing so, it is important to note the following:

Proposition 1. *Given any initial power structure $\boldsymbol{x}_0$, the equilibrium of this game is unique for all sufficiently small period length Δ.*

Proof. Found in appendix subsection A.1.1. ■

[16] Notice that since x_{it} and I_{it} "pin down" one another, the number of choice variables can be reduced to one. For the purposes of defining strategies and equilibria, x_{it} is considered the only choice variable of player i_t.

Proposition 1 guarantees that each initial power structure $\boldsymbol{x}_0$ yields a unique equilibrium path $\{\boldsymbol{x}_0^*, \boldsymbol{x}_\Delta^*, \boldsymbol{x}_{2\Delta}^*, ...\}$ when Δ becomes small. Intuitively, this uniqueness stems from the fact that increasing power by any *fixed* amount $I.$ becomes prohibitively expensive as Δ becomes small. As Δ approaches zero, players will turn out to *differentially* adjust their stocks of power in equilibrium (as opposed to making infrequent "large" adjustments). This is shown in Proposition 2, where I solve the above game and make period length Δ arbitrarily small so that that the equilibrium dynamics of $\boldsymbol{x}_t^*$ can be studied in continuous time (which is more tractable for analysis).

1.3.1 Equilibrium Power Dynamics

Let $\dot{x}_{it}^* \equiv \lim_{\Delta \to 0} \frac{x_{it}^* - x_{i,t-\Delta}^*}{\Delta}$ and let $(D_1C)^{-1}(\cdot,\cdot)$ denote the inverse function of $D_1C(\cdot,\cdot)$ with respect to its first argument, keeping its second argument fixed. With this notation in hand, the equilibrium dynamics of $\boldsymbol{x}_t^*$ are characterized in continuous time as follows.

Proposition 2. *As $\Delta \to 0$, $\dot{x}_{it}^*$ obeys the following law of motion for each $i \in \{1, \dots, N\}$:*

$$\dot{x}_{it}^* = \begin{cases} -\delta \mathbb{1}_{\mathbb{R}_{++}}(x_{it}^*), & \text{if } h(x_{it}^*, \boldsymbol{x}_{-i,t}^*) < D_1C(0, x_{it}^*) \\ 0, & \text{if } h(x_{it}^*, \boldsymbol{x}_{-i,t}^*) > D_1C(\delta, x_{it}^*) \text{ and } x_{it}^* = \chi \\ (D_1C)^{-1}\big(h(x_{it}^*, \boldsymbol{x}_{-i,t}^*), x_{it}^*\big) - \delta, & \text{otherwise,} \end{cases} \tag{1.7}$$

Proof. Found in appendix subsection A.1.1. ■

The first two parts of (1.7) correspond to the corner solutions of (1.6) while the third corresponds to the interior solution. The first part states that when the marginal benefit $h(x_{it}^*, x_{-i,t}^*)$ of power accumulation is less than the marginal cost $D_1C(I_{it}, x_{it}^*)$ of

accumulating power at any $I_{it} \geq 0$, then it is optimal for player i_t to not add any power to their stock so that it depreciates unabated ($\dot{x}^*_{it} = -\delta$) or remains at zero. The second equation implies that player i_t maintains the maximum level of power ($x^*_{it} = \chi$) when the net marginal gain of doing so is positive. Otherwise the third equation applies, and the optimal $\dot{x}^*_{it}$ equalizes the marginal benefit and marginal cost of power accumulation:

$$\underbrace{h(x^*_{it}, \boldsymbol{x}^*_{-i,t})}_{\text{marginal benefit}} = \underbrace{D_1 C(\dot{x}^*_{it} + \delta, x^*_{it})}_{\text{marginal cost}}. \tag{1.8}$$

This case most clearly illustrates the two primary forces behind players' equilibrium power accumulation behavior: first, how closely matched player i_t is with their competition[17] *positively* effects $\dot{x}^*_{it}$ through increasing marginal benefit h. Second, how much power x^*_{it} player i_t holds decreases the marginal cost $D_1 C$ of accumulating at any given rate, which has a positive effect on $\dot{x}^*_{it}$. Note that when a marginal increase in player i_t's stock of power x^*_{it} makes them marginally more closely matched with their competition, these effects work in parallel and jointly induce an increase in $\dot{x}^*_{it}$. Otherwise, they are *counterveiling* effects.

Notice that equation (1.7) is *time-invariant*; how a group's power structure evolves in equilibrium depends only on its *current* power structure ($\boldsymbol{x}^*_t = \boldsymbol{x}^*_{t'} \Leftrightarrow \dot{x}^*_{it} = \dot{x}^*_{it'}$, $\forall i, t, t'$). This is indicative of the results of the next section, which characterizes the asymptotic behavior of $\boldsymbol{x}^*_t$. In the absence of shocks that affect h or $D_1 C$, the group's initial power structure is the sole determinant of its asymptotic power structure.

Moving forward, notation will often be simplified by suppressing time subscripts and asterisks: "$\dot{x}_i$" and "x_i" should henceforth be taken to mean "$\dot{x}^*_{it}$" and "x^*_{it}," respectively. Moreover, I will often refer to "player i_t" as "player i."

[17] In the sense discussed below equation (1.4).

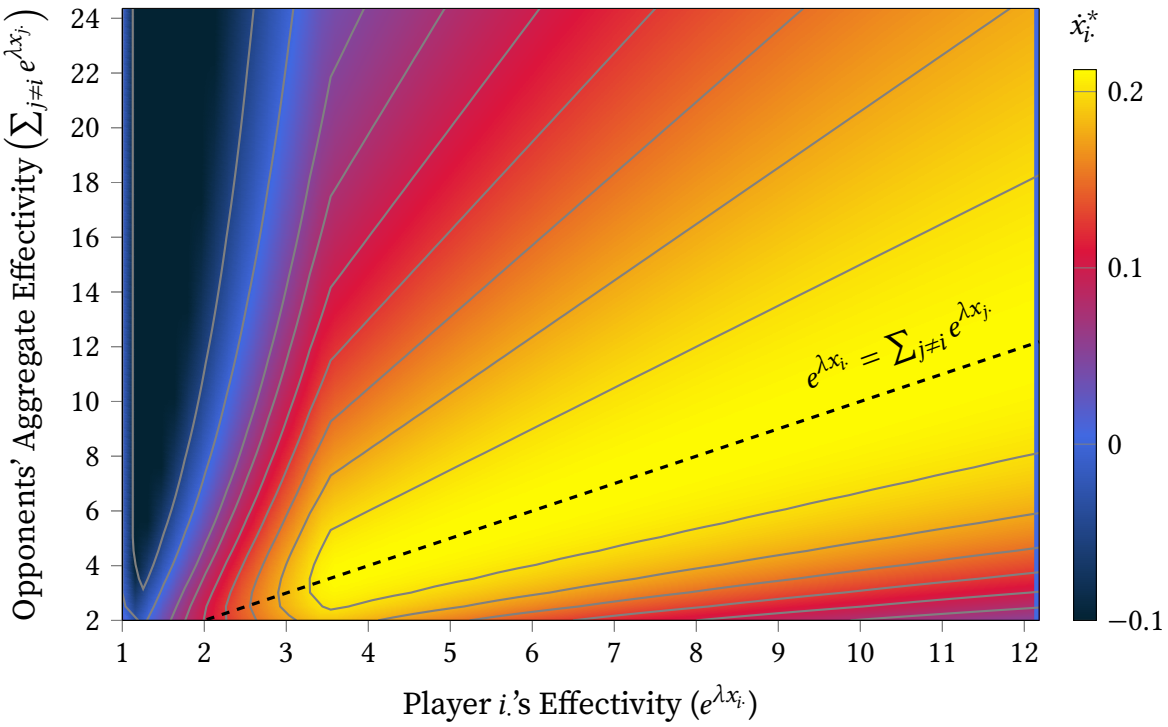


Figure 1.2: Player i.'s equilibrium power accumulation rate ($\dot{x}_i^*$) as a function of their effectivity ($e^{\lambda x_i}$) and their opponents' aggregate effectivity ($\sum_{j\neq i} e^{\lambda x_j}$).[18] This figure is produced assuming $N = 3$ players, depreciation rate $\delta = 0.1$, institutional constraint parameter $\lambda = 2.5$, power cap $\chi = 1$, and cost function $C(I_i, x_i) = I_i^2 + \max\{0.5 - x_i, 0\}I_i$.

1.3.2 Stable Power Structures

Now that the equilibrium dynamics of $\boldsymbol{x}_t$ have been fully characterized, attention is turned to the the *stable* power structures that can arise in the long run.

Definition 1. A power structure $\bar{\boldsymbol{x}} \in [0, \chi]^N$ is *stable* if

a. $\dot{x}_i = 0 \; \forall \, i$ at $\bar{\boldsymbol{x}}$, and

b. $\forall \varepsilon > 0, \exists \rho > 0$ such that if $\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}\| < \rho$, then $\|\boldsymbol{x}_t - \bar{\boldsymbol{x}}\| < \varepsilon \; \forall \, t \geq 0$ and $\lim_{t\to\infty} \|\boldsymbol{x}_t - \bar{\boldsymbol{x}}\| = 0$, where $\|\cdot\|$ denotes the Euclidean norm.

[18] This is easily derived from (1.7) using the final equality of (1.4) and noticing that $x_i = \ln(e^{\lambda x_i})/\lambda$.

Part a of this definition requires the system to be at rest at $\bar{\boldsymbol{x}}$; when this is satisfied $\bar{\boldsymbol{x}}$ is often referred to as a *steady state*. Part b requires that all trajectories that start near $\bar{\boldsymbol{x}}$ not only remain near $\bar{\boldsymbol{x}}$, but also converge to $\bar{\boldsymbol{x}}$. Proposition 3 fully characterizes the stable power structures that can arise under Assumptions 1 and 2, which will turn out to always take one of the following forms:

1. *Inclusive*, where all players hold zero or all hold χ units power. That is,

$$\bar{\boldsymbol{x}} \in \{(0, \dots, 0), (\chi, \dots, \chi)\} =: \mathscr{I}. \tag{1.9}$$

   I refer to $(0, \dots, 0)$ as *de-escalated inclusive*, and to $(\chi, \dots, \chi)$ as *escalated inclusive*.

2. *Oligarchic*, where $k \in \{2, ..., N-1\}$ players ("the oligarchs") hold χ units of power, and the remaining $N-k$ players are powerless. That is

$$\bar{\boldsymbol{x}} \in \left\{ \boldsymbol{x} \in \{0, \chi\}^N \ : \ \sum_{i=1}^{N} \mathbb{1}_{\{\chi\}}(x_i) = k \right\} =: \mathscr{O}_k. \tag{1.10}$$

   Given $k \in \{2, \dots, N-1\}$, I refer to the elements of $\mathscr{O}_k$ as *k-archic* power structures. The set of oligarchic power structures is defined as $\cup_{k=2}^{N-1} \mathscr{O}_k$.

3. *Dictatorial*, where only one player ("the dictator") holds a strictly positive amount $d \in (0, \chi]$ of power. That is,

$$\bar{\boldsymbol{x}} \in \{(d, 0, \dots, 0), \dots, (0, \dots, 0, d)\} =: \mathscr{D}_d. \tag{1.11}$$

   I refer to $\bar{\boldsymbol{x}}$ as *strong dictatorial* if $d = \chi$ and *weak dictatorial* otherwise.

With this terminology in hand, the following proposition characterizes the necessary and sufficient conditions (labeled with roman numerals) under which inclusive, oli-

garchic, and dictatorial power structures are stable (parts 1-5) and establishes that power structures outside these classes are never stable (part 6).

Proposition 3. *$\bar{\boldsymbol{x}}$ is stable only if it is inclusive, oligarchic, or dictatorial. More specifically:*

1. *The escalated inclusive power structure $(\chi,...,\chi)$ is stable if and only if*

$$h\big(\chi,(\chi,...,\chi)\big) > D_1C(\delta,\chi) \tag{I}$$

2. *The de-escalated inclusive power structure $(0,...,0)$ is stable if and only if*

$$h\big(0,(0,...,0)\big) \leq D_1C(0,0) \tag{II}$$

3. *Let $k \in \{2,...,N-1\}$. Each k-archic power structure $\bar{\boldsymbol{x}} \in \mathscr{O}_k$ is stable if and only if*

$$h\big(\chi,(\overbrace{\chi,...,\chi}^{k-1},\overbrace{0,...,0}^{N-k})\big) > D_1C(\delta,\chi) \text{ and } h\big(0,(\overbrace{\chi,...,\chi}^{k},\overbrace{0,...,0}^{N-k-1})\big) < D_1C(0,0). \tag{III}$$

4. *Let $d \in (0,\chi)$. Each weak dictatorial power structure $\bar{\boldsymbol{x}} \in \mathscr{D}_d$ is stable if and only if*

$$\begin{aligned} &h\big(\cdot,(0,...,0)\big) \text{ intersects } D_1C(\delta,\cdot) \text{ from above at } d,^{19} \text{ and} \\ &h\big(0,(d,0,...,0)\big) < D_1C(0,0) \end{aligned} \tag{IV}$$

5. *Each strong dictatorial power structure $\bar{\boldsymbol{x}} \in \mathscr{D}_\chi$ is stable if and only if*

$$h\big(\chi,(0,...,0)\big) > D_1C(\delta,\chi) \text{ and } h\big(0,(\chi,0,...,0)\big) < D_1C(0,0). \tag{V}$$

6. *No other stable power structures are possible.*

[19] I.e. $h(d,(0,...,0)) - D_1C(\delta,d) = 0$ and $\exists \varepsilon > 0$ s.t. $h(x_\cdot,(0,...,0)) - D_1C(\delta,x_\cdot) > 0 \ \forall x_\cdot \in (d-\varepsilon,d)$ and $h(x_\cdot,(0,...,0)) - D_1C(\delta,x_\cdot) < 0 \ \forall x_\cdot \in (d,d+\varepsilon)$.

Proof. Found in appendix subsection A.1.1. ■

The intuition behind each of the above conditions is quite natural, so they are only very briefly sketched here. Condition I says that each player has a strictly positive net marginal gain of maintaining χ units of power. This makes it optimal for each player to maintain χ units of power when $\boldsymbol{x} = (\chi,...,\chi)$ and – by the continuity of h and D_1C – guarantees that players accumulate power ($\dot{x}_i > 0 \ \forall i$) when $\boldsymbol{x}$ is sufficiently near to $(\chi,...,\chi)$. The *necessity* of Condition I for the stability of $(\chi,...,\chi)$ is most easily seen in the case where $h(\chi,(\chi,...,\chi) < D_1C(\delta,\chi)$: convex investment costs make it optimal for each player to optimally allow their power to depreciate when $\boldsymbol{x} = (\chi,...,\chi)$, so that it fails to be a steady state.[20]

Condition II implies that when all players are powerless, it is not optimal for any player to accumulate power (so that $\dot{x}_i = 0 \ \forall i$ at $\boldsymbol{x} = 0$). When the inequality in Condition II is strict, then $h(x_i,\boldsymbol{x}_{-i}) < D_1C(0,\boldsymbol{x}_i) \ \forall i$ when $\boldsymbol{x}$ is sufficiently close to $(0,...,0)$, since h and D_1C are continuous. At all such $\boldsymbol{x}$, every player optimally allows their power to depreciate at rate δ, eventually causing each to hold no power.[21] When Condition II fails, it follows from the convexity of investment costs that each player begins to accumulate power (so that $\dot{x}_i > 0 \ \forall i$ at $\boldsymbol{x} = (0,...,0)$).

Remark 2. The de-escalated inclusive power structure $(0,...,0)$ emerges in equilibrium only when each player lets their power fully depreciate; this represents in a certain sense the trivial case of the model (which is not ruled out by Assumptions 1 and 2).

The first (second) part of Condition III plays a similar role as Condition I (II). The first part says that the net marginal gain of maintaining χ units of power is positive

[20] When $h(\chi,(\chi,...,\chi) = D_1C(\delta,\chi)$, each player maintains their power level when $\boldsymbol{x} = (\chi,...,\chi)$, so that it is a steady state. However, if one takes $\varepsilon > 0$ units of power from each player, the power structure will not return to $\boldsymbol{x}$ in equilibrium, so that the second part of the definition of stability is violated.

[21] The explanation in the case where Condition II holds with equality is more involved; for more details, please see the proof of Proposition 3.

when faced with $k-1$ other players who also hold χ units of power, and $N-k$ players who hold no power. The second part implies that it is not optimal for powerless players to accumulate power at k-archic power structures. Notice that Condition V is essentially the $k=1$ analogue of Condition III.

Condition IV is necessary and sufficient for the stability of weak dictatorships, where player i ("the dictator") holds $x_i = d \in (0, \chi)$ units of power, and all other players hold no power. The second part of this condition makes power accumulation suboptimal for powerless players when the dictator holds d units of power. The first part of Condition IV says that the dictator player i's marginal cost $D_1C(\delta, d)$ of maintaining d units of power is equal to its marginal benefit $h(d, (0,\ldots,0))$, so that maintaining d units of power is optimal for the dictator. Furthermore, it says that $h(\cdot, (0,\ldots,0))$ intersects $D_1C(\delta,\cdot)$ from *above*, which is crucial for the second part of Definition 1 to be satisfied. When this holds, then decreasing (increasing) the dictator's power by any "small" amount causes them to optimally accumulate power (let their power depreciate) until they return to holding d units of power. However, when $h(\cdot, (0,\ldots,0))$ intersects $D_1C(\delta,\cdot)$ from *below*, any such perturbation will cause x_i to drift *away* from d in equilibrium.

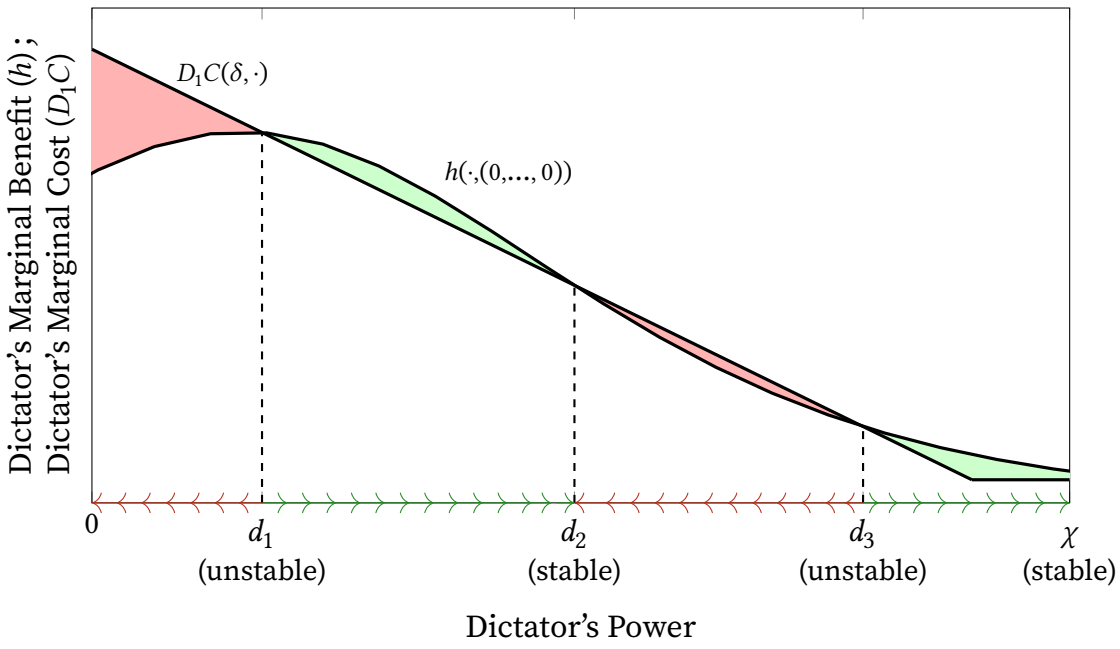


Figure 1.3: Power level $d \in (0, \chi)$ is sustained in a stable *weak* dictatorship only if the dictator's marginal benefit $h\big(\cdot,(0,\ldots,0)\big)$ of power intersects marginal cost $D_1(\delta,\cdot)$ at d from above. Power level $d = \chi$ is sustained in a stable *strong* dictatorship only if a dictator's marginal benefit $h(\chi,(0,\ldots,0)$ of maintaining χ units of power strictly outweighs the marginal cost $D_1C(\delta,\chi)$ of maintaining that power level.

Finally, the last part of Proposition 3 shows that if $\bar{\boldsymbol{x}}$ is not inclusive, oligarchic, or dictatorial, it cannot be stable. The $\bar{\boldsymbol{x}}$ outside the aforementioned three classes either have (1) two or more players with interior power levels or (2) exactly one player with an interior power level. First consider the case where $0 < \bar{x}_i < \bar{x}_j < \chi$ for some $i \neq j$ (i.e. two players hold different interior power levels at $\bar{\boldsymbol{x}}$). If such an $\bar{\boldsymbol{x}}$ is a steady state ($\dot{x}_\cdot = 0$ for all players), then (1.7) implies that for each player $\ell \in \{i, j\}$, the marginal cost $D_1C(\delta, \bar{x}_\ell)$ of maintaining $\bar{x}_\ell$ units of power is equal to its marginal benefit $h(\bar{x}_\ell, \bar{\boldsymbol{x}}_{-\ell}) = D_1C(\delta, \bar{x}_\ell)$ $\forall \ell \in \{i, j\}$. This leads to a contradiction since $D_1C(\delta, x_i) \geq D_1C(\delta, x_j)$ (Assumption 1.2) and $h(\bar{x}_i, \bar{\boldsymbol{x}}_{-i}) < h(\bar{x}_j, \bar{\boldsymbol{x}}_{-j})$ (shown in the proof).

The power structures that remain to be considered may be steady states, but never stable ones, since arbitrarily small perturbations cause $\boldsymbol{x}_t$ to eventually leave a neighborhood of $\bar{\boldsymbol{x}}$ at some t. This is achieved by giving the interior player(s) a posi-

tive amount of power (each in the same amount). This windfall of power – even when arbitrarily small – increases the net marginal gain of investment of the interior-power player(s) by the same amount, inducing each to begin accumulating power at a common, positive rate. When the aforementioned neighborhood and windfall are sufficiently small, players who held 0 or χ units of power at $\bar{\boldsymbol{x}}$ will optimally remain at their respective power levels not only after the perturbation but also as the interior-power player(s) accumulate power within the neighborhood. (This is because h and $D_1C(\delta,\cdot)$ are continuous.) That is, as the interior power players accumulate power at the same rate, the extremal power players will not move from their respective positions, and the group's power structure will eventually leave any sufficiently small neighborhood of $\bar{\boldsymbol{x}}$, making it unstable.

1.3.3 Three Player Illustration

This section illustrates the global equilibrium dynamics in the case with $N = 3$ players. In the interest of clearly visualizing these results, I focus on the case where the set of stable power structures is

$$\Big\{\underbrace{(\chi,\chi,\chi)}_{\substack{\text{escalated}\\ \text{inclusive}}},\underbrace{(d,0,0),(0,d,0),(0,0,d)}_{\text{dictatorial}},\underbrace{(\chi,\chi,0),(\chi,0,\chi),(0,\chi,\chi)}_{\substack{\text{oligarchic}\\ (2\text{-archic})}}\Big\},$$

where $d \in (0,\chi]$. Figure 1.4 visualizes the equilibrium dynamics in this case; the intuition behind these dynamics are now discussed.[22]

[22]This intuition applies for arbitrary N; the intuition behind how the de-escalated inclusive power structure is reached is discussed in Remark 2. Finally, note that additional visualizations of equilibrium dynamics are provided in Appendix A.1.3.

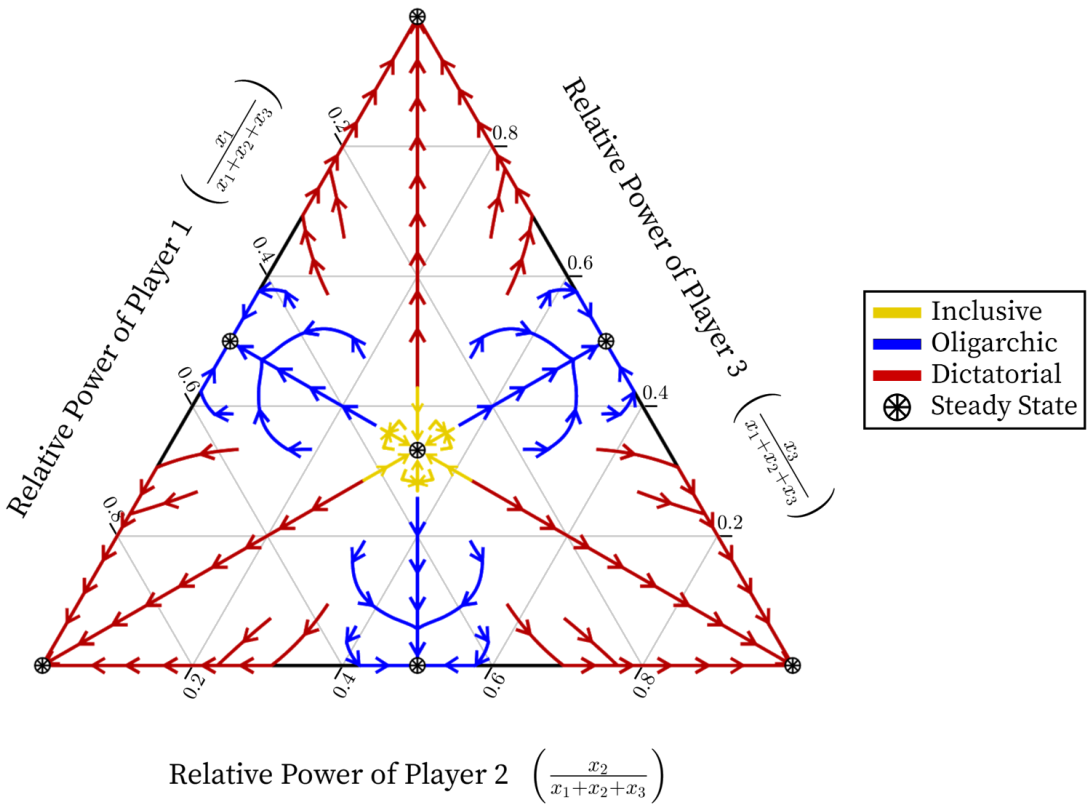


Figure 1.4: Simplex plot representation of simulated equilibrium paths produced using cost function $C(I_\cdot, x_\cdot) = I_\cdot^2 + (1 - x_\cdot)I_\cdot$, institutional constraint parameter $\lambda = 5.5$, and depreciation parameter $\delta = 0.1$.

When players' powers are initially close to one another, the *(escalated) inclusive* power structure is reached through competition. Since players' contest incentives are strongest when they are evenly matched with one another – in the sense discussed below equation (1.4) – each player begins with similarly strong power accumulation incentives. Moreover, each initially face similar power accumulation costs (and marginal costs). Consequently, players begin accumulating power at similar rates in equilibrium, resulting in each becoming more powerful but their *relative* powers remaining similar. This cycle repeats until each player reaches χ units of power.

Dictatorial power structures are reached through a qualitatively different process. One player (say, player 1) begins significantly more powerful than the rest, which thereby gives them a significant *cost advantage* at the outset (Assumption 1.2). This al-

lows them to accumulate power at a faster rate than other players in equilibrium, causing their gap in power to *widen* over time. Not only does this cause player 1's cost advantage to amplify, it also eats away at each player's power accumulation incentives. These incentives progressively weaken to the point where even matching depreciation rate δ becomes sub-optimal for players 2 and 3, so that these players eventually (optimally) allow their respective powers to *fully* depreciate, at which point player 1 has fully *consolidated power*. This is an example of the *Discouragement Effect* that is known to arise in a wide variety of dynamic contests (Konrad, 2012).

Finally, *oligarchic* (here, 2-archic) power structures are reached through a *combination* of the above two processes. These are reached when two players (say, players 1 and 2) begin closely matched to each other but outmatch the rest (here, player 3). Players 1 and 2 compete with one another, each driving the other player's power up in the same way that the escalated inclusive power structure is reached. This causes these players to "outrun" (in terms of power accumulation) player 3, who eventually allows their power to fully depreciate due to the Discouragement Effect.

While this model generates very natural equilibrium dynamics, it also yields quite novel insights, which are now presented in the next section.

1.4 Properties of Stable Power Structures

1.4.1 Stable Power Structures in Large Societies

I now turn to the main result of this chapter: as it turns out, there always exists a finite group size past which the *escalated inclusive* power structure ceases to be stable.

Proposition 4. *The escalated inclusive power structure* $(\chi,...,\chi)$ *is not stable in groups larger*

than

$$\bar{N}_\chi^I \equiv \begin{cases} \left\lceil \frac{\lambda+\sqrt{(\lambda-4D_1C(\delta,\chi))\lambda}}{2D_1C(\delta,\chi)} \right\rceil & \text{, if } \lambda/4 \geq D_1C(\delta,\chi) \\ 0 & \text{, otherwise.}^{23} \end{cases} \tag{1.12}$$

Proof. Found in appendix subsection A.1.1. ■

This result implies that in a sufficiently large society, *unchecked* political competition will *inevitably* leave a subset of its population marginalized. The takeaway of this result should not be a familiar sense of resignation, but of urgency: political competition must be regulated to make inclusivity achievable in large societies; failing to do so guarantees its impossibility.

What kinds of interventions can keep the escalated inclusive power structure stable in large societies? Before turning to this matter, it is first important to understand why the escalated inclusive power structure destabilizes in sufficiently large groups. Recall that by Proposition 3.1, the escalated inclusive power structure is stable if and only if

$$\underbrace{\frac{(N-1)\lambda}{N^2}}_{\because \text{ Equation (1.4)}} = \underbrace{h\big(\chi,(\chi,...,\chi);N\big) > D_1C(\delta,\chi).}_{\text{Condition I}}$$

The above was used to derive N_χ^I. Notice that a player's marginal benefit $h\big(\chi,(\chi,...,\chi);N\big)$ at $\boldsymbol{x} = (\chi,...,\chi)$ is decreasing[24] in N and decays to zero as N grows large, eventually falling below the marginal cost $D_1C(\delta,\chi)$ of maintaining χ units of power after the group grows larger than N_χ^I.

To see the intuition behind this, recall that players' power accumulation incentives h are increasing in how closely matched they are with their competitors (in the sense discussed after equation (1.4)). At the escalated inclusive power structure, every

[23] $(\chi,...,\chi)$ is not stable at any $N \geq 2$ when $\lambda/4 < D_1C(\delta,\chi)$. See proof for more details.

[24] I.e. $h\big(\chi,(\chi,...,\chi);N'\big) < h\big(\chi,(\chi,...,\chi);N\big) \quad \forall N' > N \geq 2.$

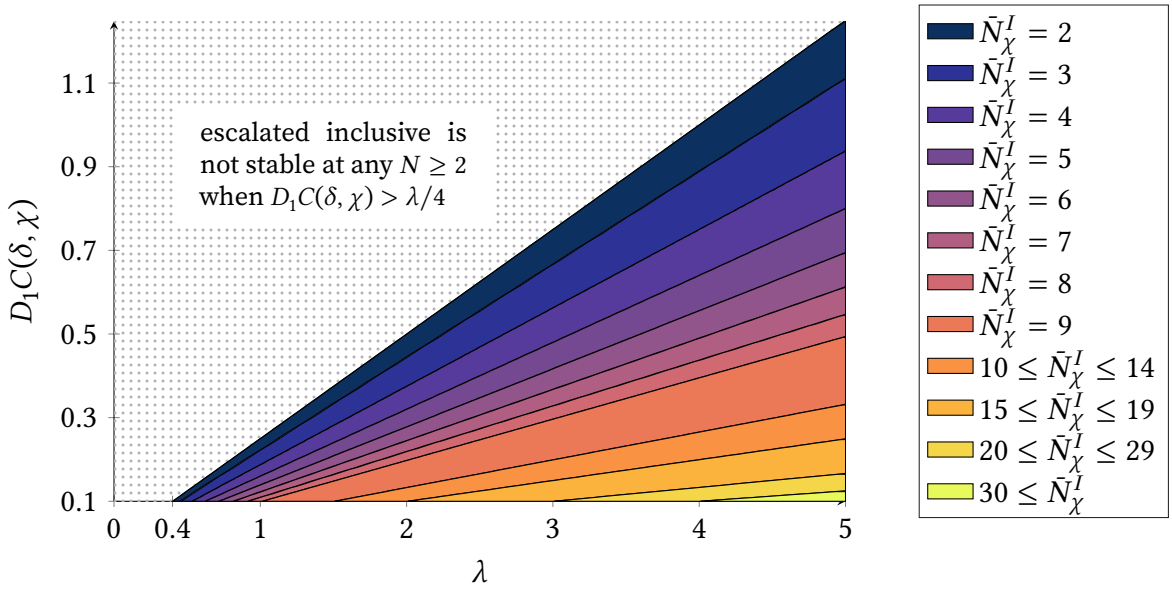


Figure 1.5: The group size $\tilde{N}_\chi^I$ past which the escalated inclusive power structure $(\chi,\dots,\chi)$ is never stable depends on noise/institutional constraint parameter λ and the marginal cost $D_1C(\delta,\chi)$ of maintaining χ units of power.

player faces $N-1$ opponents that are each as strong as they are. This is why each player has strong power accumulation incentives when N is small; in fact they are as strong as possible when $N = 2$. However, as N grows large, players become overwhelmed by their *aggregate* competition at $\boldsymbol{x} = (\chi,\dots,\chi)$. This is somewhat ironic, since – as was discussed in subsection 1.3.3 – competitive pressure is what drives $\boldsymbol{x}$ towards this power structure; when N grows large, competitive pressure is also what snuffs it out.

To return to the matter of policy interventions, I establish the following comparative statics result:

Corollary 1. *When $\lambda/4 \geq D_1C(\delta,\chi)$, $\tilde{N}_\chi^I$ is strictly increasing in λ and strictly decreasing in $D_1C(\delta,\chi)$; otherwise it is increasing in λ and decreasing in $D_1C(\delta,\chi)$.*

Proof. Found in appendix section A.1.1. ■

Lowering $D_1C(\delta,\chi)$ corresponds to decreasing the marginal cost of maintaining χ units of power; increasing λ corresponds to loosening institutional constraints on the

effectivity $e^{\lambda x_i}$ of power x_i (Remark 1). The above corollary suggests that either change increases $\bar{N}_\chi^I$, so that the escalated inclusive power structure is more robust to group size. This may seem counter-intuitive at first, since these policies seem to favor those who are already powerful.

This is indeed the case: observe that interventions that increase λ (decrease $D_1C(\delta)$) simply serve to increase (decrease) the left-hand (right-hand) side of Condition I, mentioned just above. Recalling the discussion of Proposition 3, Condition I is necessary and sufficient for the stability of $(\chi,...,\chi)$ because it ensures that players' net marginal gain of power accumulation is positive when they are all sufficiently powerful. Moreover, any finite increase (decrease) in λ ($D_1C(\delta)$) is only a temporary solution in growing societies, since they only increase $\bar{N}_\chi^I$ by a finite amount.

The hard upper bound χ on power is also a policy lever; a real world example is the U.S. Supreme Court's decision in *Citizens United v. FEC (2010)*, which effectively made the legal cap on political expenditures unbounded. Notice that increasing χ raises $\bar{N}_\chi^I$ only by diminishing $D_1C(\delta,\chi)$ (Assumption 1). Consequently, the discussion in the paragraph above also applies here; given the United States' persistent, alarming rise in inequality discussed in the introduction, raising χ may not be helpful. Moreover, it turns out that in any case this policy lever will only get you so far. As I now show in Proposition 5.1, making χ unbounded – like in the *Citizens United* decision – $\bar{N}_\chi^I$ remains finite in all but a knife-edge case where the marginal cost of maintaining an arbitrarily large amount of power becomes arbitrarily close to free. The remainder of this proposition shows that dictatorships and oligarchies are far more robust to population size.

Proposition 5. *Suppose that* $\lim_{\chi\to\infty} D_1C(\delta,\chi) > 0$.[25] *If* χ *is made arbitrarily large, then*

1. *The group size past which the escalated inclusive power structure is unstable remains*

[25] That is, some $\varepsilon > 0$ – which may be *arbitrarily small* – bounds $D_1C(\delta,\cdot)$ from below. This only leaves the knife-edge case where $D_1C(\delta,\chi)$ decays to *exactly* 0 as $\chi \to \infty$.

finite.

2. *Apart from the trivial case when* $\lambda/4 \leq \lim_{\chi\to\infty} D_1C(\delta,\chi)$,[26] *dictatorial power structures remain stable at arbitrarily large group sizes.*

3. *k-archies remain stable at arbitrarily large group sizes if k is less than or equal to*

$$\bar{k} \equiv \begin{cases} \left\lfloor \lim\limits_{\chi\to\infty} \frac{\lambda+\sqrt{(\lambda-D_1C(\delta,\chi))\lambda}}{D_1C(\delta,\chi)} \right\rfloor & \text{if } \lambda > \lim_{\chi\to\infty} D_1C(\delta,\chi) \\ 0 & \text{else.} \end{cases} \tag{1.13}$$

Otherwise, k-archies with $k \in \{\bar{k}+1, \bar{k}+2, ...\}$ *are not stable at any group size.*

Proof. Found in appendix section A.1.1. ■

When $\chi \to \infty$, the restriction of power to $[0,\chi]$ becomes relaxed by an arbitrarily large amount. The first part of this proposition shows that in all but the aforementioned knife-edge case, the escalated inclusive power structure still becomes unstable past a finite group size, and for the same reason as before: players become overwhelmed by their aggregate competition. However, notice in equation (1.12) that relaxing institutional constraints on the *effectivity* of power (i.e. making λ arbitrarily large) allows the escalated inclusive power structure to remain stable in arbitrarily large group sizes. Recalling Remark 1, this is effectively amounts to turning conflict into an all-pay auction.

Proposition 5 provides another interesting implication in its latter two parts: dictatorships and oligarchies with *sufficiently few* oligarchs are robust to group size.[27] As I discussed in the Literature Review, this reflects a stylized fact that is far from fully

[26] When $\lambda/4 \leq \lim_{\chi\to\infty} D_1C(\delta,\chi)$, the maximum attainable marginal benefit $\lambda/4$ (see Figure 1.1) is less than the marginal cost of maintaining any positive level of power (because of Assumption 1.2) $\forall N \geq 2$.

[27] Note that keeping χ *fixed* will artificially cause dictatorships and oligarchies to become unstable past a finite group size (details are provided in Propositions 11 and 12 in Appendix A.1.2). Moreover, note that Since the trivial case of this model is not ruled out by its assumptions, the de-escalated inclusive power structure may remain stable at arbitrarily large group sizes (Remark 2).

understood: power tends to fall into the hands of a few in large groups of people. In contrast to prevailing explanations, the one provided here does not rely on the particular details of political institutions; it simply stems from the nature of incentives in power accumulation competitions.

To see the intuition for why k must be sufficiently small for a k-archy to be robust to group size, recall that in k-archic power structures, each oligarch[28] is individually equally matched with $k-1$ other players. When $k \leq \bar{k}$, oligarchs face more than one – but not *too* many – closely matched opponents, which ensures strong competition incentives. Otherwise, the oligarchs become overwhelmed like the players in the escalated inclusive power structure.

This leaves one final mystery: why do dictatorships remain stable at arbitrarily large group sizes? Since dictators have no closely matched competitors, shouldn't their contest incentives be weak? This is indeed the case when N is small, but as N becomes large this story qualitatively shifts.

1.4.2 Comparative Statics of Stable Dictatorial Power

This section characterizes the comparative statics of *stable dictatorial power*, the amount of power held by the strongest player ("the dictator") in a stable dictatorship. Recall that Proposition 3 established that *weak* dictatorships (with dictatorial power $d \in (0, \chi)$) are stable if and only if

$$h\big(\cdot, (0,\ldots,0)\big) \text{ intersects } D_1C(\delta,\cdot) \text{ from above at } d, \text{ and} \qquad \text{(Condition IV)}$$

$$h\big(0,(d,0,\ldots,0)\big) < D_1C(0,0),$$

[28] Recall that the strongest players in k-archic power structures are termed "oligarchs."

and strong dictatorships (with dictatorial power $d = \chi$) are stable if and only if

$$h\big(\chi, (0,\ldots,0)\big) > D_1C(\delta, \chi) \text{ and } h\big(0, (\chi, 0,\ldots,0)\big) < D_1C(0,0). \qquad \text{(Condition V)}$$

Depending on model primitives, it is possible for no dictatorships to be stable or for *multiple* levels of power to be sustained in stable dictatorships as in Figure 1.3 in subsection 1.3.2.[29] For ease of exposition, assume throughout this subsection that *exactly one* level of power d is sustained in a stable dictatorship.[30] Analogous results hold when multiple types of dictatorship are stable, but are substantially more cumbersome to state, and offer insubstantial additional insight.

Proposition 6. *The amount of power held by dictators in stable dictatorships increases in group size N.*

Proof. Found in appendix subsection A.1.1. ■

It is natural to expect that larger group sizes lead to stronger dictators. Mechanically, this is because increasing group size N translates the dictator's marginal benefit $h(\cdot, (0,\ldots,0); N)$ rightward (equation A.50 in the proof of Proposition 6). This is illustrated in Figure 1.6a, below. Intuitively, this is because under assumption 2, powerless players have a small but *non*-zero probability of victory.[31] As a result, powerless players collectively exert competitive pressure on the dictator player. This pressure grows with the number of powerless players, thereby inducing the dictator to hold an increasingly high level of power in stable dictatorships. Note that when this level of power is in the interior of $(0, \chi)$, it is *strictly* increasing in group size N. Hence, if χ is made arbitrar-

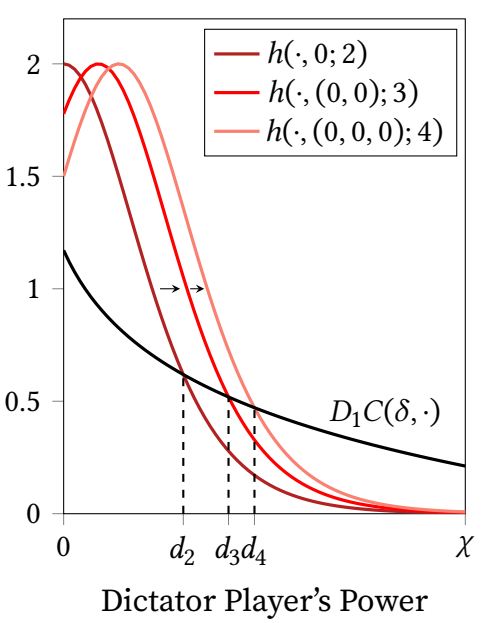
[29] Recall that the equilibrium dynamics in (1.7) are always *unique*, hence even when multiple "kinds" of dictatorships are stable. This is visualized in Figures A.1 and A.2 in Appendix A.1.3.

[30] Formally put: assume that either (1) Condition IV holds for exactly one $d \in (0, \chi)$ and Condition V fails or (2) IV fails at all $d \in (0, \chi)$ and Condition V holds.

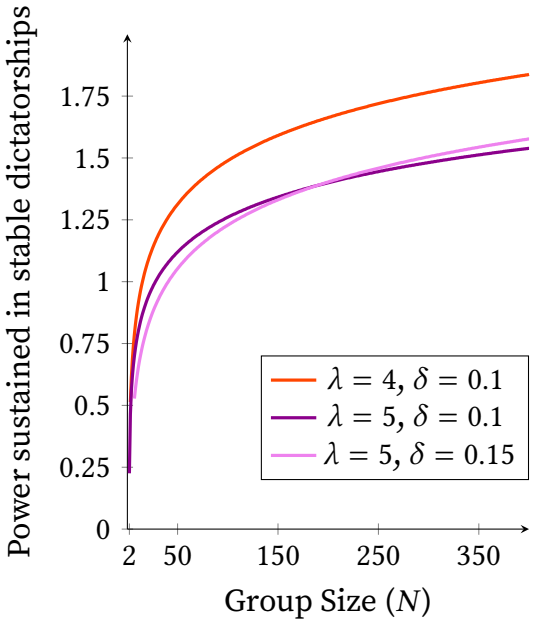
[31] As discussed in Hirshleifer (1989), this reflects the noisiness in conflicts. Figure A.5 in Appendix A.1.3 visualizes how powerless players' victory probability varies with model primitives.

ily large, the amount of power held in stable dictatorships grows without bound with group size.

As N becomes large, dictators' contest incentives – and hence optimal behavior – start to resemble those of oligarchs in 2-archies. The way in which dictators optimally respond to increases in group size is what ultimately causes their contest incentives to grow with N and allows dictatorships to be robust to group size in Proposition 5. Strong dictators emerge when (the rest of) society is *collectively* strong. While this resembles Acemoglu and Robinson's (2022b) main result, there is an added twist: non-dictator players are individually powerless, having only collective strength in numbers.



(a) Larger group sizes N induce higher levels of power held in dictatorships (denoted d_N in this figure) because it shifts the dictator's marginal benefit of power accumulation $h(\cdot, (0,\ldots,0); N)$ shifts rightward.



(b) Simulated relationship between power held by the strongest player in stable dictatorial power structures and group size using cost function $C(I., x.) = 3.25I.^2 + \max\{0.5 - x., 0\}I.$.

Figure 1.6: How larger group size induces higher levels of power in stable dictatorships.

Other comparative statics properties of the amount of power d held in stable dictatorships are given in the result below.

Proposition 7.

1. *Uniformly increasing the marginal cost of investment $D_1C(\cdot,\cdot)$ decreases d.*
2. *d is decreasing in δ.*
3. *d increasing in λ if and only if $\frac{\lambda d-1}{\lambda d+1}e^{\lambda d} < N-1$.*

Proof. Found in appendix subsection A.1.1. ■

The first two parts of this proposition consider the negative relationship between the amount of power d held by dictators in stable dictatorships and the marginal cost $D_1C(\delta, d)$ of maintaining d units of power. If the latter value were to increase – say, due to an unexpected "shock" – the dictator's marginal cost of maintaining d would outweigh the marginal benefit $h(d,(0,\dots,0))$. The dictator consequently lets their power depreciate until stabilizing at a new, lower level of power.

Conflict noise parameter λ has a less straightforward relationship with stable dictatorial power d. Increasing λ induces an increase in d if and only if they are both sufficiently small, a requirement that becomes less stringent as N increases. As λ becomes large, *simply* surpassing the other players – rather than the *amount* by which one surpasses – becomes the dominant influencing factor in winning conflicts. The role played by group size is also natural: larger N correspond to more powerless players, who always have a strictly positive probability of winning conflicts when $\lambda < \infty$. Thus, dictators in larger groups face more pressure to maintain higher levels of power in parallel fashion to Proposition 6.

1.5 Conclusion

This chapter developed an economic theory of how a society's distribution of power evolves over time. To investigate the competitive forces that underlie this evo-

lution, I studied an intergenerational power accumulation contest among multiple lineages of players, where power was modeled as an asset that increases one's chances of winning conflicts over resources. Given any initial distribution of power, this model makes a *unique* equilibrium prediction[32] of how it will evolve over time and the stable distribution to which it tends in the long run, which always falls into three classes, termed *inclusive*, *oligarchic*, and *dictatorial*.

This model also makes a far more concerning prediction: in sufficiently large societies, *unregulated* political competition inevitably leads to power falling into the hands of a few. As this turns out, this result generates a novel explanation for the longstanding empirical puzzle initiated by Michels (1915), providing a solid game theoretical foundation for his Iron Law of Oligarchy. Given this, one may worry that Michels's grim portent is unfolding before our very eyes as inequality continues to rise in nations around the world. There is indeed cause for worry, but only if nothing is done about it.

Despite Michels's (1915) assertion that "[h]istorical evolution mocks all the prophylactic measures that have been adopted for the prevention of oligarchy" (p. 406), this chapter provides a few policy implications on how to safeguard inclusivity in large societies. Relaxing institutional limits on the capacity of individual actors to influence political decision-making or policies that decrease the cost of accumulating said influence each turn out to *improve* the robustness of inclusivity to population size, but with two major caveats. The first, more obvious limitation is that these measures are local solutions in the sense that they help societies that are already sufficiently "close" to inclusivity on track to fully achieving it. The second, far more serious caveat is that such policies can only get one so far: they will never make inclusivity fully robust to population size in all but a knife-edge case. However, inclusivity can be made fully robust to population size if one is able to effectively make political conflicts into an all-pay

[32] In the absence of *shocks* to the group's power structure, the number of players, or any other model primitives that affect the costs or benefits of accumulating power.

auction.

These results – as with all of the results in this chapter – are novel in that they do not hinge on the particular details of political institutions, or on qualitative features (e.g. culture, geography, etc.) of the societies wherein they reside. Rather, they stem from two natural features of power accumulation contests: (1) the incentive to accumulate power strengthens with how closely matched one is with one's opponents and (2) the marginal cost of accumulating power diminishes with how powerful one already is. The aforementioned details come into play insofar as they affect the costs and benefits of power accumulation, and are flexibly handled by this reduced form framework, which not only makes it highly portable but also highlights the potential generality of its results. Furthermore, this framework helps unify the study of how the distribution of power evolves in societies, accommodating the endogenous emergence of an unprecedented variety of regimes ranging from dictatorship to inclusivity as well as the various "shades" of oligarchy in between.[33]

This chapter forges ahead in an emerging research area which, as Dixit (2021) discusses, is an exciting and promising one. The developments in this chapter allowed me to generate previously unattainable insights not only on the nature of political inequality in large societies, but also on the main conclusion of its foundation, Acemoglu and Robinson (2019, 2022b). While I confirm that competitive pressure is indeed what allows strong, inclusive regimes to emerge, it is also precisely what causes it to destabilize in sufficiently large societies.

Further exploring this emerging area of investigation forms part of my research agenda; as such, I conclude this chapter with a brief discussion of the directions I plan to pursue in future work. My first priority is to consider the case where longer lived agents can accumulate *productive* capital in addition to power. Along with introducing

[33]These shades of oligarchy themselves range from diarchies and triarchies to less concentrated forms that resemble gentries or the polyarchies studied in Dahl (1971).

an interesting trade-off that potentially carries novel implications for economic growth, this serves an even more important purpose: it would allow me to investigate how political and economic inequality interact and the link between their dynamics. Given the discussion in the Introduction, this is crucially important for fully understanding the dynamics of each. It may also be fruitful to study the case with multiple forms of power. As mentioned earlier, the reduced form nature of this framework was useful for clearly establishing its overarching takeaways. Now that this has been accomplished, unpacking the abstract notion of power considered here would facilitate putting data to this model. Another planned direction is to investigate the internal hierarchies that form within oligarchies by studying the case where agents can form coalitions. Finally, on the normative side I am interested in further exploring how one can reallocate power in practice. Two considerations make this a highly non-trivial task: first, a social planner may face information constraints regarding the distribution of power within the society over which they preside. The second, more serious challenge stems from the fact that in practice one cannot rely on a benevolent social planner that is completely external to society. Societies, of course, must regulate themselves, which presents an interesting institutional design problem that I plan to explore in future research.

Chapter 1 is currently being prepared for submission for publication of the material. The dissertation author, Frederick Aram Papazyan, is the sole author of this chapter.

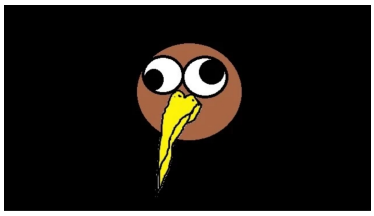
Chapter 2

Sabotage-Proof Mechanism Design

2.1 Introduction

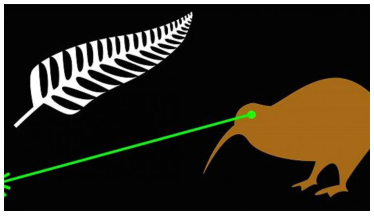
Polls – especially those conducted online – are notorious for their lack of robustness to sabotage; the derailment of online polls by internet trolls[1] is a common and well-documented occurrence, often producing amusing news articles but also potentially large costs. Take for instance the 2015–2016 New Zealand flag referendums, where an online poll was used to crowdsource a replacement for the country's national flag. The public gallery of flag submissions quickly became inundated with ridiculous, unusable flags such as those depicted in Figure 2.1, below. This process took well over a year, cost approximately 26 million New Zealand dollars, and was ultimately fruitless in producing a new flag. Similar derailments have interfered with crowdsourcing in marketing campaigns (BBC, 2016), information-gathering during the 2020 US Presidential Election (Collins and Popken, 2019; Frenkel et al., 2020; Kennedy, 2020), and even prevented the government of North Macedonia from properly counting its own population (*The Economist*, 2020).

[1] A "troll" in this context is someone who is deliberately trying to derail a poll or any other mechanism

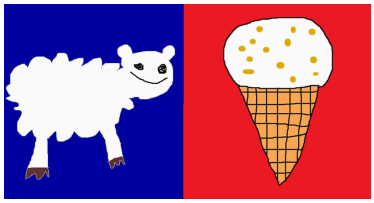


(a) "Happy Kwi [*sic*]" by Davy Lee

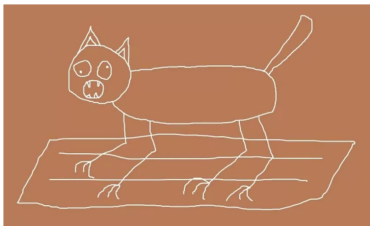
(b) "Good Flag" by James Ireland



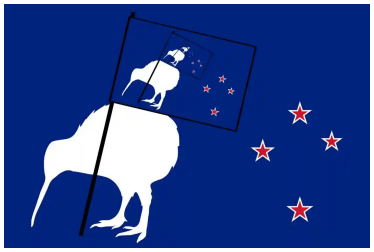
(c) "Fire the Lazar! [*sic*]" by James Gray



(d) "Sheep and Hokey Pokey" by Jesse Gibbs



(e) "Deranged Cat Raking its Garden" by Jeong Hyuk Fidan



(f) "Flag-bearing kiwi" by George George

Figure 2.1: A sample of noteworthy submissions from the New Zealand Flag Referendum gallery (New Zealand Government, 2015).

How does one optimally design voting mechanisms in the presence of saboteurs or trolls?[2] This question can be split into two parts: how does one design the entry part of the mechanism to encourage normal agents and dissuade trolls, and how does one design the voting part of the mechanism given a fixed population of participants? We aim to answer both questions in this project. However, so far our results focus on the second part. It is a natural place to start the analysis and proceed to the entry part via backward induction. For now, we consider a situation where the entry has already occurred, and analyze the designer's problem given a fixed population of participants. In this framework, we focus on analyzing a few benchmark mechanisms and characterizing the optimal mechanism as completely as possible.

We start by analyzing a simple illustrative example with two genuine agents and one troll. Each genuine agent has a private type corresponding to a bliss point over

[2]Here, we use "troll" and "saboteur" interchangeably.

actions the mechanism designer can take. The mechanism designer's objective is to maximize the welfare of the genuine agents. There is a common prior over their types. The designer gathers information through a poll and then takes an action. The troll's objective is to minimize the welfare of the genuine agents. We consider two specific mechanisms as reasonable baselines — "majority rule" and "average-of-votes" — for both methodological and empirical reasons. Choosing the outcome that was voted for on average is theoretically optimal assuming that trolls are absent, utility is quadratic, and that it is indeed possible to average over votes.[3] On the other hand, choosing the outcome with the most votes is by and large the most widely used in polls and elections. Our project offers insights into how suboptimal these two mechanisms are in the presence of trolls.

Analyzing the agents' equilibrium behavior in the illustration below, we derive the welfare implications for both mechanisms and compare with the benchmark of doing no mechanism ("no-poll benchmark"). In the example, "majority rule" performs exactly as the no-poll benchmark, whereas "average-of-votes" (a more fine-tuned mechanism) performs better when ex-ante uncertainty over types is high, but worse when it is low.

We then consider a more general model with a fixed number of agents and trolls. We derive similar predictions as in the Illustration, and show that when ex-ante uncertainty is low, the "average-of-votes" mechanism will perform poorly compared to a blind mechanism. Intuitively, this happens because under low ex-ante uncertainty there is little potential gain from gathering information from the agents, but the negative impact of trolls is still present in its full force. Consequently, when there is little information the designer can potentially gain from an informative mechanism, running a blind mechanism may be strictly better.

[3]For instance, this is possible in the Weber (1929) Problem and in other facility location problems that followed.

We derive two important properties of the optimal mechanism. First, we show that the optimal mechanism has to satisfy a *quasi-monotonicity* property, which basically requires that more votes for a given type result in an action that is closer to that type's optimal action. Interestingly, this property is always violated by the "majority rule" mechanism, which indicates that it is never optimal. Next, we show that the optimal mechanism also has to satisfy an *indifference* property, which requires the trolls to be indifferent between their messages. This allows us to reduce the search for the optimal mechanism to a narrow family of mechanisms that fulfill the property. It also allows us to generally rule out the "average-of-votes" as the optimal mechanism.

Returning to the benchmark mechanisms – "average-of-votes" and "majority rule" – we prove that they can be improved in simple ways that rely on the indifference property of the optimal mechanism. The majority rule can be improved by implementing a *supermajority* rule with a default option that is informed by the prior. This counteracts the trolls' influence in two ways: their votes are less likely to be pivotal, and the default option is the opposite of what they would vote for under the majority rule. As for the average-of-votes rule, it can be improved by a *weighted-average-of-votes* mechanism in which the option that the trolls would vote for receives a lower weight in determining the outcome than other options. This directly counteracts the trolls' influence.

Finally, we consider the limit case where the number of trolls is arbitrarily large and derive a worst-case result applicable to any (continuous) mechanism. This can model a situation where trolls have very low costs of entry (and possibly submitting multiple votes). We show that given any continuous mechanism, trolls can achieve the worst-case outcome under it if they are sufficiently numerous. A natural corollary follows: if number of trolls is potentially unlimited, the best mechanism for the designer to implement is a *blind* (or no-poll) mechanism, which ignores messages from agents and chooses the ex-ante best outcome. This speaks to the observed tendency of online

polls being shut down or cancelled when a large influx of trolls occurs, and suggests that such action may indeed be optimal in these circumstances.

At the end of the chapter, we discuss these insights in more detail and outline future extensions. Most notably, we plan to consider the entry part of the designer's problem and see how the presence of trolls affects the optimal population of agents that the designer wants to attract. We also plan to analyze environments where the designer's preferences are not perfectly aligned with the genuine agents'. In this scenario, it is plausible that the trolls' desire to hurt the designer could inadvertently improve the genuine agents' welfare. For example, in an auction setting, saboteurs could decrease the auctioneer's expected revenue, which may be beneficial to the genuine bidders.

Literature Review Our focus in this chapter is on the design of *polling* mechanisms in the presence of *adversarial saboteurs* among the voters. To our knowledge, there are papers that incorporate a strict subset of these considerations, but not all of them. For instance Chorppath and Alpcan (2011), Liu et al. (2017), Yang et al. (2017), Brahma et al. (2022), and Jiang et al. (2022) focus on mechanism design with malicious/adversarial agents in *non*-voting settings. In the literature on electoral competition, (Invernizzi, 2020) studies sabotage *within* parties and (Hirsch and Kastellec, 2022) studies studies sabotage between parties. We instead focus on the *other* side of the ballot-box (voters) in *polls* (i.e. without strategic candidates).

This chapter is also related to the literature on the design of false name-proof[4] mechanisms when *anonymous* agents can participate more than once (e.g. by creating multiple identifiers, botting, etc.). Such mechanisms were originally studied in *combinatorial-auction* settings (Yokoo (2003, 2008), Yokoo et al. (2001, 2004, 2006), and Rastegari et al. (2007)) and only more recently in voting games albeit without saboteurs

[4]I.e. mechanisms where agents do not have an incentive to participate more than once, even if they are able to do so.

(Conitzer (2008), Bachrach and Elkind (2008), Aziz et al. (2011), Elkind et al. (2011), and Fioravanti and Massó (2022)).

This chapter is also related to Lambert and Shoham (2008), who studies how to design a *survey* mechanism that elicits truthful opinions, and to Gary-Bobo and Jaaidane (2000), who study a polling mechanism design problem. However, neither paper allows for saboteurs amongst the voting population as in this chapter. Sabotage is a consideration more often seen in (dynamic and static) contests (discussed in Chowdhury and Gürtler (2015)). Most relevant is Ishida (2012), which considers the problem of designing sabotage-proof dynamic contests.

2.2 Illustration

Suppose there are three agents, $i \in \{1, 2, 3\}$. Agents 1 and 2 are "genuine" or "normal" (interchangeable) agents. Each of them has a type $\theta_i \in \{\gamma_1, \gamma_2\}$, with an i.i.d. prior distribution characterized by $\mathbb{P}(\theta_i = \gamma_1) = p$, where $p \in (0; 1)$. For simplicity, assume $\gamma_1 = 1$ and $\gamma_2 = 2$. Agent 3 is a "troll" or "saboteur" (interchangeable) and will be described below.

There is a mechanism designer who wants to maximize the well-being of agents 1 and 2. The utility function of an agent of type θ_i is given by

$$u_i(a, \theta_i) = -(a - \theta_i)^2,$$

where $a \in \mathbb{R}$ is the *expected* action taken by the designer.[5] The objective function of the designer is given by

$$V(a) = \mathbb{E}\left[\sum_{i=1}^{2} u_i(a, \theta_i)\right].$$

[5]Notice that these agents are risk *neutral*. If a had instead represented the designer's *realized* action, these agents would be risk *averse*.

Agent 3 is a "troll", or a "saboteur", whose goal is to reduce the well-being of agents 1 and 2 as much as possible. He does not know the agents' types, but knows the prior distribution, just as the mechanism designer. In a sense, his objective is entirely opposite of the designer's objective.[6] Hence, no matter what mechanism the designer creates, agent 3 will participate in a way that ex-ante minimizes $V(a)$. Agents 1 and 2 are aware of this and can take it into account when choosing how to behave.

In order to maximize the well-being of agents 1 and 2, the designer can create a *mechanism*, which consists of a message set M and an outcome rule $x : M^3 \to \mathbb{R}$. We limit our attention to direct mechanisms in which agents report their types, i.e. $M = \{1, 2\}$. The choice of a mechanism then boils down to choosing the outcome function.

There are two baseline mechanisms we will consider. The first will be referred to as "majority rule", where the designer implements the action equal to the mode of observed messages (and randomizes in case of a tie). It closely matches the design of online polls discussed in the Introduction. The second mechanism will be referred to as "average-of-messages rule", where the designer implements the action equal to the average of observed messages.[7] This mechanism is optimal in the absence of trolls, given the quadratic-loss utility function of the normal agents.

As we will see, both mechanisms have their comparative strengths and weaknesses when it comes to solving the designer's problem. In short, the majority rule is less susceptible to the influence of trolls, since they have to be pivotal in order to affect the outcome. On the other hand, the average-of-votes rule incorporates more information from the normal agents and has the potential to better match the average type of the agents. However, that potential can be limited by the increased influence of trolls, who no longer need to be pivotal in order to affect the outcome. In fact, we

[6] Note that the designer does not care about the troll's well-being. One way to interpret this assumption is that the troll comes from outside the population of agents that the designer cares about, e.g. a foreigner participating in a poll about purely domestic matters.

[7] Recall that $\gamma_1 = 1$ and $\gamma_2 = 2$, so messages are just real numbers.

show that this mechanism might perform worse than doing no mechanism at all — a "blind mechanism" benchmark, where the designer always takes the ex-ante best action. This occurs when there is little information that can be gained through the poll to begin with, in which case the negative influence of trolls outweighs the positive gain from more information.

The next two subsections outline a detailed analysis of the two baseline mechanisms.

Majority rule

As implied by its name, this outcome rule selects the *mode* of the received messages when the mode is unique; otherwise, we assume that the outcome rule uniformly randomizes between the choices tied for first. Formally, if $\mathbf{m}$ represents the vector of observed messages, then

$$x(\mathbf{m}) \equiv \mathcal{U}\{\text{mode}(\mathbf{m})\}. \tag{2.1}$$

How do the genuine agents and the troll behave in this mechanism?

Lemma 1. *Fix mechanism $M = \{1, 2\}$ and $x(\mathbf{m}) \equiv \mathcal{U}\{mode(\mathbf{m})\}$, and assume that genuine agents always break indifference in favor of telling the truth. Then any BNE of the resulting game is for genuine agents to tell the truth and for the troll to tell either $m_3 = 1$ or $m_3 = 2$ (he is indifferent).*

The proof of this lemma can be found in Appendix A.2.1. Intuitively, any agent's vote matters only when that agent is pivotal. For genuine agents, that means that their decision matters only when the other agent and the troll split votes, in which case the agent strictly prefers to tell the truth.[8] For the troll, his decision matters only when the

[8]Our assumption about the indifference-breaking rule eliminates nonsensical equilibria where all agents always say the same message.

genuine agents are (truthfully) splitting the vote, in which case he is indifferent between saying 1 (and hurting type 2) or saying 2 (and hurting type 1). Importantly, this is not the case when there is more than one troll, and we plan to consider this case in subsequent work.

Let us now consider the welfare implications of the "majority rule" mechanism. Given the BNE described in Lemma 1 (for argument's sake, assume $m_3 = 2$), the ex-ante welfare of agents is equal to

$$V_{MVW} = p^2 \cdot 0 + 2p(1-p) \cdot \left(-(0)^2 - (1)^2\right) + (1-p)^2 \cdot 0 = -2p(1-p) \tag{2.2}$$

Our benchmark for welfare is the no-poll scenario, under which the designer does not create any mechanism and simply takes the ex-ante best action. That action should maximize the objective function,

$$V_{NP}(a) = -p^2 \cdot 2\,(a-1)^2 - 2p(1-p) \cdot \left((a-1)^2 + (a-2)^2\right) - (1-p)^2 \cdot 2\,(a-2)^2\,.$$

which has the corresponding first order condition,

$$-4p^2(a-1) - 4p(1-p) \cdot (2a-3) - 4(1-p)^2(a-2) = 0$$

Solving this first order condition for a yields $a = 2 - p$ as the designer's ex-ante best action.[9]

Under this action, we can show that the agents' welfare (after a few algebraic

[9] Note that $\frac{\partial^2}{\partial a^2} V_{NP}(A) = -4p^2 - 8(1-p)p - 4(1-p)^2$

simplifications) is given by

$$\begin{aligned} V_{NP} &= -2p^2(1-p)^2 - 2p(1-p)((1-p)^2 + p^2) - 2(1-p)^2p^2 \\ &= -2p(1-p). \end{aligned}$$

Note that this is exactly the same as V_{MVW}. This implies that running a "majority rule" mechanism leads to the same welfare as running no mechanism at all! The presence of the troll completely nullifies the effectiveness of the mechanism in conveying information to the designer.

Average-of-votes

This outcome rule simply takes the average of the messages received by the mechanism designer. Formally,

$$x(\mathbf{m}) \equiv \frac{1}{N}\sum_{i=1}^{N} m_i \tag{2.3}$$

Similarly to the "majority rule" discussion, let us find the equilibrium in the resulting game between the genuine agents and the troll.

Lemma 2. *Fix mechanism $M = \{1, 2\}$ and $x(\mathbf{m}) \equiv \frac{1}{N}\sum_{i=1}^{N} m_i$, and assume that genuine agents always break indifference in favor of telling the truth. Then the unique BNE of the resulting game involves the genuine players telling the truth and the troll playing the following strategy:*

$$m_{3T}^* = \begin{cases} 1, \ \textit{if } p < \frac{1}{2} \\ \{1, 2\}, \ \textit{if } p = \frac{1}{2} \\ 2, \ \textit{if } p > \frac{1}{2}. \end{cases}$$

The proof of this lemma can be found in Appendix A.2.2. Intuition behind it

is similar to that of Lemma 1, with the exception that now the troll is not indifferent between messages because he is able to affect the outcome in all cases (as opposed to only the cases where he is pivotal in the "majority rule" mechanism).

Let us now consider the welfare implications of the "average-of-votes" mechanism. Assume $p > \frac{1}{2}$ (to pin down the exact message of the troll). Given the BNE described in Lemma 2 , the ex-ante welfare of the genuine agents is equal to

$$\begin{aligned} V_{AM} &= p^2 \cdot 2\left(-\left(\frac{1}{3}\right)^2\right) + 2p(1-p)\cdot\left(-\left(\frac{2}{3}\right)^2 - \left(-\frac{1}{3}\right)^2\right) + (1-p)^2 \cdot 0 \\ &= -\frac{2}{9}p^2 - \frac{10}{9}p(1-p) = -\frac{2}{9}p\,(5-4p)\,. \end{aligned}$$

Let us now compare it to the no-poll benchmark. We already know that under it the designer's problem is exactly the same as the one already considered in the "majority rule" discussion, so the optimal action is $a = 2 - p$ and the attained welfare is

$$V_{NP} = -2p(1-p).$$

When is this outcome better than the one provided by "average-of-votes" mechanism? Note:

$$V_{AM} < V_{NP} \;\Leftrightarrow\; -\frac{2}{9}p\,(5-4p) < -2p(1-p) \;\Leftrightarrow\; p > \frac{4}{5}.$$

That is, if p is sufficiently high, the no-poll benchmark (as well as the "majority rule" mechanism) provides higher welfare than the "average-of-votes" mechanism. Intuitively, this happens because there is little ex-ante uncertainty over the distribution of types of the genuine agents, meaning that the no-poll benchmark performs relatively well. This also means that there is little information that could potentially be gained from running the "average-of-votes" mechanism, while the negative effect of the troll's presence still remains in full force. Therefore, if ex-ante uncertainty over the type distribution

is sufficiently low, the "average-of-votes" mechanism loses to the no-poll benchmark and to the "majority rule" mechanism. On the other hand, if the ex-ante uncertainty is sufficiently high, there is a lot of information to be gained from the "average-of-votes" mechanism, so it is worth choosing it over the considered alternatives.

2.3 Model

In this section, we introduce a general model of voting mechanism design with a finite number of voters and trolls (or saboteurs). We focus on the case of two types in the interest of clearly presenting our results.[10] We will partially characterize the optimal mechanism by showing that it must satisfy a particular "indifference property" for the saboteurs. We will also analyze the performance of two benchmark mechanisms — *majority rule* and *average-of-votes rule*. Finally, we will finish by discussing how these benchmark mechanisms can be improved in simple ways to account for the presence of the saboteurs.

2.3.1 Setting

There is a designer that can take a public action $a \in \mathbb{R}$ and N "genuine"/"normal" voters. Each voter i has a type $\theta_i \in \{\gamma_1, \gamma_2\} =: \Gamma$ that is i.i.d. with $\mathbb{P}(\theta_i = \gamma_1) = p$. This is common knowledge. A voter of type θ has a standard quadratic-loss utility function

$$u(a, \theta) = -(a - \theta)^2.$$

We make the assumption of a specific functional form for tractability of analysis. In general, we can assume any single-peaked utility function (so that it has a bliss point).

[10] Our main results in subsections 2.3.2 and 2.3.3 can be straightforwardly extended to the case with an arbitrary, finite number of types.

The designer's objective function is to maximize the expected aggregate welfare of the genuine voters voters:

$$U(a) = \mathbb{E}\left[\sum_{i=1}^{N} u(a, \theta_i)\right].$$

In addition to the N genuine agents, the voting population also contains T trolls (or saboteurs). Each troll agent has objective function that is diametrically opposed to that of the designer:

$$u_T(a) \equiv -\mathbb{E}\left[\sum_{i=1}^{N} u(a, \theta_i)]\right] = -U(a).$$

Hence, their goal is to minimize the designer's objective function.

In order to choose a, the designer picks a voting mechanism which specifies a message set M and an outcome rule $g : M^{N+T} \to \mathbb{R}$. We will focus on *direct* mechanisms that allow voters to submit a report of their type. Formally, $M = \{\gamma_1, \gamma_2\}$. The messages of voters and trolls are indistinguishable to the designer, but she knows that there are N voters and T trolls. Given that there are two types, we can express a mechanism's outcome rule as a mapping $g : \{0, 1, \ldots, N + T\} \to [\gamma_1, \gamma_2]$ from the number of votes for $\theta = 1$ into an outcome $a \in [\gamma_1, \gamma_2]$.[11]

The timing of the model is as follows. Nature draws the types of the voters, $\{\theta_i\}_{i=1}^{N}$. The designer announces and commits to a mechanism g. Voters and trolls submit messages to the mechanism. The outcome is picked according to g and submitted messages, and payoffs realize.

Before we proceed to the analysis of the optimal mechanism, it is useful to analyze a benchmark case where there are no trolls. Suppose $T = 0$. The designer then faces a straightforward problem of eliciting types of the voters and picking the best action. Given that the voters vote sincerely, the designer will observe $\{\theta_i\}_{i=1}^{N}$. Maximizing

[11] We allow for "compromise" outcomes, in which the designer picks action $a \in (\gamma_1, \gamma_2)$.

aggregate welfare:

$$\max_{a\in[1,2]} -\sum_{i=1}^{N}(a-\theta_i)^2 \implies a^* = \frac{1}{N}\sum_{i=1}^{N}\theta_i.$$

In other words, the optimal mechanism without the trolls is the average-of-votes. Hereafter, we will describe a mechanism by its outcome rule $g(k)$ for all $k \in \{0, 1, \dots, N+T\}$, where k is the number of votes for the lower type, γ_1.

Lemma 3. *If $T = 0$, the optimal mechanism is the average-of-votes rule, i.e. the outcome rule is*

$$g(k) = \frac{k}{N}\gamma_1 + \frac{N-k}{N}\gamma_2.$$

It is useful to examine what happens to the performance of the average-of-votes mechanism (denoted by g_{av}). We will also compare its performance to that of a "blind mechanism" (g_b), which ignores the votes and always picks the ex-ante best action. Using the prior, we can find that action to be

$$g_b(k) = \arg\max_a \mathbb{E}_{\theta_i}\left[\sum_{i=1}^{N} -(a-\theta_i)^2\right] \implies g_b(k) = p\gamma_1 + (1-p)\gamma_2.$$

Suppose there are some trolls, $T \geq 1$. The average-of-votes mechanism is now defined as

$$g_{av}(k) = \frac{k}{N+T}\gamma_1 + \frac{N+T-k}{N+T}\gamma_2.$$

We can show that the trolls' best strategy under this mechanism is to vote for the ex-ante *less likely* type. For concreteness, assume $p > \frac{1}{2}$, which makes γ_2 the less likely type.

Lemma 4. *Assume $p > \frac{1}{2}$. Under g_{av}, the trolls optimally vote for $\theta = \gamma_2$.*

Since the average-of-votes mechanism does not account for the trolls' presence, the designer's expected welfare is smaller than when $T = 0$. However, sometimes trolls not only reduce the effectiveness of the mechanism, but can completely overturn any

welfare improvement that it generates in their absence.

Lemma 5. *The expected welfare under g_{av} is strictly lower than under the blind mechanism g_b if and only if*

$$p > \frac{N + 2T}{N + 2T + T^2} =: \bar{p}.$$

Note that if $T = 0$, this inequality turns into $p > 1$. This is expected: without trolls, the average-of-votes mechanism will always outperform the blind mechanism. The lemma also sheds light on the circumstances when ignoring information from the voters can be beneficial. When p is large (close to 1), there is little ex-ante uncertainty about the average type in the voter population. As a result, there is little benefit in gathering information about voters' types in the first place. This leads a mechanism that doesn't put any weight on the prior to perform worse than picking the ex-ante best action.

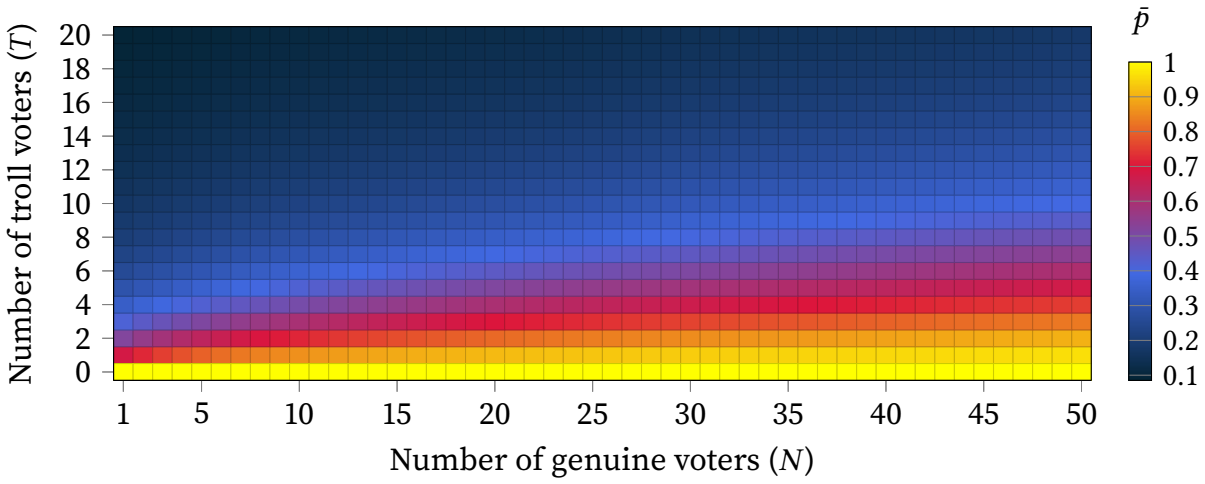


Figure 2.2: How $\bar{p} := \frac{N+2T}{N+2T+T^2}$ from Lemma 5 varies with N and T.

Figure 2.2 demonstrates how $\bar{p}$ depends on N and T. If we focus on a specific level curve, we can observe that when T needs to increase at a slower rate than N to maintain the same level of $\bar{p}$. Technically, T needs to grow at the rate proportional to

$\sqrt{N}$.

2.3.2 Properties of the Optimal Mechanism

This section derives two properties of the optimal mechanism in the given setting — *quasi-monotonicity* and an *indifference condition*. The first property puts reasonable bounds on the optimal mechanism and allows us to rule out the majority rule as the optimal mechanism. The second property puts a strict restriction on the optimal mechanism that generally rules out the "average-of-votes" mechanism. Additionally, we can use the insights of the indifference condition to suggest simple improvements to both of these benchmark mechanisms.

Suppose that the designer observes k votes given for type $\theta = \gamma_1$. Given that there are T trolls in total, the true number of genuine agents of type $\theta = \gamma_1$ can be from $\max\{k-T, 0\}$ to $\min\{k+T, N\}$. This allows us to put some reasonable bounds on the optimal mechanism's outcome rule. The following proposition describes them formally.

Proposition 1. *If g is an optimal mechanism, then for any $k \in \{0, 1, \dots, N+T\}$*

$$\frac{k}{N}\gamma_1 + \frac{N-k}{N}\gamma_2 \;\le\; g(k) \;\le\; \frac{\max\{k-T, 0\}}{N}\gamma_1 + \frac{\min\{N-k+T, N\}}{N}\gamma_2.$$

This property can be described as quasi-monotonicity with respect to votes. Roughly speaking, the optimal mechanism's action should be increasing in the number of votes that are cast for type $\theta = \gamma_2$. There may be local non-monotonicity, but overall the outcome rule should fall within the given bounds. This property is clearly satisfied by the "average-of-votes" mechanism, but it is not satisfied by the "majority rule" mechanism. Therefore, we can conclude that the latter mechanism is not optimal under any prior p and any number of voters or trolls.

Can the "average-of-votes" mechanism be optimal, then? The next property of

the optimal mechanism sheds some light on this.

Proposition 2. *Under the optimal mechanism, the trolls are indifferent between sending* $m = \gamma_1$ *and* $m = \gamma_2$.

We refer to this as the indifference property, since the optimal mechanism has to keep the trolls indifferent between all messages. Intuitively, when the trolls are not indifferent, the designer can slightly adjust the outcome rule under the mechanism without changing the trolls' best reply. This allows the designer to improve aggregate welfare under that strategy (and potentially worsen it under other strategies). This bit-by-bit optimization remains possible until the designer reaches a mechanism in which the trolls are indifferent between sending either message. Notably, this result can be readily generalized to a setting with more than two types; then, the optimal mechanism must keep the trolls indifferent between *all* messages.

The indifference property has two main benefits. First, it allows us to rule out a lot of potential mechanisms and focus on a narrow family of those that keep the trolls indifferent. "average-of-votes" mechanism generally does not satisfy the property. The only situation where it does is where $p = 0.5$. In that case, the trolls are indifferent between sending either messages, and it turns out that the "average-of-votes" is the optimal mechanism in that case. However, in any other situation the mechanism is not optimal because it violates the indifference property.

Second, the indifference property is a useful tool for reducing the computational complexity of searching for the optimal mechanism. In a general setting with N voters, T trolls and k types, a mechanism has to specify k^{N+T} outcomes. Proposition 2 puts $\frac{k(k-1)}{2}$ equations that restrict these variables. As a result, it reduces the dimensionality of the set of mechanisms that one needs to search through.

The impact of the indifference property can be seen visually in Figure 2.3, which depicts the designer's expected utility for the case $N = 2$ and $T = 1$ and for various priors

p. Recall that in this case, a mechanism is characterized by $\{g(0), g(1), g(2), g(3)\}$, where $g(k)$ is the outcome conditional on observing k votes for γ_1. Correspondingly, the axes in each subfigure capture $g(1)$ and $g(2)$ through the weights placed on γ_1.[12]

We can observe that the designer's utility is generally higher along a dotted line in each subfigure. This dotted line depicts the set of mechanisms that satisfy the indifference property from Proposition 2. As can be seen, the designer's utility is generally higher the closer the mechanism is to the dotted line. Intuitively, the closer a mechanism is to making trolls indifferent, the better (on average) it is. We can also see the computational impact of the indifference property, which reduces the dimensionality of the set of candidate mechanisms from 2 (a square) to 1 (a line).

We are currently investigating the designer's problem under the indifference restriction in order to see whether we can explicitly derive the optimal mechanism.

[12] For instance, if the weight placed on γ_1 under 1 vote for γ_1 is equal to 0.4, that means $g(1) = 0.4\gamma_1 + 0.6\gamma_2$.

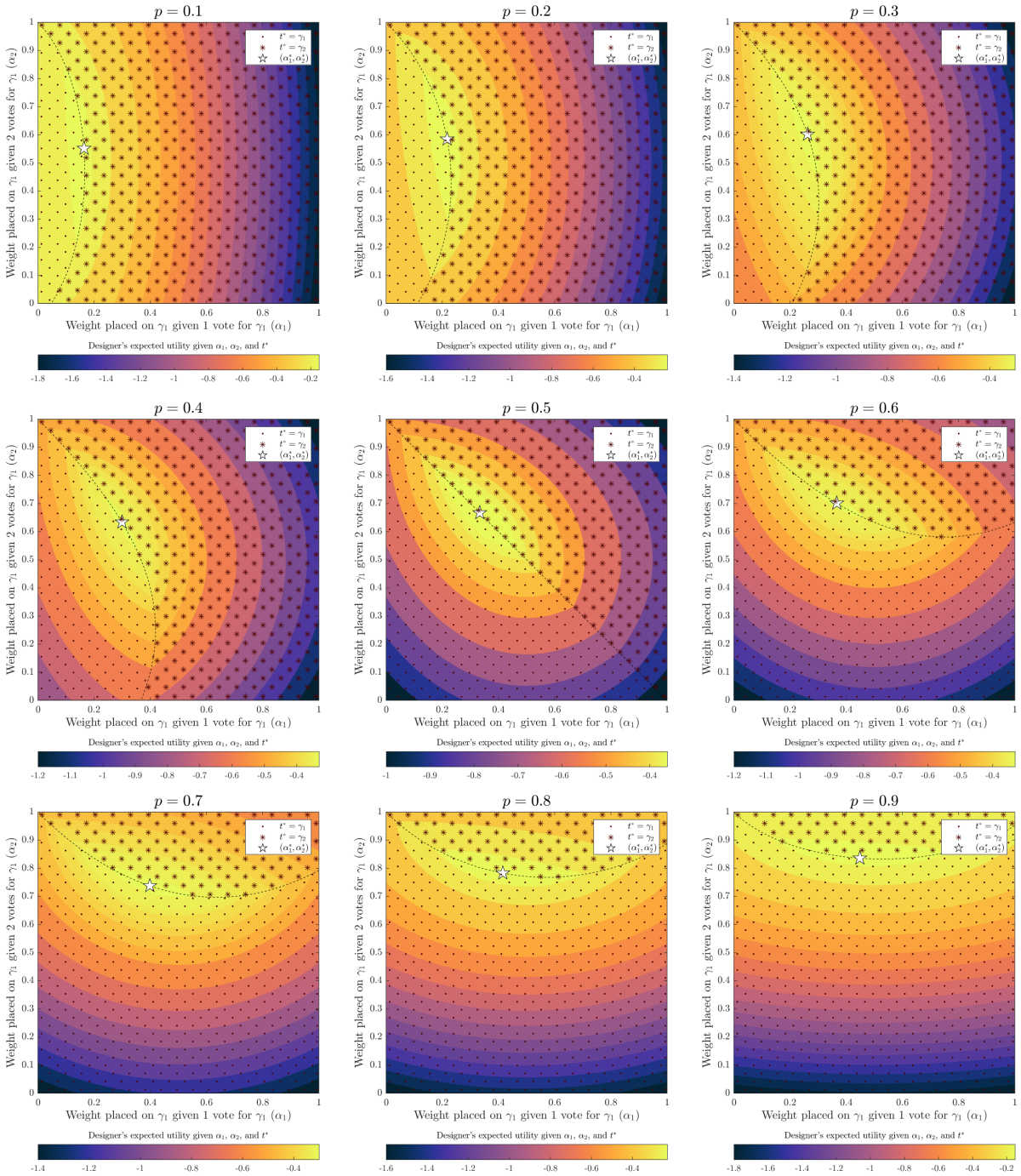


Figure 2.3: Graphical illustration of the optimal mechanism for the case with $N = 2$ genuine voters and $T = 1$ troll voter, for various levels of p. The white star denotes the optimal mechanism. The dotted line that the star is situated on is the family of mechanisms that make the troll indifferent between messages.

2.3.3 Improving Benchmark Mechanisms

In this section, we will propose ways to improve two benchmark mechanisms — majority rule and average-of-votes rule — by using the indifference property.

Under both mechanisms, the trolls' optimal strategy is the same. They can vote for the more likely type γ_1[13] or the less likely type γ_2. To decrease the expected aggregate welfare, trolls should vote for the *less likely* type. This introduce a *bias against the prior* into the mechanisms' outcomes. In order to improve aggregate welfare, the designer should tweak the mechanisms in a way that introduces *bias towards the prior*. We will show an intuitive way to do that for both benchmark mechanisms.

First, consider the majority rule. A common modification of this mechanism is a *supermajority* rule, which makes it harder to affect the outcome of the vote by small deviations in the vote distribution. One needs to also specify what happens if the supermajority is not reached, which we refer to as the *default option*. This is where the designer can put some bias towards the prior and offset the influence of the trolls. We will show that modifying the majority rule in this way can improve its expected welfare.

Formally, let g_{mr} be the majority-rule mechanism and $g_{smr}^{\alpha,x}$ be an α-supermajority rule with default outcome x. The outcome x is implemented if neither γ_1 nor γ_2 gets enough votes to meet the threshold α. Naturally, we only consider $\alpha > \frac{1}{2}$. Before we proceed to the result, recall that $p \geq \frac{1}{2}$, i.e., type $\theta = \gamma_1$ is ex-ante more likely in the voter population. The result will assume that there are at least 3 trolls in order to avoid a trivial case where changing the supermajority rule does not change the expected outcome.

Proposition 3. *Suppose $T \geq 3$. There exists $\overline{\alpha} > \frac{1}{2}$ such that the expected welfare under mechanism $g_{smr}^{\alpha,\gamma_1}$ is strictly larger than the expected welfare under g_{mr} for any $\alpha \in \left(\frac{1}{2}, \overline{\alpha}\right)$.*

Intuitively, the supermajority rule limits the trolls' influence in two ways. First,

[13] Recall that we assume $\mathbb{P}(\theta_i = \gamma_1) = p > \frac{1}{2}$.

it makes it less likely that their vote is pivotal. Second, it incorporates a bias towards the prior in its default option. Note that in the Proposition, the supermajority rule picks $a = \gamma_1$ (the more likely type from ex-ante perspective) as its default option. This works to counteract the trolls' influence in the cases where they were previously pivotal.

Now we turn our attention to the average-of-votes mechanism. The mechanism's performance suffers in a similar fashion to the majority rule—trolls vote for the less likely type and bias the outcome against the prior. One natural way to adjust the mechanism is by changing the weights assigned to each vote. In the benchmark mechanism, votes for γ_1 and γ_2 receive equal vote in determining the outcome. The designer can introduce a bias towards the prior by assigning a larger weight in the outcome to the vote for the more likely type, γ_1.

Formally, let $g_{am}(\beta)$ be the weighted-average-of-votes rule in which votes for γ_1 receive weight β and weights for γ_2 receive weight 1. For example, if there are k votes for γ_1 and $N + T - k$ votes for γ_2, the outcome under g_{am}^{β} would be

$$g_{am}^{\beta}(k) = \frac{k\beta}{k\beta + (N + T - k)}\gamma_1 + \frac{N + T - k}{k\beta + (N + T - k)}\gamma_2.$$

Note that $\beta = 1$ corresponds to the benchmark average-of-votes rule.

Proposition 4. *There exists $\overline{\beta} > 1$ such that the expected welfare under g_{am}^{β} is strictly higher than the expected welfare under g_{am}^{1} for any $\beta \in \left(1, \overline{\beta}\right)$.*

The weighted-average-of-votes rule counters the trolls' influence in a direct way—by decreasing the weight of votes for the troll-preferred option in determining the outcome.

2.4 Limit Environment: Trolls Ruin Everything

In this section, we will consider the effect of the trolls on a mechanism's outcome when the number of trolls becomes large. For this purpose, we will restrict attention to direct mechanisms that map distribution of votes into a distribution over outcomes. Put more formally, we consider a direct mechanism that is characterized by the outcome rule, $g : \Delta\Gamma \to \Delta\Gamma$.[14] Note that the argument of the mechanism is a distribution over votes (which are types due to directness of the mechanism), whereas the output of the mechanism is a distribution over outcomes (which are types by a reasonable assumption).[15]

Let $g(\Delta\Gamma)$ denote the set of possible distributions over outcomes that mechanism g can generate. Let

$$t = \min_{x \in g(\Delta\Gamma)} V(x) = \min_{x \in g(\Delta\Gamma)} \mathbb{E}\left[\sum_{i=1}^{N} u_i(x, \theta_i)\right]$$

be the worst utility (from ex-ante perspective) that the mechanism may generate for the designer. Let π_t be the distribution of votes that produces that outcomes, i.e. $V(g(\pi_t)) = t$. This is the ideal scenario for the trolls, given that their aim is to minimize the designer's objective function. The following result focuses on trolls' ability to manipulate the mechanism to produce that scenario.

Before we introduce the result, let us introduce some notations. We will call a mechanism g *continuous* if $g : \Delta\Gamma \to \Delta\Gamma$ is a continuous mapping. Let $p(\theta, T)$ denote a distribution of votes that is produced when normal agents have types $\theta = (\theta_1, \dots, \theta_N)$ and trolls' vote distribution is $\pi(T)$, which is defined as follows:

$$\pi(T) = \min_{\pi \in F(T)} |\pi - \pi_t|,$$

[14] Recall that $\Gamma = \{\gamma_1, \gamma_2\}$ denotes the type space.

[15] Alternative way to view this is to map a distribution of votes into an outcome that is in the span of Γ, when such an outcome can be defined. This is the case for the "average-of-votes" mechanism.

where $F(T) = \{\pi \in \Delta\Gamma \mid \forall \gamma_i,\ \pi(\gamma_i) = \frac{k}{T} \text{ for some } k \in \mathcal{N}\}$. Given these notations, we have the following result:

Proposition 5. *Fix a continuous mechanism $g : \Delta\Gamma \to \Delta\Gamma$. Then for any θ and any $\epsilon > 0$, there exists $\bar{T}$ such that if $T > \bar{T}$, we have $|V(g(\pi(\theta, T))) - t| < \epsilon$.*

This result can be interpreted as follows. Fixing any continuous mechanism, there will be a distribution of votes π_t that produces the (ex-ante) worst outcome under this mechanism. If trolls are sufficiently numerous, they can get the actual distribution of votes arbitrarily close to π_t no matter the distribution of normal agents' types. And hence, due to continuity of the mechanism, they can get its outcome $g(\pi(\theta, T))$ arbitrarily close to $g(\pi_t)$.

Letting T grow to potentially unlimited extent may seem implausible at first, since it relies on trolls being able to enter in large numbers at little to no cost. However, this is a common feature of open-access online polls that have been discussed in the Introduction as part of our motivation. Several of those polls have clear signs of "botting", which is a practice of creating dozens and hundreds fake accounts or entries in order to participate in a poll. In these circumstances, having T grow arbitrarily large is not a strange assumption, and may have actually contributed to the organizers' decision to shut down those polls, as we will see below.

Note that Proposition 5 places the designer into worst-case analysis territory of mechanism design. She knows that trolls, provided that they are sufficiently numerous, may get the outcome of any mechanism arbitrarily close to the worst case of that mechanism. Given this, she may evaluate mechanisms based on their worst case alone and pick the best mechanism based on that evaluation.

One option that is always open to the designer is to simply pick a distribution over outcomes that is best from the ex-ante perspective. We will refer to this as a *blind* mechanism, since it does not gather any information from the agents and picks a distribution

over outcomes based on the prior alone. Formally, a blind mechanism is characterized by function $g_b : \Delta\Gamma \to \Delta\Gamma$ such that $\forall \pi \in \Delta\Gamma,\ g_b(\pi) \in \arg\max_{x \in \Delta\Gamma} V(x)$.

Lemma 6. *Let g_b be a blind mechanism and g be any other continuous mechanism. Let π_t be the worst-case distribution of votes for g. Then $V(g(\pi_t)) \leq V(g_b(\cdot))$.*

This result easily follows from the definition of a blind mechanism. Since g_b always maps any distribution of votes into $\arg\max_{x \in \Delta\Gamma} V(x)$, and since $g(p_t)$ is the worst-case outcome for g, it must be that $V(g(p_t)) \leq \max_{x \in \Delta\Gamma} V(x)$. This implies $V(g(p_t)) \leq V(g_b(\cdot))$.

Lemma 6 indicates that when the designer expects too many trolls to participate in her mechanism, her best option might be to "shut down" the mechanism and run a blind one. This relates our analysis to motivational examples in the Introduction, where most online polls that were infiltrated by trolls were shut down by organizers. Our analysis provides theoretical rationale for such a decision, and also suggests when it is optimal. In particular, it is optimal to run a blind mechanism when costs of entering the mechanism are very low for trolls, i.e. in situations where trolls will be sufficiently numerous to bias the outcome towards the worst-case scenario.

2.5 Discussion and Next Steps

We have studied two baseline mechanisms of gathering information in the presence of trolls: "average-of-votes" and "majority rule". Moreover, the "majority rule" mechanism performs exactly the same as the no-poll benchmark, meaning that it does not give the designer any additional information that he can use. We have also shown that it is not clear which mechanism is comparatively better: if the ex-ante uncertainty over the best action is high, a more flexible option ("average-of-votes") is better, and vice versa. This suggests that more simple, rigid mechanisms may perform better in

the presence of trolls when there is relatively little ex-ante uncertainty over what the best action is. And if the opposite is the case, then more flexible mechanisms may perform better. We plan to study this conjecture in a more general setting and determine its truth value.

Another important question to ask is what role commitment plays in the presence of trolls. In classical papers on limited commitment, it has been shown that a mechanism designer generally performs better when he has access to full commitment. However, in the environment where some agents actively try to sabotage his mechanisms, the power to commit may actually *hurt* the designer's objective. If he commits to a certain mechanism, then trolls will be able to take full advantage of it. They might not be able to do so if the designer does not fully commit to a mechanism. This discussion can be related to the examples from Introduction, where many online polls were rejected or shut down after it was discovered that trolls had a major influence over the outcome. If the designer picks a mechanism where he has to best respond to his beliefs about the agents' types, it imposes an interesting constraint on the trolls — they have to conceal themselves among legitimate messages in order to avoid being "detected". We plan to formalize this analysis and determine whether it is indeed true that full-commitment mechanisms perform worse (on some metric) than limited-commitment mechanisms.

It is also important to consider environments where the designer's objective is not aligned with that of the normal agents, e.g. auction or classical principal-agent setting. Apart from doing the same analysis of trolls' influence on the welfare of normal agents, we could also now disentangle two alternative models of trolls' preferences — anti-agents and anti-designer. In the first case, the trolls wish to hurt the aggregate welfare of the normal agents, which is similar to our current example. In the second case, trolls wish to hurt the designer as much as possible, which could have non-trivial effects

on the welfare of normal agents.

To sum up, our future analysis includes studying full-commitment mechanisms more generally and check whether the insights we have gained so far hold there. In addition, we plan to compare full-commitment mechanisms with limited-commitment mechanisms and see whether the latter generally perform better against trolls' influence than the former. We also plan to extend analysis to the case where the designer's preferences are not aligned with preferences of normal agents.

Chapter 2 is currently planned for submission for publication of the material. Danil Dmitriev and the dissertation author, Frederick Aram Papazyan, are co-authors of this chapter.

Chapter 3

Strategic Misdirection

3.1 Introduction

From political interviews and debates to courtroom examinations and hearings, eliciting information about a state from others regularly requires deciding between multiple lines of inquiry (information sources) whose informativeness is *ex ante* uncertain. Take for instance a regulator who gauges the efficacy of a novel vaccine developed by pharmaceutical company, who discloses the results of various studies they ran. Suppose that each study measured the effect of their vaccine on a distinct metric, each of which may – or *may not* – be relevant to the vaccine's immunization ability due to the presence of a confounding factor that the company privately knows to be present/absent. Even if study results are verifiable – which prevents the company from lying about them – the relevance of these results may not be possible to verify. This poses a non-trivial challenge to the regulator: even if study results are verifiable, and all evidence is disclosed, to what extent can the regulator learn? Would it ever be beneficial for the company to fully disclose all of their results, even when they have a positive result they know is uninformative and a negative result they know to be informative?

To study these questions I consider an evidence disclosure game played by a Bayesian Sender and Receiver who have a common prior over a payoff-relevant state, a confound, and the signals produced over two information sources, which are all binary and have a commonly-known information structure. Given the realizations of the state and the confound, one source produces a perfectly informative signal of the state, while the other source produces a signal that is completely uninformative about the state (being drawn independently of the realization of the state). Only the Sender observes the realizations of each source's signal and the confound (hence, they know which source is ir/relevant).[1] After observing these realizations, the Sender chooses whether to disclose the realization from the relevant source, the irrelevant source, or the realizations from both sources. After observing what the Sender disclosed, the Receiver updates their beliefs using Bayes' rule whenever possible. Finally, the Receiver takes a real-valued action that affects the payoffs of both players. The Receiver is assumed to face a quadratic loss utility function (so that their bliss point is the state) while the Sender's utility function is a linear, strictly increasing function of the action chosen by the Receiver. (Hence, Senders always have an incentive to positively influence the Receiver's action choice.)

I show that it is possible to observe full disclosure of information sources in perfect Bayesian equilibrium. However, I also show that in any perfect Bayesian equilibrium, Senders who see opposing draws of each sources signal must pool. Hence, when sources produce signals that oppose one another, the Receiver necessarily does not update about the relevance of each source in equilibrium (in the sense that their posterior belief about the confound is equal to their prior). Consequently, it is possible for the Receiver's posterior belief over the state to be equal to their prior, even after all information is disclosed.

[1] Hence, they can perfectly infer the realization of the state.

Related Literature This chapter is related to Liang and Mu (2020), who study how myopic agents can fall into "learning traps" when the informativeness of signals is *ex ante* uncertain. The main departure of this model from Liang and Mu (2020) is that the Receiver above relies on a *strategic* information provider (namely, the Sender). In contrast, the information provider in their model is non-strategic in that it always reveals the source requested by the Receiver. A related framework is Liang et al. (2022). This is largely a special case of Liang and Mu (2020) where confounding variables are absent (as a result, optimizing agents always learn efficiently).

The communication game I consider also has parallels with Bull and Watson (2019) and Shishkin (2021). In those papers, a Sender possibly observes an information source and then chooses whether to reveal it or to stay silent. Another paper that studies signals of uncertain informativeness is Acemoglu et al. (2016); they show that this causes asymptotic agreement between two Bayesian learners to become fragile.[2] Hendricks and McAfee (2006) is related in that it focuses on a specific type of diversionary tactic known as a feint. Finally, Sobel (2020) is relevant to this chapter through its definition of "deception."

Liang and Mu (2020) is part of a burgeoning collection of papers that study the attentional misdirection of agents who learn "in isolation."[3] These papers focus on how agents can be misdirected due to myopia (Liang and Mu, 2020), misspecification Spiegler (2016, 2020, 2021), or the use of improper updating rules (Schwartzstein (2014) and He (2018)). The results of Koçak (2018) also heavily depend on assuming agents use improper updating rules, but *does* feature a Sender, distinguishing itself from the aforementioned papers. That being said, the model he considers is qualitatively quite

[2]Note that in contrast to their paper, a common-prior assumption *is* made below for simplicity. Relaxing this assumption would likely have a highly non-trivial and interesting effect on the Sender's ability to influence the Receiver using irrelevant information; I plan to explore this in the future.

[3]I.e. without a strategic intermediary – such as the Sender of this chapter – between the agent and information sources.

different from the one considered here.

As a disclosure game with verifiable information, this chapter is related to Milgrom (1981), Grossman (1981), Shishkin (2021), Martini (2022), and Titova (2022). Apart from the information structure of the game considered here, the setup I consider is standard in this literature.

This chapter is also related to the literature on narratives (Glazer and Rubinstein (2021), Eliaz and Spiegler (2020), Bénabou et al. (2019), and Lang (2020)) as well as to papers that study how models are used to create spurious correlations (Spiegler et al., 2021) or to persuade (Schwartzstein and Sunderam, 2021). Bayesian networks – which are considered in most of the just aforementioned papers – are also relevant to the present chapter, since the unknowns and sources in this chapter form a Bayesian network. Pearl (1985, 2000) are seminal works of this literature. The type of Bayesian network considered in this chapter is somewhat similar to the commonly-used "Noisy-OR gate" class of Bayesian network. Oniśko et al. (2001) consider the problem of learning the parameters of a Bayesian network with a small data set using a Noisy-OR gate.

The rest of this chapter is structured as follows: section 3.2 establishes the model, section 3.3 contains the definition of perfect Bayesian equilibrium and the results of this chapter, and section 3.4 contains concluding remarks. Proofs of all of the results are found in Appendix A.3.

3.2 Model

Two players named Sender and Receiver play a one-shot strategic communication game. The following information structure is common-knowledge: first, a (*payoff-*

relevant) *state* $\omega \in \{L, H\}$ is drawn according to the following probabilities:

$$\mathbb{P}\{\omega = H\} \equiv \mu \in (0,1); \quad \mathbb{P}\{\omega = L\} \equiv 1 - \mu. \tag{3.1}$$

I assume that $-\infty<L<H<\infty$; I interpret L as a Low draw and H as a High draw. Next, a *confounding variable* $C \in \{1, 2\}$ is drawn (independently of the state ω) according to the following probabilities:

$$\mathbb{P}\{C = 1\} \equiv \gamma_1 \in (0,1); \quad \mathbb{P}\{C = 2\} \equiv \gamma_2 = 1 - \gamma_1. \tag{3.2}$$

Afterwards, two information sources $i \in \{1, 2\}$ each produce a signal $X_i \in \{L, H\}$ that are conditionally independent given (ω, C),[4] and are drawn according to the following *conditional* probability distribution, where I assume that $\nu_i \in (0,1)\ \forall i \in \{1, 2\}$.

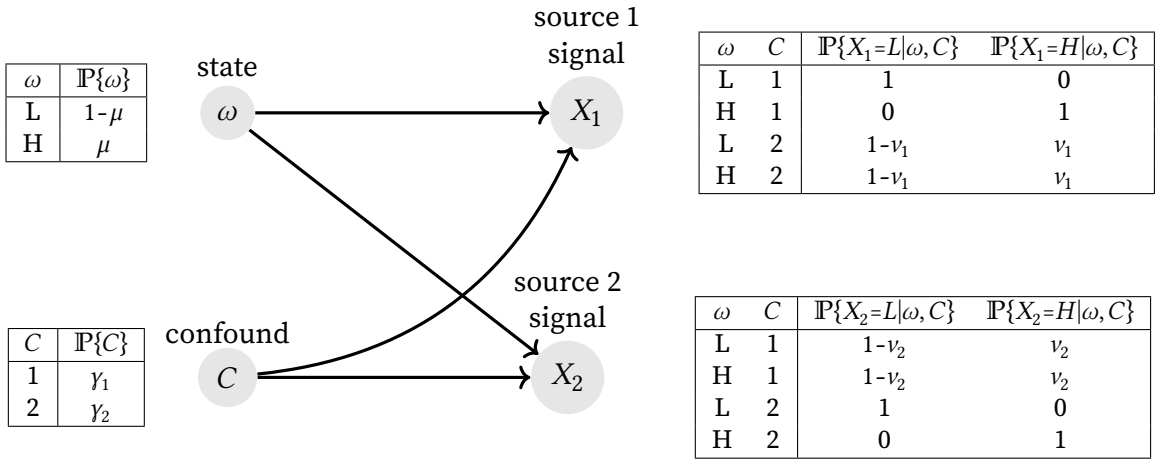
Table 3.1: Conditional probability distribution of source i's signal, X_i, given state ω and confound C, where $i \in \{1, 2\}$.

		$\mathbb{P}\{X_i{=}L\|\omega, C\}$	$\mathbb{P}\{X_i{=}H\|\omega, C\}$
$\omega = L$	$C = i$	1	0
$\omega = H$	$C = i$	0	1
$\omega = L$	$C \neq i$	$1\text{-}\nu_i$	ν_i
$\omega = H$	$C \neq i$	$1\text{-}\nu_i$	ν_i

Notice that if $C = c$, then source $i = c$ produces a *perfectly informative* signal for ω while the other source $i' \neq c$ produces a signal that is *completely uninformative* about ω. Hence, C pins down which source is *relevant* to learning about the state, ω, and which sourse is *irrelevant*.

Before (ω, C, X_1, X_2) are drawn, players have a common prior $\pi(\cdot)$ over (ω, C, X_1, X_2)

[4] I.e. $\mathbb{P}\{X_1, X_2|\omega, C\} = \mathbb{P}\{X_1|\omega, C\} \cdot \mathbb{P}\{X_2|\omega, C\}\ \forall (X_1, X_2, \omega, C)$, where $\mathbb{P}\{X_i|\omega, C\}$ is given in Table 3.1.



ω	$\mathbb{P}\{\omega\}$
L	$1-\mu$
H	μ

C	$\mathbb{P}\{C\}$
1	γ_1
2	γ_2

ω	C	$\mathbb{P}\{X_1=L\|\omega,C\}$	$\mathbb{P}\{X_1=H\|\omega,C\}$
L	1	1	0
H	1	0	1
L	2	$1-\nu_1$	ν_1
H	2	$1-\nu_1$	ν_1

ω	C	$\mathbb{P}\{X_2=L\|\omega,C\}$	$\mathbb{P}\{X_2=H\|\omega,C\}$
L	1	$1-\nu_2$	ν_2
H	1	$1-\nu_2$	ν_2
L	2	1	0
H	2	0	1

Figure 3.1: Graphical representation of the conditional in/dependence structure of the state ω, confound C, and source signals X_1 and X_2, using a Bayesian network.

that conforms to the above description; formally, it is given by

$$\pi(\omega{=}\hat{\omega}, C{=}c, X_1{=}x_1, X_2{=}x_2) \equiv \mathbb{P}\{\omega{=}\hat{\omega}\}\mathbb{P}\{C{=}c\}\prod_{i=1}^{2}\mathbb{P}\{X_i{=}x_i|\omega{=}\hat{\omega}, C{=}c\} \tag{3.3}$$

where $\mathbb{P}\{\omega = \cdot\}$ and $\mathbb{P}\{C = \cdot\}$ and respectively given in equations (3.1) and (3.2), and $\mathbb{P}\{X_i = \cdot|\omega = \cdot, C = \cdot\}$ is given in Table 3.1. The Receiver does not directly observe ω,C,X_1 or X_2.

The timing of the game proceeds as follows. First, state $\omega \in \{L, H\}$ and confound $C \in \{1, 2\}$ are randomly (and independently) drawn according to equations (3.1) and (3.2), respectively. Given realization $(\omega, C) = (\hat{\omega}, c)$, source $i \in \{1, 2\}$ then produces signal $X_i \in \{L, H\}$ according to the conditional probabilities in Table 3.1. Let x_i denote the realization of X_i. R does not directly observe $\omega, C, X_1,$ or X_2. S observes realizations of each source signal $(X_1, X_2) = (x_1, x_2)$, and the realization of the confound $C = c$. I henceforth refer to $\theta = (x_1, x_2, c)$ as the Sender's *type*. After observing their type, the Sender forms a posterior belief $\pi^S(\cdot|x_1, x_2, c)$ over (ω, C, X_1, X_2) using Bayes' rule. Note that since the Sender knows the identity of the relevant source (c) and the realization

of each source (x_1, x_2) they can infer that $\omega = x_c$ with probability 1. Therefore, upon observing their type, the Sender's posterior belief is degenerate:

$$\pi^s(\omega, C, X_1, X_2 | C{=}c, X_1{=}x_1, X_2{=}x_2\} = \mathbb{1}\{(\omega, C, X_1, X_2) = (x_c, c, x_1, x_2)\}. \tag{3.4}$$

After observing their type, S then chooses which signal(s) – if any – to disclose. Formally, they send a message m from the following set:

$$\mathcal{M}_{x_1,x_2} \equiv \{m_{x_1\emptyset}, m_{\emptyset x_2}, m_{x_1 x_2}\}. \tag{3.5}$$

I assume it is common knowledge that given $(X_1, X_1) = (x_1, x_2)$, the Sender can only send messages from the above set $\forall (x_1, x_2) \in \{L, H\}^2$. Since message $m = m_{x_1 x_2}$ is available to the Sender if and only if the Sender observed $(X_1, X_2) = (x_1 x_2)$, it is commonly understood as "$X_1 = x_1$ *and* $X_2 = x_2$." Message $m = m_{x_1\emptyset}$ is available to the Sender if and only if $(X_1, X_2) \in \{(x_1 L), (x_1, H)\}$, and is hence commonly understood as "$X_1 = x_1$." Similarly, message $m = m_{\emptyset x_2}$ is available to the Sender only if $(X_1, X_2) \in \{(L, x_2), (H, x_2)\}$, and is hence commonly understood as "$X_2 = x_2$."

Remark 3. Notice that source signal realizations are verifiable information, while the identity of the relevant source (pinned down by realization $C = c$) is not verifiable information. That is: the Sender cannot "lie" about signal realizations but they can disclose signal realizations they know to be uninformative about the state.[5]

After observing the Sender's message m, the Receiver forms a posterior belief $\pi^R(\cdot|m)$ over (ω, C, X_1, X_2) using Bayes' rule whenever possible. Let $\pi^R_\omega(\cdot|m)$ denote the

[5]For example, a Sender that observed $(X_1, X_2, C) = (x_1, x_2, 1)$ cannot send message $m_{z_1 z_2}$ if $z_i \in \{L, H\} \setminus \{x_i\}$ for any $i \in \{1, 2\}$, but they are free to send message m_{x_1,x_2} or $m_{\emptyset x_2}$.

marginal distribution of ω under posterior belief $\pi^R(\cdot|m)$, which is given by

$$\pi^R_\omega(\tilde{\omega}|m) \equiv \sum_{\tilde{c}=1}^{2} \sum_{\tilde{x}_1 \in \{L,H\}} \sum_{\tilde{x}_2 \in \{L,H\}} \pi^R(\omega{=}\tilde{\omega}, C{=}\tilde{c}, X_1{=}\tilde{x}_1, X_2{=}\tilde{x}_2|m). \tag{3.6}$$

Furthermore, let $\mu^R(m) \equiv \pi^R(H|m)$ denote the probability assigned to the event, $\omega = H$, under posterior belief $\pi^R(\cdot|m)$. After forming their posterior belief, R takes an action $a \in \mathbb{R}$; R and S then respectively earn the following payoffs:

$$u_R(a, \omega) = -(a - \omega)^2; \quad u_S(a) = a. \tag{3.7}$$

Therefore, given posterior belief π^R, R earns *expected* payoff

$$\mathbb{E}_{\omega \sim \pi^R_\omega(\cdot|m)} [u_R(\omega, a)] = (1 - \mu^R(m))[-(L - a)^2] + \mu^R(m)[-(H - a)^2]. \tag{3.8}$$

Since R faces a quadratic loss utility function, they essentially want to match the state in the sense that their expected utility (given posterior belief π^R) is maximized when $a = (1 - \mu^R(m))L + \mu^R(m)H = \mathbb{E}_{\omega \sim \pi^R_\omega(\cdot|m)}[\omega]$. However, S has a state-independent payoff and simply wants R to take as high an action as possible.

3.3 Equilibrium

The Sender's *disclosure strategy* $\sigma : \{L, H\}^2 \times \{1, 2\} \to \Delta\left(\mathcal{M}_{x_1,x_2}\right)$ maps their type $\theta \equiv (x_1, x_2, c)$ to a distribution $\sigma(\cdot|x_1, x_2, c)$ over messages m in $\mathcal{M}_{x_1,x_2}$. Since source signal realizations are verifiable information, we necessarily have $\sigma(m_{z_1 z_2}|x_1, x_2, c) = 0$ if $z_1 \in \{L, H\} \backslash \{x_1\}$ or $z_2 \in \{L, H\} \backslash \{x_2\}$. I now turn attention to the Receiver's beliefs and strategies.

Let the set of all messages that are possible for the Receiver to observe be denoted by

$$\mathcal{M} \equiv \bigcup_{(x_1,x_2)\in\{L,H\}^2} \mathcal{M}_{x_1,x_2} = \{m_{L\varnothing}, m_{H\varnothing}, m_{\varnothing L}, m_{\varnothing H}, m_{LL}, m_{LH}, m_{HL}, m_{HH}\}. \tag{3.9}$$

After observing message $m \in \mathcal{M}$, R forms a posterior belief $\pi^R(\cdot|m)$ over (ω, C, X_1, X_2) using Bayes' rule, whenever possible. When using Bayes' rule is possible, $\pi^R(\cdot|m)$ is given as follows:

$$\pi^R(\omega, C, X_1, X_2|m) = \frac{\sigma(m|x_1, x_2, c)\pi(\omega, C, X_1, X_2)}{\sum_{\tilde{\omega}\in\{L,R\}} \sum_{\tilde{c}\in\{1,2\}} \sum_{\tilde{x_1}\in\{L,R\}} \sum_{\tilde{x_2}\in\{L,R\}} \sigma(m|\tilde{x_1}, \tilde{x_2}, \tilde{c})\pi(\tilde{\omega}, \tilde{c}, \tilde{x_1}, \tilde{x_2})} \tag{3.10}$$

Since information about signal realizations is verifiable, note that R has a degenerate belief over X_i if it is disclosed by the Sender.[6] The Receiver's *action strategy*

$$\alpha : \Delta(\{L, H\}) \times \mathcal{M} \to \Delta(\mathbb{R})$$

maps the Receiver's posterior beliefs $\pi^R(\cdot|m)$ and the Sender's message m to a distribution over actions $a \in \mathbb{R}$.

The Sender's expected utility from playing disclosure strategy σ after observing $(X_1, X_2, C) = (x_1, x_2, c)$, keeping the Receiver's action strategy α fixed, is given by

$$V_s(\alpha, \sigma|x_1, x_2, c) = \mathbb{E}_{m\sim\sigma(\cdot|x_1,x_2,c)}\left[\mathbb{E}_{a\sim\alpha(\cdot|m,\pi^R(\cdot|m))}\left[u_S(a)\right]\right]. \tag{3.11}$$

The Receiver's expected utility from playing action strategy α after having observed mes-

[6] Put in more formal terms: the Receiver's posterior belief $\pi^R(\cdot|m)$ is degenerate over X_1 if they observe message $m \in \{m_{L\varnothing}, m_{H\varnothing}, m_{LH}, m_{LL}, m_{HL}, m_{HH}\}$ and it is degenerate over X_2 if they observe message $m \in \{m_{\varnothing L}, m_{\varnothing H}, m_{LH}, m_{LL}, m_{HL}, m_{HH}\}$.

sage $m \in \mathcal{M}$ is given by

$$V_R(\alpha, \sigma|m) = \mathbb{E}_{a\sim\alpha(\cdot|m,\pi^R(\cdot|m))}\left[\mathbb{E}_{\omega\sim\pi^R_\omega(\cdot|m)}\left[u_R(a,\omega)\right]\right]. \tag{3.12}$$

In this chapter, I focus on perfect Bayesian equilibrium which is defined as follows.

Definition 2. A *perfect Bayesian equilibrium* consists of a Sender's disclosure strategy σ^*, Receiver's action strategy α^*, and Receiver's posterior belief $\pi^R(\cdot|\cdot)$ over (ω, C, X_1, X_2) given $m \in \mathcal{M}$ such that

(i) $\sigma^*(\cdot|x_1, x_2, c) \in \arg\max_{\sigma(\cdot|x_1,x_2,c)\in\Delta(\mathcal{M}_{x_1x_2})} V_s(\alpha^*, \sigma|x_1, x_2, c)\ \forall x_1, x_2, c$

(ii) $\alpha^*\left(\cdot|m, \pi^R(\cdot|m)\right) \in \arg\max_{\alpha(\cdot|m,\pi^R(\cdot|m))\in\Delta(\mathbb{R})} V_R(\alpha, \sigma^*|m)\ \forall m, \pi^R(\cdot|m)$

(iii) $\pi^R(\cdot|m)$ is formed using Bayes' rule and prior belief π (in equation 3.3) whenever possible.

(iv) $\pi^R(\omega, C, X_1 \neq x_1, X_2|m) = 0$ if $m \in \{m_{x_1\emptyset}, m_{x_1L}, m_{x_1H}\}$ for some $x_1 \in \{L, H\}$, and
$\pi^R(\omega, C, X_1, X_2 \neq x_2|m) = 0$ if $m \in \{m_{\emptyset x_2}, m_{Lx_2}, m_{Hx_2}\}$ for some $x_2 \in \{L, H\}$.

In words: (i) says that the Sender's disclosure strategy maximizes their expected payoff given the Receiver's strategy and the Sender's type (x_1, x_2, c), for every type of Sender; (ii) says that, given the Sender's strategy, the Receiver's strategy maximizes their expected payoff, given any message they receive and any posterior belief they hold; (iii) says that the Receiver's posterior belief over (ω, C, X_1, X_2) given message m is calculated using Bayes' rule whenever possible; (iv) says that the Receiver's posterior beliefs on *and off* the equilibrium path are degenerate over any source signal that is revealed. This last condition ensures that whenever the Receiver observes, say, message $m_{H\emptyset}$, they are certain that $X_1 = H$, regardless of whether $m_{H\emptyset}$ was observed on or off the equilibrium path. This reflects the fact that it is common knowledge that source

signal realizations are verifiable in the sense that the Sender can only send messages from $\mathcal{M}_{x_1x_2}$ given that they observed $(X_1, X_2) = (x_1, x_2)$.

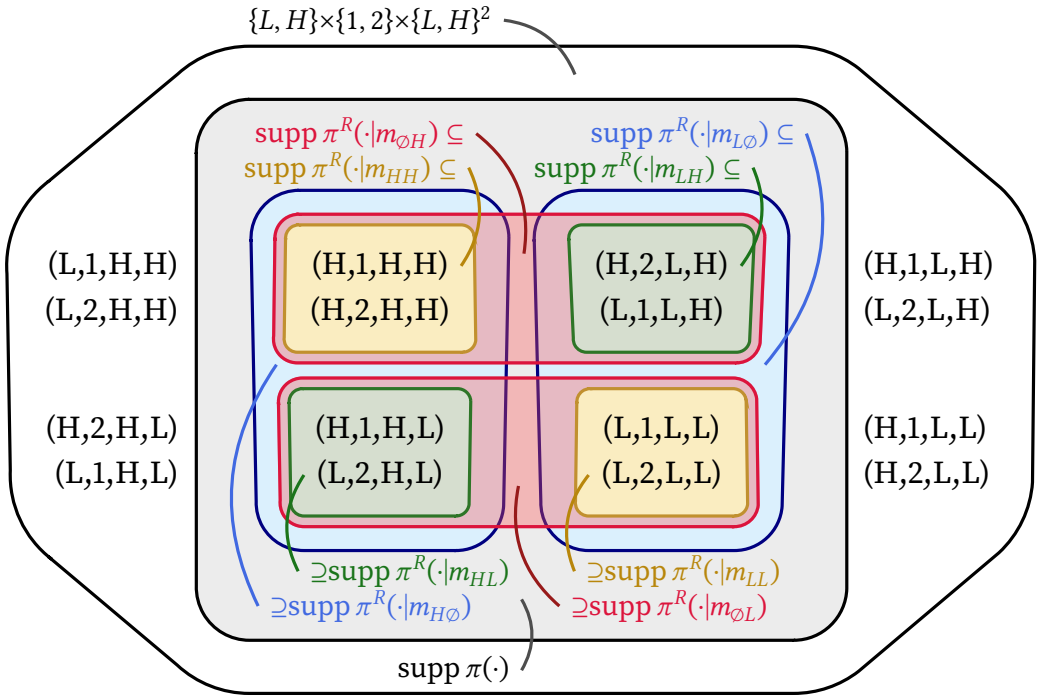


Figure 3.2: Visualization of the support of agents' common prior $\pi(\cdot)$ and how the support of the Receiver's posterior belief $\pi^R(\cdot|m)$ varies with $m \in \mathcal{M}$. Each quatruple above is an element $(\hat{\omega}, c, x_1, x_2)$ of $\{L,H\}\times\{1,2\}\times\{L,H\}^2$.

Figure 3.2 summarizes how observing (on- or off-path) message $m \in \mathcal{M}$ restricts the Receiver's posterior belief $\pi^R(\cdot|m)$ over (ω, C, X_1, X_2). This yields two observations that are useful to note before turning our attention to characterizing perfect Bayesian equilibria.

Observation 1. In any perfect Bayesian equilibrium, $\mu^R(m_{HH}) = 1$ and $\mu^R(m_{LL}) = 0$ must hold, regardless of whether $m \in \{m_{LL}, m_{HH}\}$ is sent on or off the equilibrium path.

While Figure 3.2 demonstrates the *mechanical* reason for why this observation is true, the intuition is as follows: when S discloses that both source signals are High (Low) draws, R can perfectly infer the state must be High (Low), because one of the sources must be relevant, and the relevant source is a perfectly informative signal of the state.

Observation 2. If $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$ is sent off the equilibrium path, then there exists a $\pi^r(\cdot|m)$ that satisfies condition (iv) of Definition 2 such that $\mu^R(m) = \psi$, for any chosen $\psi \in [0, 1]$.

Informally put, the above observation says that "anything goes," for the Receiver's posterior beliefs about the state after observing any message $m \in \setminus\{m_{LL}, m_{HH}\}$ off the equilibrium path. That is, when m is a message that corresponds to full disclosure of conflicting evidence (i.e. $m \in \{m_{HL}, m_{LH}\}$) or partial disclosure ($m \in \{m_{\varnothing H}, m_{\varnothing L}, m_{H\varnothing}, m_{L\varnothing}\}$). Notice that in Figure 3.2, R's posterior $\pi^R(\cdot|m)$ formed after observing each such message has a support that is permitted (by condition (iv) of Definition 2) to include H and L realizations of the state. The intuition is most easily seen in the case of full disclosure of conflicting evidence: after observing, say, message $m = m_{HL}$ off the equilibrium path, condition (iv) of Definition 2 requires that

$$\pi^R(H, 1, H, L|m_{HL}) + \pi^R(L, 2, H, L|m_{HL}) = 1.$$

Hence, $\mu^R(m_{HL})$ is pinned down by R's belief over the identity of the relevant source (which is determined by C). Notice that given the above equation, any $\mu^R(m_{HL}) \in [0, 1]$ can be achieved while complying with condition (iv) of Definition 2. As we will see below, the common-knowledge information structure will also yield structure to R's off path beliefs, which will allow for them to be interpreted in terms of the information structure.

Remark 4. Under the setup of this chapter, assuming that beliefs satisfy condition (iv) in Definition 2 is equivalent to assuming the *consistency* requirement from Kreps and Wilson (1982). This is shown in Lemma 7, which is found in appendix section A.3.2. Because of this equivalence, all perfect Bayesian equilibria of the game considered in this chapter are also *sequential equilibria* (which are defined in Kreps and Wilson (1982)).

In this chapter, I focus on pure strategy perfect Bayesian equilibrium, which I henceforth refer to simply as "equilibrium" in the interest of concision. Condition (ii) of Definition 2 thence requires $\alpha^*(\cdot|m, \pi^R(\cdot|m))$ to place probability 1 on action $a = a^*(\cdot|m, \pi^R(\cdot|m))$ given by

$$\begin{aligned} a^*(\cdot|m, \pi^R(\cdot|m)) &\equiv \arg\max_{a \in \mathbb{R}} \Big\{ \overbrace{\big[1 - \mu^R(m)\big]\left[-(L-a)^2\right] + \mu^R(m)\left[-(H-a)^2\right]}^{\mathbb{E}_{\omega \sim \pi^R_\omega(\cdot|m)}\left[u^R(a,\omega)\right]} \Big\} \\ &= \underbrace{\big[1 - \mu^R(m)\big]\, L + \mu^R(m) H}_{\mathbb{E}_{\omega \sim \pi^R_\omega(\cdot|m)}[\omega]} \end{aligned} \tag{3.13}$$

for every $m \in \mathcal{M}$ and $\mu^R(m)$, both on and off the equilibrium path. Therefore, condition (i) requires that $\sigma^*(\cdot|x_1, x_2, c)$ place probability 1 on message $m = s^*_{x_1 x_2 c}$ such that

$$s^*_{x_1 x_2 c} \in \arg\max_{m \in \mathcal{M}_{x_1 x_2 c}} \Big\{ \overbrace{\big[1 - \mu^R(m)\big]\, L + \mu^R(m) H}^{u_s\left(a^*(\cdot|m, \pi^R(\cdot|m))\right)} \Big\} = \arg\max_{m \in \mathcal{M}_{x_1 x_2 c}} \left\{ \mu^R(m) \right\} \tag{3.14}$$

holds for every (x_1, x_2, c). The above equality implies that in equilibrium each type (x_1, x_2, c) of Sender, behaves *as if* they choose the message $s = m^*_{x_1 x_2 c} \in \mathcal{M}_{x_1 x_2}$ that maximizes the Receiver's posterior belief $\mu^R(m)$ that the state is high after observing m, taking the messages chosen by other Senders as given.[7] For readers' convenience, probability tables for $\mu^R(\cdot)$ are provided in Tables A.1-A.6 in appendix section A.3.1.

In Proposition 8, I show that full disclosure of *source signals* is always possible in equilibrium, for any choice of model primitives.

Proposition 8. *There is a pure-strategy equilibrium where each type* (x_1, x_2, c) *sends message* $m_{x_1 x_2}$ *with probability 1.*

[7] Note that the equality in (3.14) holds because $\big[1 - \mu^R(m)\big]\, L + \mu^R(m) H = L + (H - L)\mu^R(m)$ and $H > L$.

The proof of this proposition is found in subsection A.3.3 of Appendix A.3, and the intuition is as follows: Types $(H, H, 1)$ and $(H, H, 2)$ each send m_{HH}. Since $\mu^R(m_{HH}) = 1$, neither type has a profitable deviation. Types $(L, L, 1)$ and $(L, L, 2)$ each send m_{LL}. Since $\mu^R(m_{LL}) = 0$, it follows that in order for neither $(L, L, 1)$ nor $(L, L, 2)$ to have a profitable deviation, we must have $\mu^R(m_{L\varnothing}) = 0 = \mu^R(m_{\varnothing L})$. Notice that these equalities are equivalent to requiring the following:

$$\text{supp}\, \pi^R(\cdot|m_{\varnothing L}) \subseteq \{(L, 1, L, L), (L, 2, L, L), (L, 2, H, L)\};$$

$$\text{supp}\, \pi^R(\cdot|m_{L\varnothing}) \subseteq \{(L, 1, L, L), (L, 2, L, L), (L, 1, L, H)\}.$$

In words, this is saying that when R observes partial disclosure of a single Low draw off the equilibrium path, they believe (with certainty) that the Sender is concealing a Low draw from the relevant source. Recalling Figure 3.2, this is equivalent to saying that R believes that the Sender is concealing a High draw from the relevant source with probability zero. Hence, R believes that a deviation to disclosing a single L realization must come from a type of S that observed an L realization of the relevant source.

Finally, we turn our attention to S types who saw conflicting signal realizations. Types $(H, L, 1)$ and $(H, L, 2)$ each send $m = m_{HL}$ with probability 1, and types $(L, H, 1)$ and $(L, H, 2)$ each send $m = m_{LH}$ with probability 1. As a result, we have

$$\mu^R(m_{HL}) = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1\text{-}\mu)};\ \mu^R(m_{LH}) = \frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu + \gamma_1\nu_2(1\text{-}\mu)}, \tag{3.15}$$

by Tables A.2 and A.1, respectively. Therefore, in order for neither $(H, L, 1)$ nor $(H, L, 2)$ to have a profitable deviation, we must have

$$\mu^R(m_{H\varnothing}) \leq \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1\text{-}\mu)} = \mu^R(m_{HL}). \tag{3.16}$$

Similarly, in order for neither $(L, H, 1)$ nor $(L, H, 2)$ to have a profitable deviation, we must have

$$\mu^R(m_{\emptyset H}) \leq \frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu + \gamma_1\nu_2(1\text{-}\mu)} = \mu^R(m_{LH}). \tag{3.17}$$

As mentioned earlier, the common-knowledge information structure of this game allows for an interpretation of these inequalities. Notice that one can rearrange the inequality $\mu^R(m_{H\emptyset}) \leq \mu^R(m_{HL})$ in (3.16) to arrive at the following equivalent equality:

$$\frac{\pi^R(L, 2, H, L|m_{HL})}{\pi^R(H, 1, H, L|m_{HL})} \leq \frac{\pi^R(L, 2, H, L|m_{H\emptyset})}{\pi^R(H, 1, H, L|m_{H\emptyset})}. \tag{3.16$'$}$$

Notice that the left-hand (right-hand) term in the above inequality is the *posterior odds ratio* of S's type being $(x_1, x_2, c) = (H, L, 2)$ to their type being $(x_1, x_2, c) = (H, L, 1)$ under the posterior belief R forms after observing on-path message m_{HL} (off-path message $m_{H\emptyset}$). In words, the inequality in (3.16′) says that when the Receiver observes only partial disclosure (of a High draw from the first source) off the equilibrium path, they believe that it is more likely that the Sender is concealing a Low draw from the relevant source than a Low draw from the irrelevant source. Hence, the Receiver believes that Senders of type $(H, L, 1)$ are more likely than those of type $(H, L, 2)$ to deviate from sending message m_{HL} to message $m_{H\emptyset}$.

Notice that the full source disclosure equilibrium necessarily involves pooling of types $(x_1, x_2, 1)$ and $(x_1, x_2, 2)$ $\forall x_1, x_2$, so that the Receiver's posterior belief over C is the same as their prior after every message sent on the equilibrium path. As mentioned just above, this does not affect the Receiver's ability to learn if $x_1 = x_2$: after observing m_{HH} (m_{LL}), the Receiver's posterior belief assigns probability 1 to the event $\omega = H$ ($\omega = L$). However, this *can* impede the Receiver's ability to learn when the Sender holds contradictory evidence ($x_1 \neq x_2$). In fact, the Receiver can potentially learn very little (or even nothing) about about the state after observing full disclosure of contradictory

evidence.[8]

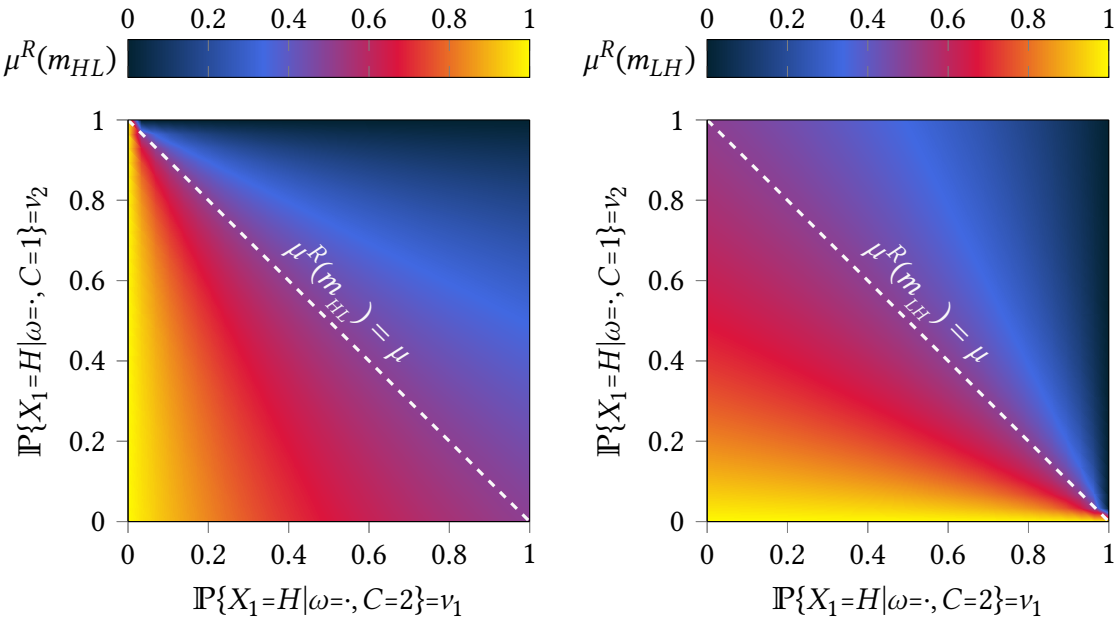


Figure 3.3: Heat maps visualizing $\mu^R(m_{HL})$ (left-hand side) and $\mu^R(m_{LH})$ (right-hand side) in the full source disclosure equilibrium when $\mu{=}0.5{=}\gamma_1{=}\gamma_2$, for various combinations of $(\nu_1, \nu_2) \in (0,1)^2$.

Figure 3.3 illustrates the extreme case where there is maximal *ex ante* uncertainty regarding the state ($\mu = 0.5$) and the identity of the ir/relevant source ($\gamma_1 = 0.5 = \gamma_2$). Notice that when $\nu_2 = 1 - \nu_1$, then the Receiver's posterior belief about the state after observing message m_{HL} or m_{LH} is equal to their prior about the state. Furthermore, when $\nu_2 = 1 - \nu_1$ and $\mu = 0.5 = \gamma_1$, the probability that the Receiver observes either of these messages on the equilibrium path is given by 0.5. That is, for these parameters, there is a 50% chance that the Receiver learns nothing about the state in equilibrium, despite full disclosure of evidence. This is of course a knife-edge case, but it turns out that there are a non-negligible set of "nearby" cases. In order to discuss this in more precise terms, a few definitions are needed.

Let the Kullback–Leibler divergence (henceforth referred to as "KL divergence")

[8] In the sense that $\mu^R(m_{HL})$ and $\mu^R(m_{LH})$ are each "close" to μ.

of the Receiver's prior and posterior belief about the state be denoted by

$$D(\mu^R(m), \mu) \equiv \left(1 - \mu^R(m)\right) \ln\left(\frac{1-\mu^R(m)}{1-\mu}\right) + \mu^R(m) \ln\left(\frac{\mu^R(m)}{\mu}\right).$$

Since KL divergence is a type of statistical distance, larger values of $D(\mu^R(m), \mu)$ correspond to larger differences between μ and $\mu^R(m)$. Hence, for each $\varepsilon > 0$, the inequality

$$\max\left\{D\left(\mu^R(m_{HL}), \mu\right), D\left(\mu^R(m_{LH}), \mu\right)\right\} < \varepsilon \tag{3.18}$$

is a condition that intuitively requires that the Receiver's posterior belief about ω after observing $m \in \{m_{LH}, m_{HL}\}$ to be sufficiently similar to their prior belief about ω. For each $\psi \in [0, 1]$, the inequality

$$\sum_{\hat{\omega} \in \{L,H\}} \sum_{c \in \{1,2\}} \sum_{(x_1,x_2) \in \{(L,H),(H,L)\}} \pi(\hat{\omega}, c, x_1, x_2) > \psi \tag{3.19}$$

is a condition that requires the probability of the event $X_1 \neq X_2$ to be larger than ψ. Notice that this is equal to the probability that the Receiver receives message $m \in \{m_{LH}, m_{HL}\}$ in the full source disclosure equilibrium.

For each (ψ, ε), let $\lambda(\varepsilon, \psi)$ denote the Lebesgue measure of the set

$$\Phi_{\psi\varepsilon} \equiv \left\{(\mu, \gamma_1, \nu_1, \nu_2) \in (0,1)^4 \ : \ \text{(3.18) and (3.19) both hold given } (\psi, \varepsilon)\right\}. \tag{3.20}$$

Intuitively, $\lambda(\varepsilon, \psi) > 0$ implies that given (ψ, ε), inequalities (3.18) and (3.19) hold for a non-trivial subset of parameters $(\mu, \gamma_1, \nu_1, \nu_2)$, and larger values of $\lambda(\varepsilon, \psi)$ correspond to larger sizes of this subset. As Figure 3.4 shows, $\lambda(\varepsilon, \psi)$ can be non zero when ε is small and ψ up to roughly 0.5. Hence, this illustrates that for a non-trivial subset of model parameters, there is a "not small" probability that Receiver learns only a "small" amount

of information about the state in the full-disclosure equilibrium.

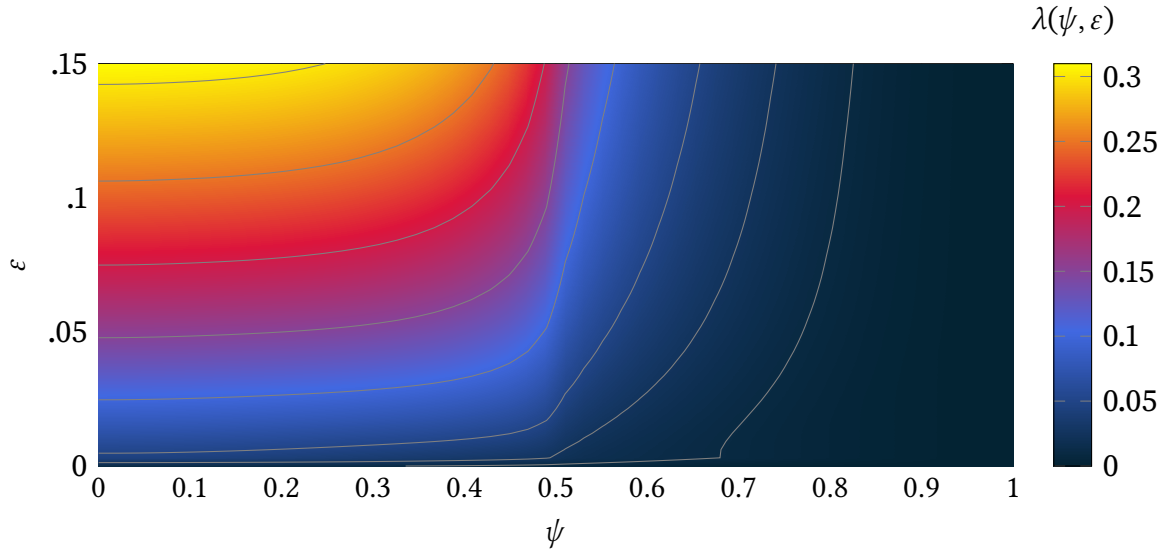


Figure 3.4: Numerically computed Lebesgue measure $\lambda(\psi, \varepsilon)$ of the set $\Phi_{\psi\varepsilon}$ defined in equation (3.20) for various combinations of (ψ, ε) in $(0, 1) \times (0, 0.25)$.

As I show in the next proposition, *all* equilibria of this game must have pooling among types $(H, L, 1)$ and $(H, L, 2)$ as well as pooling among types $(L, H, 1)$ and $(L, H, 2)$.

Proposition 9. *Let $(x_1, x_2) \in \{(L, H), (H, L)\}$ be arbitrarily fixed. Then, there does not exist any equilibrium where type $(x_1, x_2, 1)$ sends message $m \in \mathcal{M}_{x_1,x_2}$ with probability 1 and type $(x_1, x_2, 2)$ sends a different message $m' \in \mathcal{M}_{x_1,x_2} \setminus \{m\}$ with probability 1.*

The proof of this proposition is found in subsection A.3.3 of Appendix A.3. The proof of why $(L, H, 1)$ and $(L, H, 2)$ must pool in equilibrium is sketched below.[9]

Consider a pure strategy of the Sender where type $(L, H, 1)$ sends message $m_{\varnothing H}$ and type $(L, H, 2)$ sends message $m_{L\varnothing}$. If this strategy has type (H, H, r') send message $m_{\varnothing H}$ for some $r' \in \{1, 2\}$, then $\mu^R(m_{\varnothing H}) \in (0, 1)$. Intuitively, $\mu^R(m_{\varnothing H})$ must be strictly less than 1 because there is a strictly positive probability that $m_{\varnothing H}$ was sent by a Sender of type $(L, H, 1)$, who observed a Low draw of the relevant source (which implies that the

[9]The reasoning behind why $(H, L, 1)$ and $(H, L, 2)$ must pool in equilibrium is very similar.

state is Low with probability 1). However since $\mu^R(m_{HH}) = 1$ on and off the equilibrium path (by condition (iv) of Definition 2), it then follows from equation (3.14) that type (H, H, r') has a profitable deviation to message m_{HH} since $\mu^R(m_{\varnothing H}) < 1 = \mu^R(m_{HH})$.

If the aforementioned strategy has neither type $(H, H, 1)$ nor $(H, H, 2)$ sending message $m_{\varnothing H}$, then by construction type $(L, H, 1)$ is the only type that sends $m_{\varnothing H}$ (since type $(L, H, 2)$ sends message $m_{L\varnothing}$). Consequently, the Receiver can perfectly infer that the Sender's type is $(L, H, 1)$ after receiving message $m_{\varnothing H}$. Therefore, they can perfectly infer that $\omega = L$ (since this type observed a Low draw from the relevant source) so that $\mu^R(m_{\varnothing H}) = 0$ in this case. In contrast, $\mu^R(m_{L\varnothing}) > 0$ since message $m_{L\varnothing}$ is sent by type $(L, H, 2)$, who observed a High draw from the relevant source. Thus in light of equation (3.14), it is evident that type $(L, H, 1)$ has a profitable deviation to message $m_{L\varnothing}$ since $\mu^R(m_{H\varnothing}) = 0 < \mu^R(m_{\varnothing L})$.

For all other pure strategies where type $(L, H, 1)$ sends message $m \in \mathcal{M}_{LH}$ and type $(L, H, 2)$ sends a different message $m' \in \mathcal{M}_{LH} \setminus \{m\}$, it turns out that $\mu^R(m) < \mu^R(m')$, so that type $(L, H, 1)$ has a profitable deviation to the message m' sent by type $(L, H, 2)$. Intuitively $\mu^R(m) < \mu^R(m')$ holds in all of these cases because after observing message m (message m'), R's posterior belief shifts mass towards the event that S saw a Low (High) draw of the relevant source.

Notice that as a result of Proposition 9, it is not possible to have full unraveling in equilibrium. Moreover, the Receiver's posterior belief about C is *generically* equal to their prior with probability 1 given that the Sender observed conflicting signal realizations $(x_1, x_2) \in \{(L, H), (H, L)\}$, which is the case where knowing the identity of the relevant source is most crucial. Thus, the potential for the Receiver to learn very little about the state when the Sender observed conflicting signals is not just a feature of the full source disclosure equilibrium discussed in Proposition 8: it generically holds in any equilibrium of this game.

I now show that it is possible to observe partial disclosure of sources in equilibrium. The following proposition shows that there exist – for certain model parameters – equilibria where types $(L, H, 1)$ and $(L, H, 2)$ send message $m_{L\varnothing}$ with probability 1 and types $(H, L, 1)$ and $(H, L, 2)$ send message $m_{\varnothing L}$ with probability 1.

Proposition 10.

(a) There exists an equilibrium such that $\sigma^*(m_{L\varnothing}|\theta) = 1\ \forall\theta \in \{(L, H, 1), (L, H, 2), (L, L, 1), (L, L, 2)\}$ *and* $\sigma^*(m_{\varnothing L}|\theta) = 1\ \forall\theta \in \{(H, L, 1), (H, L, 2)\}$ *if and only if* $\frac{\gamma_1+\gamma_2(1-\nu_1)}{\gamma_2(1-\nu_1)} \leq \frac{\gamma_2\nu_1}{\gamma_1(1-\nu_2)}$.

(b) There exists an equilibrium such that $\sigma^*(m_{L\varnothing}|\theta) = 1\forall\theta \in \{(L, H, 1), (L, H, 2)\}$ *and* $\sigma^*(m_{\varnothing L}|\theta) = 1\ \forall\theta \in \{(H, L, 1), (H, L, 2), (L, L, 1), (L, L, 2)\}$ *if and only if* $\frac{\gamma_1\nu_2}{\gamma_2(1-\nu_1)} \geq \frac{\gamma_1(1-\nu_2)+\gamma_2}{\gamma_1(1-\nu_2)}$.

(c) There exists an equilibrium such that $\sigma^*(m_{L\varnothing}|\theta) = 1\forall\theta \in \{(L, H, 1), (L, H, 2), (L, L, 1)\}$ *and* $\sigma^*(m_{\varnothing L}|\theta) = 1\ \forall\theta \in \{(H, L, 1), (H, L, 2), (L, L, 2)\}$ *if and only if* $\frac{\gamma_2}{\gamma_1(1-\nu_2)} = \frac{\gamma_1}{\gamma_2(1-\nu_1)}$.

(d) There exists an equilibrium such that $\sigma^*(m_{L\varnothing}|\theta) = 1\forall\theta \in \{(L, H, 1), (L, H, 2), (L, L, 2)\}$ *and* $\sigma^*(m_{\varnothing L}|\theta) = 1\ \forall\theta \in \{(H, L, 1), (H, L, 2), (L, L, 1)\}$ *if and only if* $\frac{\gamma_2\nu_1}{\gamma_1(1-\nu_2)} = \frac{\gamma_1\nu_2+\gamma_2(1-\nu_1)}{\gamma_2(1-\nu_1)}$.

The proof for Proposition 10 is sound in appendix section A.3.5. Notice that the first two parts – where $(L, L, 1)$ and $(L, L, 2)$ *pool* – correspond to equilibria that exist for a set of model parameters with non-zero Lebesgue measure. In contrast, the latter two parts – where $(L, L, 1)$ and $(L, L, 2)$ *separate* – correspond to equilibria that exist only in a knife-edge case.

Before discussing the reasoning behind this result, it is useful to note that since types $(L, H, 1)$ and $(L, H, 2)$ pool on message $m_{L\varnothing}$ and types $(H, L, 1)$ and $(H, L, 2)$ pool on message $m_{\varnothing L}$, both

$$\mu^R(m_{L\varnothing}) \in (0, 1) \text{ and } \mu^R(m_{L\varnothing}) \in (0, 1) \tag{3.21}$$

hold regardless of the messages sent by types $(L, L, 1)$ and $(L, H, 2)$.[10] This has two consequences: first, this makes it never optimal for types $(L, L, 1)$ and $(L, L, 2)$ to send message m_{LL}, since $\mu^R(m_{LL}) = 0$ on and off the equilibrium path by condition (iv) of Definition 2. The second implication is that any equilibrium with the aforementioned pooling behavior must have types $(H, H, 1)$ and $(H, H, 2)$ send message m_{HH} with probability 1.[11]

In the equilibrium described in part (a), $m_{L\varnothing}$ is sent by all types that observed $X_1 = L$ while $m_{\varnothing L}$ is sent only by types $(H, L, 1)$ and $(H, L, 2)$. Neither type $(L, L, 1)$ nor $(L, L, 2)$ has a profitable deviation to $m_{\varnothing L}$ if and only if $\mu^R(m_{L\varnothing}) \geq \mu^R(m_{L\varnothing})$, which is what yields the inequality in part (a):

$$\overbrace{\frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1+\gamma_2(1-\nu_1)}{\gamma_2(1-\nu_1)}\right)}}^{\mu^R(m_{L\varnothing})} \geq \overbrace{\frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2\nu_1}{\gamma_1(1-\nu_2)}\right)}}^{\mu^R(m_{\varnothing L})} \Leftrightarrow \frac{\gamma_1+(1-\gamma_1)(1-\nu_1)}{(1-\gamma_1)(1-\nu_1)} \leq \frac{(1-\gamma_1)\nu_1}{\gamma_1(1-\nu_2)}. \tag{3.22}$$

The above restriction places an upper bound (which depends on (ν_1, ν_2)) on the probability γ_1 that the first source is relevant. Mechanically, this is because the latter inequality in (3.22) can be rearranged as $\gamma_1 \leq \bar{\gamma}(\nu_1, \nu_2)$ for some function $\bar{\gamma}(\nu_1, \nu_2)$ which is visualized in the left-hand panel of Figure 3.5. This makes intuitive sense: in part (a) types $(L, L, 1)$ and $(L, L, 2)$ pool with types $(L, H, 1)$ and $(L, H, 2)$ at message $m_{L\varnothing}$. Hence, when the probability that the first source is relevant (γ_1) is sufficiently large, types $(L, L, 1)$ and $(L, L, 2)$ are better off pooling with types $(H, L, 1)$ and $(H, L, 2)$ at message $m_{L\varnothing}$.

Notice that no further restrictions on model parameters are required to sustain an equilibrium that satisfies the description in part (a) of Proposition 10: one only has to specify off path posterior beliefs $\{\pi^R(\cdot|m_{H\varnothing}), \pi^R(\cdot|m_{HL}), \pi^R(\cdot|m_{\varnothing H}), \pi^R(\cdot|m_{LH})\}$, for the

[10] This is because the disclosure of a single, Low source signal realization (i.e. message $m \in \{m_{L\varnothing}, m_{\varnothing L}\}$) can come from the Sender that observed a Low realization of the relevant source ($\Rightarrow \mu^R(m) < 1 \ \forall m \in \{m_{L\varnothing}, m_{\varnothing L}\}$) or a High realization of the relevant source ($\Rightarrow \mu^R(m) > 0 \ \forall m \in \{m_{L\varnothing}, m_{\varnothing L}\}$).

[11] This is because if message $m \in \{m_{H\varnothing}, m_{\varnothing H}\}$ is sent by type $(H, H, 1)$ or $(H, H, 2)$, then $\mu^R(m) = 1$; this implies that type $(L, H, 1)$, $(L, H, 2)$, $(H, L, 1)$, or $(H, L, 2)$ has a profitable deviation to m (because of (3.14) and (3.21)).

Receiver that satisfy the following:

$$\mu^R(m) \leq \mu^R(m_{\varnothing L}) \; \forall m \in \{m_{H\varnothing}, m_{HL}\}; \quad \mu^R(m) \leq \mu^R(m_{L\varnothing}) \; \forall m \in \{m_{\varnothing H}, m_{LH}\}. \tag{3.23}$$

The former inequality precludes profitable deviations for types $(H, L, 1)$ and $(H, L, 2)$; the latter precludes profitable deviations for types $(L, H, 1)$ and $(L, H, 2)$.

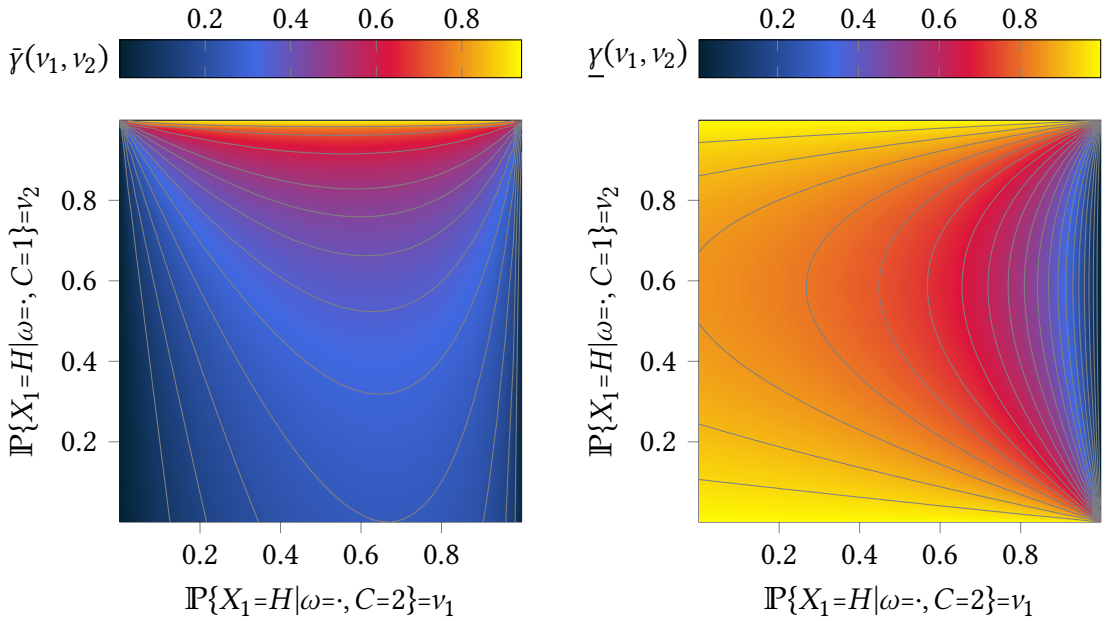


Figure 3.5: Heat maps visualizing $\bar{\gamma}(\nu_1, \nu_2)$ (left-hand side) and $\underline{\gamma}(\nu_1, \nu_2)$ (right-hand side).

The intuition for part (b) is very similar. The inequality in this part is derived from the condition that ensures that types $(L, L, 1)$ and $(L, L, 2)$ do not have a profitable deviation to $m_{L\varnothing}$:

$$\frac{\overbrace{\mu^R(m_{L\varnothing})}}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1\nu_2}{\gamma_2(1-\nu_1)}\right)} \leq \frac{\overbrace{\mu^R(m_{\varnothing L})}}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1(1-\nu_2)+\gamma_2}{\gamma_1(1-\nu_2)}\right)} \Leftrightarrow \frac{\gamma_1\nu_2}{(1-\gamma_1)(1-\nu_1)} \geq \frac{\gamma_1(1-\nu_2)+(1-\gamma_1)}{\gamma_1(1-\nu_2)}, \tag{3.24}$$

where $\mu^R(m_{L\varnothing})$ is given in row 6 of Table A.5 and $\mu^R(m_{\varnothing L})$ is given in the last row of Table A.6. Off path beliefs are specified like in (3.23). In symmetric fashion to part (a), the

condition in (3.24) places a *lower* bound on γ_1; that is, it can be rearranged as $\gamma \geq \underline{\gamma}(\nu_1, \nu_2)$ for some function $\underline{\gamma}(\nu_1, \nu_2)$ which is visualized in the right-hand panel of Figure 3.5.

The proofs for parts (c) and (d) follow very similar reasoning, but with one qualitative difference: since types $(L, L, 1)$ and $(L, L, 2)$ separate, precluding profitable deviations for both types now requires $\mu^R(m_{L\varnothing}) = \mu^R(m_{\varnothing L})$.[12] This is what yields the condition in parts (c) and (d), as before. Unlike the previous parts, these parts are knife-edge cases.

As one may suspect, the Receiver's *ex ante* expected welfare is lower under the partial disclosure equilibria of Proposition 10 than in the full source disclosure equilibrium of Proposition 8. When all sources were disclosed, the Receiver could fully unravel the state whenever the Sender observed $X_1 = X_2$. However, in the partial disclosure equilibria of Proposition 10, the Receiver is only able to fully unravel the state when $X_1 = H = X_2$.

I demonstrate this by comparing the Receiver's expected welfare under the full disclosure equilibrium and the partial disclosure equilibrium in part (a) of Proposition 10. Let the prior probability that $(X_1, X_2) = (x_1, x_2) \in \{L, H\}^2$ be denoted by

$$\pi_{\mathbf{X}}(x_1, x_2) \equiv \sum_{\omega \in \{L,H\}} \sum_{c \in \{1,2\}} \pi(\omega, c, x_1, x_2).$$

In the full source disclosure equilibrium, the Receiver observes message $m_{x_1 x_2}$ with probability $\pi_{\mathbf{X}}(x_1, x_2)$ for all $(x_1, x_2) \in \{L, H\}^2$. Let $\pi^{R^*}(\cdot|\cdot)$ denote the Receiver's posterior beliefs in the full source disclosure equilibrium, and let $\mu^{R^*}(\cdot)$ denote the associated probability placed on the event $\omega = H$. The Receiver's expected welfare under the full

[12] $\mu^R(m_{L\varnothing})$ is given in the penultimate row of Table A.5 in part (c) and the third-to-last row of this table in part (d); $\mu^R(m_{\varnothing L})$ is given in the third-to-last row of Table A.6 in part (c) and the penultimate row of this table in part (d).

source disclosure equilibrium is then given by

$$W^R_{FD} \equiv -\frac{1}{(H\text{-}L)^2}\Big[\pi_{\mathbf{X}}(H,L)\left(1\text{-}\mu^{R^*}(m_{HL})\right)\mu^{R^*}(m_{HL})+\pi_{\mathbf{X}}(L,H)\left(1\text{-}\mu^{R^*}(m_{LH})\right)\mu^{R^*}(m_{LH}) \\ \pi_{\mathbf{X}}(H,H)\cdot 0+\pi_{\mathbf{X}}(L,L)\cdot 0\Big]. \tag{3.25}$$

Note that the last two terms above are equal to zero because $\mu^{R^*}(m_{HH}) = 1$ and $\mu^{R^*}(m_{LL}) = 0$. As we saw in (3.15),

$$\mu^{R^*}(m_{HL}) = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2\nu_1(1\text{-}\mu)};\ \mu^{R^*}(m_{LH}) = \frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+\gamma_1\nu_2(1\text{-}\mu)},$$

In the partial disclosure equilibrium in part (a) of Proposition 10, types $(H,L,1)$ and $(H,L,2)$ pool on message $m_{\varnothing L}$, types $(H,H,1)$ and $(H,H,2)$ pool on message m_{HH}, and all other types pool on message $m_{L\varnothing}$. Hence, the Receiver observes message $m_{\varnothing L}$ with probability $\pi_{\mathbf{X}}(H,L)$, message m_{HH} with probability $\pi_{\mathbf{X}}(H,H)$, and message $m_{L\varnothing}$ with probability $1-\pi_{\mathbf{X}}(H,L)-\pi_{\mathbf{X}}(H,H)$. Notice that this last probability is equal to the prior probability that $X_1 = L$. Let $\pi^{R^{**}}(\cdot|\cdot)$ denote the Receiver's posterior beliefs in the aforementioned partial disclosure equilibrium, and let $\mu^{R^{**}}(\cdot)$ denote the associated probability placed on the event $\omega = H$. As before, $\mu^{R^{**}}(m_{HH}) = 1$ and by (3.22)

$$\mu^{R^{**}}(m_{\varnothing L}) = \frac{\gamma_1(1\text{-}\nu_2)\mu}{\gamma_1(1\text{-}\nu_2)\mu+\gamma_2\nu_1(1\text{-}\mu)};\quad \mu^{R^{**}}(m_{L\varnothing}) = \frac{\gamma_2(1\text{-}\nu_1)\mu}{\gamma_2(1\text{-}\nu_1)\mu+[\gamma_1+\gamma_2(1\text{-}\nu_1)]\,(1\text{-}\mu)}.$$

The Receiver's expected welfare under the aforementioned partial disclosure equilibrium is given by

$$W^R_{PD} \equiv -\frac{1}{(H\text{-}L)^2}\Big[\pi_{\mathbf{X}}(H,L)\left(1-\mu^{R^{**}}(m_{\varnothing L})\right)\mu^{R^{**}}(m_{\varnothing L})+\pi_{\mathbf{X}}(H,H)\cdot 0 \\ +[1-\pi_{\mathbf{X}}(H,L)-\pi_{\mathbf{X}}(H,H)]\left(1\text{-}\mu^{R^{**}}(m_{L\varnothing})\right)\mu^{R^{**}}(m_{L\varnothing})\Big]. \tag{3.26}$$

Notice that $\mu^{R^*}(m_{HL}) = \mu^{R^{**}}(m_{\varnothing L})$. This is because m_{HL} is only sent by types $(H, L, 1)$ and $(H, L, 2)$ in the full disclosure equilibrium and message $m_{\varnothing L}$ is only sent by these types in the partial disclosure equilibrium of Proposition 10, part (a). Consequently, the Receiver's expected welfare gain $W^R_{FD} - W^R_{PD}$ when all sources are disclosed is given by

$$\begin{aligned} W^R_{FD} - W^R_{PD} = -\tfrac{1}{(H-L)^2}\Big[& \pi_{\mathbf{X}}(L,H)\left(1-\mu^{R^*}(m_{LH})\right)\mu^{R^*}(m_{LH}) \\ & -\left[1-\pi_{\mathbf{X}}(H,L)-\pi_{\mathbf{X}}(H,H)\right]\left(1-\mu^{R^{**}}(m_{L\varnothing})\right)\mu^{R^{**}}(m_{L\varnothing})\Big]. \end{aligned} \tag{3.27}$$

After some straightforward algebraic manipulation, we arrive at

$$W^R_{FD} - W^R_{PD} = \frac{(1-\nu_1)\gamma_2\mu}{(H-L)^2} \cdot \frac{(1-\nu_2)\gamma_1 + (1-\nu_1)\gamma_2}{[\gamma_1\nu_2(1-\mu)+\gamma_2\mu(1-\nu_1)]\cdot[\gamma_1(1-\mu)+\gamma_2(1-\nu_1)]} > 0.$$

The strict inequality follows from the fact that each term in the middle expression above is strictly positive since $\mu, \gamma_1, \gamma_2, \nu_1, \nu_2 \in (0,1)$ and $H > L$. This is illustrated in Figure 3.6.

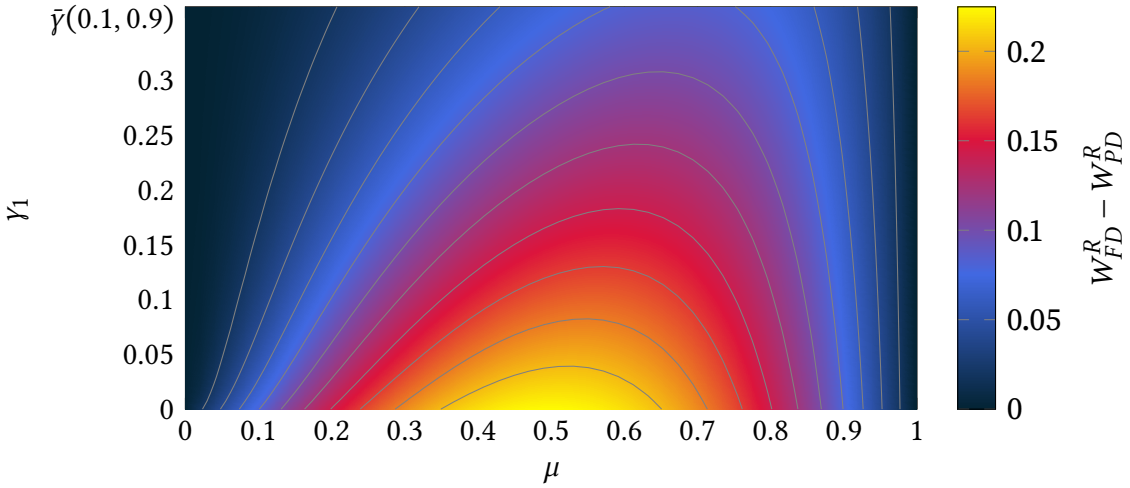


Figure 3.6: Receiver's welfare gain under the full disclosure equilibrium relative to the partial disclosure equilibrium in part (a) of Proposition 10. This figure was generated assuming $L = 0$, $H = 1$, $\nu_1 = 0.1$, $\nu_2 = 0.9$ and $(\mu, \gamma_1) \in (0,1)^2$.

As the above figure demonstrates, the Receiver's expected welfare gain from full

source disclosure $W^R_{FD} - W^R_{PD}$ is large when there is high prior uncertainty about the state (i.e. when μ is near 0.5) and low prior uncertainty that the second source is relevant (i.e. when γ_1 is small). The intuition behind the former relationship is relatively straightforward: in expectation more information is transmitted to the Receiver when all source signals are disclosed; the value of this additional information grows with the prior uncertainty about the state. To understand the effect of γ_1, recall that in the partial disclosure equilibrium of Proposition 10, part (a), all Senders that observed $X_1 = L$ pool on $m_{L\varnothing}$. Hence, upon observing $m_{L\varnothing}$, the Receiver can only infer that $X_1 = L$. Hence, the informational value of this signal diminishes with γ_1.

3.4 Concluding remarks

In this chapter, I studied an evidence disclosure game where the Sender has access to a piece of evidence they know to be perfectly informative about a state, and one that they know to be perfectly uninformative. In Proposition 8, I showed that full disclosure of evidence is possible in (pure strategy perfect Bayesian) equilibrium. However, since this necessarily requires pooling amongst Senders that observe the same conflicting evidence, the Receiver cannot fully unravel which piece of disclosed evidence is informative. Consequently, the Receiver can potentially gain little to no information about the state after observing fully disclosed conflicting evidence when there is high *ex ante* uncertainty regarding which piece of evidence is informative.

Full evidence disclosure is not necessarily ideal for the Receiver's ability to learn about the state; suppressing irrelevant evidence would be beneficial for the Receiver's ability to learn. This would also be beneficial for Senders who observed favorable, relevant evidence and unfavorable, irrelevant evidence. However, I show in Proposition 9 that Senders who observe the same conflicting evidence always pool in equilibrium,

since the informativeness of evidence is not verifiable information. As a result, the same learning failure discussed above generically takes place in equilibrium. Proposition 10 showed that under certain conditions on model primitives, even full source disclosure can fail to hold in equilibrium.

Chapter 3 is currently planned for submission for publication of the material. The dissertation author, Frederick Aram Papazyan, is the sole author of this chapter.

Appendix A

Supplemental Material

A.1 Appendix of Chapter 1

A.1.1 Proofs

Proof of Propositions 1 and 2

Arbitrarily fix $\boldsymbol{x}_{t-\Delta} \in [0, \chi]^N$. Player i_t faces the following optimization problem:

$$\begin{cases} \max\limits_{x_{it}, I_{it}} & H(x_{it}, \boldsymbol{x}_{-i,t}) - \Delta \cdot C(I_{it}, x_{i,t-\Delta}) \\ s.t. & x_{it} = \Delta \cdot I_{it} + x_{i,t-\Delta} - \delta\Delta \\ & 0 \leq x_{it} \leq \chi \\ & I_{it} \geq 0 \end{cases} \tag{A.1}$$

Where $H(x_{it}, \boldsymbol{x}_{-i,t}) = \frac{e^{\lambda x_i}}{\sum_{j=1}^{N} e^{\lambda x_j}}$ by Assumption 2. Notice that the optimization problem in (A.1) is equivalent to the optimization problem

$$\begin{cases} \max\limits_{x_{it}} & H(x_{it}, \boldsymbol{x}_{-i,t}) - \Delta \cdot C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) \\ \text{s.t.} & 0 \leq x_{it} \\ & x_{it} \leq \chi \\ & x_{i,t-\Delta} - \delta\Delta \leq x_{it} \end{cases} \tag{A.1′}$$

wherein x_{it} is the only choice variable.[1] Optimization problem (A.1′) is now solved. I form the Lagrangian

$$\begin{aligned} \mathcal{L} = \; & H(x_i, \boldsymbol{x}_{-i}) - \Delta \cdot C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) \\ & + x_{it}\mu_1 + (\chi - x_{it})\mu_2 + (x_{it} - \delta\Delta - x_{i,t-\Delta})\mu_3 \end{aligned} \tag{A.2}$$

where μ_1, μ_2, and μ_3 respectively denote the Lagrange multiplier of the first, second, and third inequality constraints of (A.1′). The first order condition is given by

$$h(x_{it}, \boldsymbol{x}_{-i,t}) = D_1 C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) - \mu_1 + \mu_2 - \mu_3. \tag{FOC}$$

Recall that $h(x_{it}, \boldsymbol{x}_{-i,t}) \equiv \frac{\partial}{\partial x_{it}} H(x_{it}, \boldsymbol{x}_{-i,t})$ (equation (1.4)) and $D_1 C$ denotes the partial derivative of C with respect to its first argument. The Karush Kuhn-Tucker (KKT) conditions are given by (CS_1), (CS_2), and (CS_3), which respectively correspond to the first, second,

[1] This is achieved by rearranging the equality constraint of (A.1)

$$x_{it} = \Delta \cdot I_{it} + x_{i,t-\Delta} - \delta\Delta \;\Leftrightarrow I_{it} = \frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta$$

and substituting out I_{it}. Notice that the third inequality constraint of (A.1′) can be rewritten as $0 \leq \frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta$.

and third inequality constraints of (A.1′).

$$x_{it} \geq 0,\ \mu_1 \geq 0,\ x_{it}\mu_1 = 0 \tag{CS$_1$}$$

$$x_{it} \leq \chi,\ \mu_2 \geq 0,\ (\chi - x_{it})\mu_2 = 0 \tag{CS$_2$}$$

$$x_{it} \geq x_{i,t-\Delta} - \delta\Delta,\ \mu_3 \geq 0,\ \left(x_{it} - \delta\Delta - x_{i,t-\Delta}\right)\mu_3 = 0 \tag{CS$_3$}$$

Suppose the first two constraints of (A.1′) are both binding. This implies that $x_{it} = 0$ and $x_{it} = \chi$, which is a contradiction since $\chi > 0$. Therefore it is never possible for the first two constraints of (A.1′) to both be binding.

Now suppose that the first constraint of (A.1′) is slack, but the second two constraints are binding. This implies that $x_{it} = \chi$ and $x_{it} = x_{i,t-\Delta} - \delta\Delta$, which in turn jointly imply that $x_{i,t-\Delta} = \chi + \delta\Delta > \chi$, which is a contradiction. It follows that this case is never possible.

I now turn to the case where only the second constraint of (A.1′) is binding. In this case, $\mu_1 = \mu_3 = 0$, $\mu_2 \geq 0$, $x_{it} = \chi$, and $x_{it} > x_{i,t-\Delta} - \delta\Delta$. The latter two imply that $x_{i,t-\Delta} < \chi + \delta\Delta$, which is non-restrictive since $x_{i,t-\Delta}$ must take values in $[0, \chi]$. In this case, it follows from (FOC) that

$$h(\chi, \boldsymbol{x}_{-i}) \geq D_1 C\left(\frac{\chi - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) \tag{A.3}$$

Since h is a bounded function and $D_1 C$ is strictly increasing in its first argument, it follows that there exists a sufficiently small $\tilde{\Delta} > 0$ such that $\forall \Delta < \tilde{\Delta}$ the above inequality holds only at $x_{i,t-\Delta} = \chi$.

Now consider the case where only the third constraint of (A.1′) is binding. In this case $\mu_1 = \mu_2 = 0$, $\mu_3 \geq 0$, $x_{it} \in (0, \chi)$, $x_{it} = x_{i,t-\Delta} - \delta\Delta$. The latter two imply that

$x_{i,t-\Delta} \in (\delta\Delta, \chi]$. In this case, it follows from (FOC) that

$$h(x_{it}, \boldsymbol{x}_{-i,t}) \leq D_1 C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) \tag{A.4}$$

In the case where the first and third constraints of (A.1′) are binding while the second constraint is slack, we have $\mu_1 \geq 0$, $\mu_2 = 0$, $\mu_3 \geq 0$, $x_{it} = 0$, and $x_{it} = x_{i,t-\Delta} - \delta\Delta$. It follows from the latter two that $x_{i,t-\Delta} = \delta\Delta$. It follows from (FOC) that

$$h(0, \boldsymbol{x}_{-i,t}) \leq D_1 C(0, \delta\Delta) \tag{A.5}$$

Next, consider the case where only the first constraint of (A.1′) is binding. Here, we have $\mu_1 \geq 0 = \mu_2 = \mu_3$, $x_{it} = 0$, and $x_{it} > x_{i,t-\Delta} - \delta\Delta$. The latter two imply that $x_{i,t-\Delta} \in [0, \delta\Delta)$. It follows from (FOC) that

$$h(0, \boldsymbol{x}_{-i,t}) \leq D_1 C\left(\delta - \frac{x_{i,t-\Delta}}{\Delta}, x_{i,t-\Delta}\right) = D_1 C\left(\frac{\delta\Delta - x_{i,t-\Delta}}{\Delta}, x_{i,t-\Delta}\right) \tag{A.6}$$

Finally, attention is turned to the *interior* case where all constraints of (A.1′) are slack. Here, $\mu_i = 0\ \forall i \in \{1, 2, \ldots, N\}$, $x_{it} \in (0, \chi)$, and $x_{it} > x_{i,t-\Delta} - \delta\Delta$. Equation (FOC) then implies that x_{it} satisfies the following equality:

$$h(x_{it}, \boldsymbol{x}_{-i,t}) = D_1 C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right). \tag{A.7}$$

Let $D_1 h$ denote the partial derivative of h with respect to its first argument. Note that $h(\cdot, \boldsymbol{x}_{-i,t})$ and $D_1 h(\cdot, \boldsymbol{x}_{-i,t})$ are bounded given any $\boldsymbol{x}_{-i,t} \in [0, \chi]^{N-1}$ and recall that $D_1 C(\cdot, x_{i,t-\Delta})$ is strictly increasing given any $x_{i,t-\Delta} \in [0, \chi]$. It then follows that there exists a sufficiently small $\hat{\Delta} > 0$ such that given any fixed $\Delta < \hat{\Delta}$, a unique $x_{it} \in \left(\max\{x_{i,t-\Delta} - \delta\Delta, 0\}, \chi\right)$

satisfies (A.7) and that at said x_{it}

$$D_1 h(x_{it}, \boldsymbol{x}_{-i,t}) - D_{11} C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) < 0, \tag{A.8}$$

where $D_{11}C(z, \boldsymbol{x}_{i,t-\Delta}) \equiv \frac{\partial^2}{\partial I^2} C(I, x_{i,t-\Delta})\big|_{I=z}$.

Given the above, for every $\Delta < \min\{\tilde{\Delta}, \hat{\Delta}\}$, the maximizer x_{it}^* of (A.1′) is unique and satisfies

$$\begin{cases} x_{it}^* = 0, & \text{if } h(0, \boldsymbol{x}_{-i,t}) \leq D_1 C\left(\frac{\delta\Delta - x_{i,t-\Delta}}{\Delta}, x_{i,t-\Delta}\right) \\ & \quad \text{and } x_{i,t-\Delta} \in [0, \delta\Delta] \\ x_{it}^* = x_{i,t-\Delta} - \delta\Delta, & \text{if } h(x_{it}, \boldsymbol{x}_{-i,t}) \leq D_1 C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) \\ & \quad \text{and } x_{i,t-\Delta} \in (\delta\Delta, \chi] \\ x_{it}^* = \chi, & \text{if } h(\chi, \boldsymbol{x}_{-i}) \geq D_1 C\left(\frac{\chi - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right) \\ h(x_{it}, \boldsymbol{x}_{-i,t}) = D_1 C\left(\frac{x_{it} - x_{i,t-\Delta}}{\Delta} + \delta, x_{i,t-\Delta}\right), & \text{otherwise} \end{cases}$$

Sending $\Delta \to 0$ yields

$$\begin{cases} \dot{x}_{it} = 0, & \text{if } h(0, \boldsymbol{x}_{-i,t}) \leq D_1 C\,(0,0) \text{ and } x_{it} = 0 \\ & \qquad \text{or} \\ & h(\chi, \boldsymbol{x}_{-i,t}) \geq D_1 C(\delta, \chi) \text{ and } x_{it} = \chi \\ \dot{x}_{it} = -\delta, & \text{if } h(x_{it}, \boldsymbol{x}_{-i,t}) \leq D_1 C\,(0, x_{it}) \text{ and } x_{it} \in (0, \chi] \\ \dot{x}_{it} = (D_1 C)^{-1}\left(h(x_{it}, \boldsymbol{x}_{-i,t}), x_{it}\right) - \delta, & \text{otherwise} \end{cases}$$

Note that under Assumption 1 $(D_1 C)^{-1}$ is guaranteed to be a well-defined function (Jittorntrum (1978, Theorem 1); Kumagai (1980)). ■

Proof of Proposition 3

Parts 1, 2, 3, 4, and 5, respectively establish the necessary and sufficient conditions under which escalated inclusive, de-escalated inclusive, oligarchic, weak dictatorial, and strong dictatorial power structures are stable. Part 6 then shows that all other power structures will fail to be stable.

To establish the stability of inclusive and oligarchic power structures, I use Lyapunov's Direct/Second Method (Lyapunov, 1892, 1992).[2] According to this method, $\bar{\boldsymbol{x}} \in [0, \chi]^N$ is stable if there exists a continuous, differentiable function $\Lambda : \mathbb{R}^N \to \mathbb{R}$ and an ε-ball of $\bar{\boldsymbol{x}}$, $B_\varepsilon(\bar{\boldsymbol{x}})$, such that the following hold:

1. $\Lambda(\bar{\boldsymbol{x}}) = 0$ and $\Lambda(\boldsymbol{x}_t) > 0 \; \forall \boldsymbol{x}_t \in B_\varepsilon(\bar{x}) \setminus \{\bar{\boldsymbol{x}}\}$.

2. $\frac{d}{dt}\Lambda(\boldsymbol{x}_t) < 0 \; \forall \boldsymbol{x} \in B_\varepsilon(\bar{x}) \setminus \{\bar{\boldsymbol{x}}\}$

Λ is often referred to as a *Lyapunov function* and thought of as an "energy function."

$$\Lambda(\boldsymbol{x}_t) \equiv \frac{1}{2}\sum_{i=1}^{N}(\bar{x}_i - x_{it})^2 \tag{A.9}$$

whose time-derivative is

$$\dot{\Lambda}(\boldsymbol{x}_t) \equiv \frac{d}{dt}\Lambda(\boldsymbol{x}_t) = -\sum_{i=1}^{N}(\bar{x}_i - x_{it})\dot{x}_{it}. \tag{A.10}$$

Intuitively, Lyapunov's Direct Method amounts to showing that the energy of the system *strictly* decreases to zero along all trajectories starting sufficiently close to a steady state $\bar{\boldsymbol{x}}$.

In what follows, let N be arbitrarily fixed. Let $\mathbf{1}_k$ and $\mathbf{0}_k$ respectively denote $(1, \dots, 1) \in \mathbb{R}^k$ and $(0, \dots, 0) \in \mathbb{R}^k$ $(k = 1, \dots, N)$. Let $\boldsymbol{e}_i \in \mathbb{R}^N$ denote the i^{th} standard ba-

[2] For more information on Lyapunov's Direct Method, please see LaSalle (1960), La Salle and Lefschetz (2012), and Chiang and Alberto (2015).

sis vector ($i = 1, \dots, N$). Finally, let $B_\varepsilon(\boldsymbol{x})$ denote the ε-ball centered at $\boldsymbol{x}$, where $\varepsilon > 0$ and $\boldsymbol{x} \in \mathbb{R}^N$. I now proceed with the proof of Proposition 3, proving each part in turn.

Proof of Part 1 Here, I show that the escalated inclusive power structure $\bar{\boldsymbol{x}} = (\chi, \dots, \chi)$ is stable if and only if Condition I

$$h\big(\chi, (\chi, \dots, \chi)\big) > D_1 C(\delta, \chi)$$

holds. I first suppose that this condition holds. It then follows from (1.7) that $\dot{x}_i = 0$ $\forall i \in \{1, \dots, N\}$ so that the first part of Definition 1 is satisfied. I now turn to showing that the remaining part of this definition is satisfied. Notice that Condition I, along with the continuity of h and $D_1 C$, imply that for some $\varepsilon > 0$

$$h(x_i, \boldsymbol{x}_{-i}) > D_1 C(\delta, x_i), \forall i \in \{1, \dots, N\} \text{ and } \forall \boldsymbol{x} \in B_\varepsilon(\chi \mathbf{1}_N). \tag{A.11}$$

The dynamics of $\boldsymbol{x}$ in $B_\varepsilon(\mathbf{1}_N) \cap [0, \chi]^N$ are therefore given by

$$\dot{x}_i = \begin{cases} 0, & \text{if } x_i = \chi \\ (D_1 C)^{-1}\left(h(x_i, \boldsymbol{x}_{-i}), x_i\right) - \delta, & \text{otherwise} \end{cases} \qquad (i = 1, \dots, N). \tag{A.12}$$

Arbitrarily fix $i \in \{1, \dots, N\}$. Then at every $\boldsymbol{x} \in B_\varepsilon(\chi \mathbf{1}_N) \cap \left\{\boldsymbol{x} \in [0, \chi]^N : x_i < \chi\right\}$, we have $\dot{x}_i > 0$ because

$$h(x_i, \boldsymbol{x}_{-i}) > D_1 C(\delta, x_i) \Leftrightarrow (D_1 C)^{-1}(h(x_i, \boldsymbol{x}_{-i}), x_i) - \delta > 0 \Leftrightarrow \dot{x}_i > 0. \tag{A.13}$$

where the last equivalence directly follows from (A.12). Evaluating (A.9) and (A.10) at $\bar{x} = \chi \mathbf{1}_N$ yields

$$\Lambda(\boldsymbol{x}) = \frac{1}{2}\sum_{1}^{N}(\chi - x_i)^2; \quad \dot{\Lambda}(\boldsymbol{x}) = -\sum_{1}^{N}(\chi - x_i)\dot{x}_i.$$

Clearly $\Lambda(\chi \mathbf{1}_N) = 0$ and $\dot{\Lambda}(\boldsymbol{x}_t) < 0$ in $B_\varepsilon(\chi \mathbf{1}_N) \cap [0, \chi]^N \setminus \{\mathbf{1}_N\}$. Thus $\boldsymbol{x} = \chi \mathbf{1}_N$ is stable. Now let us suppose that

$$h(\chi, \chi \mathbf{1}_{N-1}) \leq D_1 C(\delta, \chi),$$

meaning that Condition I is violated. Then $h(\alpha, \alpha \mathbf{1}_{N-1}) \leq D_1 C(\delta, \alpha)\ \forall \alpha \in [0, \chi)$. Fix some $\varepsilon > 0$ and consider $\boldsymbol{x}_0 = (\chi - \rho)\mathbf{1}_N$, where $\rho \in (0, \varepsilon)$. System (1.7) then implies that

$$h(\chi - \rho, (\chi - \rho)\mathbf{1}_{N-1}) = D_1 C(\delta, \chi - \rho).$$

It then follows that $\dot{x}_{it} = 0\ \forall i \in \{1, ..., N\}$ and $\forall t \geq 0$ at $\boldsymbol{x}_t = (\chi - \rho)\mathbf{1}_N$, thereby making $\chi \mathbf{1}_N$ not stable. If instead we had

$$h(\chi - \rho, (\chi - \rho)\mathbf{1}_{N-1}) < D_1 C(\delta, \chi - \rho),$$

then (1.7) would imply that $\dot{x}_{i0} < 0\ \forall\ i$. Moreover, since $x_{i0} = x_{j0}\ \forall i, j$, we also know that $\dot{x}_{i0} = \dot{x}_{j0}\ \forall i, j$. It then clearly follows that $x_{it} = x_{jt} < \chi$ and $\dot{x}_{it} = \dot{x}_{jt} < 0\ \forall i, j \in \{1, ..., N\}$ while $\boldsymbol{x}_t \in B_\varepsilon(\chi \mathbf{1}_N) \cap [0, \chi]^N$. Hence $\exists t \geq 0$ at which $\boldsymbol{x}_t$ is not in this ε-ball centered at $\chi \mathbf{1}_N$, thereby making $\chi \mathbf{1}_N$ not stable.

Proof of Part 2 Here, I show that the de-escalated inclusive power structure $\bar{\boldsymbol{x}} = \mathbf{0}_N$ is stable if and only if Condition II

$$h(0, \mathbf{0}_{N-1}) \leq D_1 C(0, 0)$$

holds. First suppose that it does; this implies that $\exists \alpha \in (0, \chi]$ s.t. $h(\alpha, \alpha \mathbf{1}_{N-1}) < D_1 C(\delta, \alpha)$. To see why, first note that $D_1 C(0,0) < D_1 C(\delta, 0)$ since $\delta > 0$ and by Assumption 1 $D_1 C$ is strictly increasing in its first argument. Since $h(a, a\mathbf{1}_{N-1}) = h(0, \mathbf{0}_{N-1})\ \forall a > 0$ and $D_1 C$ is assumed to be continuous, the statement in (A.1.1) follows.

Assumptions 1 and 2 respectively imply that $D_1 C(\delta, \alpha') \leq D_1 C(\delta, \alpha)$ and

$$h(\alpha', \alpha' \mathbf{1}_{N-1}) = h(\alpha, \alpha \mathbf{1}_{N-1})\ \forall \alpha \in [0, \alpha').$$

It then follows that $h(\alpha, \alpha \mathbf{1}_{N-1}) < D_1 C(\delta, \alpha)$ for all such α. By the continuity of h and $D_1 C$, this in turn implies that $\exists \varepsilon > 0$ such that

$$h(x_i, \boldsymbol{x}_{-i}) < D_1 C(0, x_i),\ \forall i, \text{ and } \forall \boldsymbol{x} \in B_\varepsilon(\mathbf{0}_N) \cap [0, \chi]^N. \tag{A.14}$$

Therefore, the dynamics in $B_\varepsilon(\mathbf{0}_N) \cap [0, \chi]^N$ are given by $\dot{x}_i = -\delta \mathbb{1}_{\mathbb{R}_{++}}(x_i)$ for all $i \in \{1, ..., N\}$ .Clearly $\dot{x}_i = 0\ \forall i$ at $\mathbf{0}_N$, so that the first part of Definition 1 is satisfied. Clearly all trajectories in $B_\varepsilon(\mathbf{0}_N) \cap [0, \chi]^N$ approach the origin in the limit, and this can again be verified using Lyapunov's Direct method. Evaluating (A.9) and (A.10) at $\bar{\boldsymbol{x}} = \mathbf{0}_N$:

$$\Lambda(\boldsymbol{x}) = \frac{1}{2} \sum_1^N x_i^2; \quad \dot{\Lambda}(\boldsymbol{x}) = \sum_1^N x_i \dot{x}_i$$

The first function is zero at $\boldsymbol{x} = \mathbf{0}_N$ and the latter is strictly negative at every $\boldsymbol{x} \in B_\varepsilon(\mathbf{0}_N) \cap [0, \chi]^N \setminus \{\mathbf{0}_N\}$. Hence, $\mathbf{0}_N$ is stable.

Now suppose instead that $h(0, \mathbf{0}_{N-1}) > D_1 C(0,0)$, which violates Condition II. It then directly follows from (1.7) that $\dot{x}_i > 0\ \forall i \in \{1, \dots, N\}$ at $\boldsymbol{x} = \mathbf{0}_N$, which violates the first part of Definition 1. ■

Proof of Part 3 Fix some $k \in \{2, ..., N-1\}$. I show that $(\chi \mathbf{1}_k, \mathbf{0}_{N-k})$ is stable if and only if Condition III

$$h\big(\chi, (\overbrace{\chi, ..., \chi}^{k-1}, \overbrace{0, ..., 0}^{N-k})\big) > D_1 C(\delta, \chi) \text{ and } h\big(0, (\overbrace{\chi, ..., \chi}^{k}, \overbrace{0, ..., 0}^{N-k-1})\big) < D_1 C(0, 0)$$

holds for k. By symmetry, this is without loss, and the proof where $\tilde{\boldsymbol{x}}$ is instead an arbitrarily fixed element of

$$\left\{ \boldsymbol{x} \in \{0, \chi\}^N \ : \ \sum_{i=1}^{N} x_i = k\chi \right\} \tag{A.15}$$

is identical, apart from having more complicated, cumbersome notation.

First, suppose that Condition III holds. Recall that this requires the following two inequalities to hold:

$$h\big(\chi, (\chi \mathbf{1}_{k-1}, \mathbf{0}_{N-k})\big) > D_1 C(\delta, \chi) \tag{A.16a}$$

$$h\big(0, (\chi \mathbf{1}_k, \mathbf{0}_{N-k-1})\big) < D_1 C(0, 0) \tag{A.16b}$$

where (A.16a) and (A.16b) respectively correspond to Condition III.1 and III.2. It follows from Condition III.1 and equation 2 of (1.7) that $\dot{x}_i = 0 \ \forall i \in \{1, ..., k\}$ at $\boldsymbol{x} = (\chi \mathbf{1}_k, \mathbf{0}_{N-k})$. It follows from III.2 and line 1 of (1.7) that $\dot{x}_i = 0 \ \forall i \in \{k+1, \dots, N\}$. Therefore, part (a) of the definition of stability (Definition 1) is satisfied.

Maintaining the supposition that Condition III holds, I now show $\boldsymbol{x} = (\chi \mathbf{1}_k, \mathbf{0}_{N-k})$ satisfies the second definition of stability, again using Lyapunov's Direct Method. Since the inequalities in (A.16a) and (A.16b) hold at $(\chi \mathbf{1}_k, \mathbf{0}_{N-k})$, and since h and $D_1 C$ are continuous functions, these inequalities hold in an $\varepsilon-$neighborhood of $\boldsymbol{x} = (\chi \mathbf{1}_k, \mathbf{0}_{N-k})$. Let $\mathbb{U} \equiv B_\varepsilon((\chi \mathbf{1}_k, \mathbf{0}_{N-k})) \cap [0, \chi]^N$. It then follows from 1.7 that at every $\boldsymbol{x} \in \mathbb{U}$ and for all

$i \in \{1, \dots, N\}$:

$$\dot{x}_i = \begin{cases} \left[(D_1 C)^{-1}\left(h(x_{it}, \boldsymbol{x}_{-i,t}), x_i\right) - \delta\right] \mathbb{1}_{\mathbb{R}_{++}}(x_i) & \text{if } i \leq k \\ -\delta \mathbb{1}_{(0,\infty)}(x_i) & \text{if } i > k. \end{cases} \tag{A.17}$$

Notice that for all $i \leq k$, $\dot{x}_i > 0$ in $\mathbb{U} \setminus \{\boldsymbol{x} : x_i = \chi\}$ because

$$h(x_{it}, \boldsymbol{x}_{-i,t}) > D_1 C(\delta, x_i) \Leftrightarrow (D_1 C)^{-1}\left(h(x_{it}, \boldsymbol{x}_{-i,t}), x_i\right) - \delta > 0.$$

As before, $\Lambda((\chi \mathbf{1}_k, \mathbf{0}_{N-k})) = 0$. $\frac{d}{dt}\Lambda(\tilde{\boldsymbol{x}}) < 0 \; \forall \tilde{\boldsymbol{x}} \in \mathbb{U} \setminus \{(\chi \mathbf{1}_k, \mathbf{0}_{N-k})\}$ because

$$\frac{d}{dt}\Lambda(\tilde{\boldsymbol{x}}_t) = -\sum_1^N (x_i - \tilde{x}_i)\dot{x}_i = -\left[\sum_1^k (\chi - \tilde{x}_i)\dot{x}_i + \sum_{k+1}^N (-\tilde{x}_i)\dot{x}_i\right] < 0.$$

Note that the above inequality is *strict* because at any $\boldsymbol{x} \in \mathbb{U} \setminus \{(\chi \mathbf{1}_k, \mathbf{0}_{N-k})\}$ either $x_i < \chi$ for some $i \leq k$ or $x_i > 0$ for some $i > k$; note that in $\mathbb{U}$, $\forall i > k \; \dot{x}_i = 0$ when $x_i = 0$ and strictly negative otherwise. Thus, $(\chi \mathbf{1}_k, \mathbf{0}_{N-k})$ is stable.

I now suppose that the following inequalities hold:

$$h(1, (\chi \mathbf{1}_{k-1}, \mathbf{0}_{N-k})) \leq D_1 C(\delta, \chi), \tag{A.18}$$

$$h(0, (\chi \mathbf{1}_k, \mathbf{0}_{N-k-1})) < D_1 C(0, 0). \tag{A.19}$$

I now show that $(\chi \mathbf{1}_k, \mathbf{0}_{N-k})$ is not stable. Note that by the continuity of h, $\exists \varepsilon > 0$ such that

$$h(\alpha, (\alpha \mathbf{1}_k, \mathbf{0}_{N-k-1})) < D_1 C(0, 0) \; \forall \alpha \in (\chi - \varepsilon, \chi]. \tag{A.20}$$

Define the following subset $\mathbb{L}$ of $B_\varepsilon(\bar{\boldsymbol{x}})$,

$$\begin{aligned}\mathbb{L} \equiv \quad & \left\{\boldsymbol{x} \in [0,\chi]^N \,:\, x_i = \alpha \in (\chi-\varepsilon,\chi)\ \forall i \leq k,\ x_i = 0\ \forall i > k\right\} \\ & \left\{\boldsymbol{x} \in [0,\chi]^N \,:\, \boldsymbol{x} = \alpha \sum_{i=1}^{k} \boldsymbol{e}_i \text{ for some } \alpha \in (\chi-\varepsilon,\chi)\right\}\end{aligned} \tag{A.21}$$

Notice that by construction $\dot{x}_i = 0\ \forall i \in \{k+1,\dots,N\}$ at any $\boldsymbol{x} \in \mathbb{L}$. Furthermore, at any $\boldsymbol{x} \in \mathbb{L}$ we also have $\dot{x}_i = \dot{x}_j < 0\ \forall i,j \in \{1,\dots,k\}$. This is because

$$\frac{\partial}{\partial\alpha} h(\alpha,(\alpha\mathbf{1}_{k-1},\mathbf{0}_{N-k})) = \frac{\lambda^2 e^{\alpha\lambda}(N-k)[(k-2)e^{\alpha\lambda}+(N-k)]}{[N+(e^{\alpha\lambda}-1)k]^3} > 0 \tag{A.22}$$

with the inequality following from the facts that $N > k \geq 2$, $\lambda > 0$, and the fact that $e^{\alpha\lambda} > 1\ \forall(\alpha,\lambda) \in (0,\infty)^2$. (In words: at all states in $\mathbb{L}$, powerless lineages $(k+1,\dots,N)$ maintain zero power, and all other lineages let their power depreciate at the same rate.) It then follows that if $\boldsymbol{x}_0 \in \mathbb{L}$, then there exists a $\tau > 0$ such that $\boldsymbol{x}_\tau \notin B_\varepsilon(\bar{\boldsymbol{x}})$. At every $t \in [0,\tau)$ $\dot{x}_{it} = \dot{x}_{jt} < 0\ \forall i,j \in \{1,\dots,k\}$ and $\dot{x}_{it} = 0\ \forall i \in \{k+1,\dots,N\}$, so that at future time $t' \in (t,\tau)$, $\boldsymbol{x}_{t'} \in \mathbb{L}$. $\boldsymbol{x}_t$ moves along $\mathbb{L}$ in this fashion at all $t \in [0,\tau)$ until it leaves the ε-ball of $\bar{\boldsymbol{x}}$ at time τ. This violates the definition of stability, specifically the second part of Definition 1.

This completes the proof that $(\chi\mathbf{1}_k,\mathbf{0}_{N-k})$ is stable if and only if Condition III holds for k. As mentioned earlier, this was without loss of generality by symmetry; it then follows that every element of

$$\left\{\boldsymbol{x} \in \{0,\chi\}^N \,:\, \sum_{i=1}^{N} x_i = k\chi\right\}$$

is stable if Condition III holds for k and no element of the above set is stable otherwise.

■

Proof of Part 4 Fix $i \in \{1, ..., N\}$ and $d \in (0, \chi)$. I show that every weak dictatorial power structure in $\{\boldsymbol{e}_i\}_{i=1}^N$ is stable if and only if Condition IV

$$h\big(\cdot, (0,..., 0)\big) \text{ intersects } D_1C(\delta, \cdot) \text{ from above at } d, \text{ and}$$

$$h\big(0, (d, 0,..., 0)\big) < D_1C(0, 0)$$

holds at d; recall that by "intersect from above at d" I mean that $h\big(\cdot, (0,..., 0)\big) - D_1C(\delta, \cdot)$ is zero and strictly decreasing at d. Suppose that Condition IV holds at $d \in (0, \chi)$. I now show that $\tilde{\boldsymbol{x}} = d\boldsymbol{e}_i$ is then stable. First, notice that by the first part of this condition, $\dot{x}_i = 0$ solves

$$h(d, \mathbf{0}_{N-1}) = D_1C(\dot{x}_i + \delta, d) \tag{A.23}$$

by construction. The second part of IV dictates that $h(0, (d, \mathbf{0}_{N-2}) < D_1C(0, 0)$, so it then follows from the first line of (1.7) that $\dot{x}_j = 0 \; \forall j \neq i$ at $\tilde{\boldsymbol{x}} = d\boldsymbol{e}_i$. Hence the first part of the definition of stable is satisfied.

I now turn to the second part of said definition. Assume that Condition IV holds for some $i \in \{1, ..., N\}$. Then there exists an $d \in (0, \chi)$ at which $h(\cdot, \mathbf{0}_{N-1})$ intersects $D_1C(\delta, \cdot)$ from above at $x_i = d$. By the second part of this condition, we have

$$h(0, (d, \mathbf{0}_{N-2})) < D_1C(0, 0).$$

The stability of $d\boldsymbol{e}_i$ is now established not by Lyapunov's Direct Method but rather by using the definition of stability itself. That is, I must show that there exists a neighborhood of this point in which all trajectories beginning in said neighborhood approach $d\boldsymbol{e}_i$ in the limit.

Since $h(\cdot, \mathbf{0}_{N-1})$ intersects $D_1C(\delta, \cdot)$ from above at $x_i = d$, it follows from the conti-

nuity of h and D_1C that $\exists\tilde{\varepsilon}_1 > 0$ such that the following inequalities hold.

$$h(x_i, \mathbf{0}_{N-1}) < D_1C(\delta, x_i)\ \forall x_i \in (d, d+\tilde{\varepsilon}_1) \tag{A.24}$$

$$h(x_i, \mathbf{0}_{N-1}) > D_1C(\delta, x_i)\ \forall x_i \in (d-\tilde{\varepsilon}_1, d) \tag{A.25}$$

Furthermore, by the second part of Condition IV and the continuity of h and D_1C, $\exists\tilde{\varepsilon}_2 > 0$ such that

$$h(0, (d, \mathbf{0}_{N-2})) < D_1C(0,0)\ \forall x_i \in (d-\tilde{\varepsilon}_2, d+\tilde{\varepsilon}_2). \tag{A.26}$$

Let $\tilde{\varepsilon} \equiv \min\{\tilde{\varepsilon}_1, \tilde{\varepsilon}_2\}$. It then follows from (1.7) that any trajectory starting at an initial conition $\boldsymbol{x}_0$ such that $x_{i0} \in (d-\tilde{\varepsilon}, d+\tilde{\varepsilon})$ and $x_{j0} = 0\ \forall j \neq i$ asymptotes towards $d\boldsymbol{e}_i$.

Now let us consider trajectories starting *away* from the x_i-axis. Throughout, keep in mind the following:

$$h(x_j, \boldsymbol{x}_{-j}) < D_1C(0, x_j) \Rightarrow h(x_j, \boldsymbol{x}_{-j}) < D_1C(\delta, x_j)\ \forall \boldsymbol{x},\ \forall j \neq i. \tag{A.27}$$

That is, when the first inequality is true, the second one is also true. Below, for $j \neq i$, I prove statements using the first inequality only. This helps me exploit the continuity of h and D_1C.[3] Now, let

$$g_j(x_j, \boldsymbol{x}_{-j}) := h(x_j, \boldsymbol{x}_{-j}) - D_1C(0, x_j)\ \forall j \neq i \tag{A.28}$$

and

$$g_i(x_i, \boldsymbol{x}_{-i}) := h(x_i, \boldsymbol{x}_{-i}) - D_1C(\delta, x_i). \tag{A.29}$$

[3]Recall the presence of $\mathbb{1}_{\mathbb{R}_{++}}(x_j)$ inside $D_1C(\cdot,\cdot)$ in (1.7).

For each $j \neq i$, fix some $\varepsilon_j \in \left(0, \left|g_j(0,(d,\mathbf{0}_{N-1}))\right|\right)$. Since $g_j \in \mathscr{C}$, $\exists \zeta_j > 0$ such that

$$\left|g_j(x_j, \boldsymbol{x}_{-j}) - g_j(0,(d,\mathbf{0}_{N-1}))\right| < \varepsilon_j \text{ when } |x_\ell| < \zeta_j \ \forall \ell \neq i \text{ and } |x_i - d| < \zeta_j.$$

Now, let $\varepsilon = \min\{\{\varepsilon_j\}_{j\neq i}\}$ and $\zeta = \min\{\{\zeta\}_{j\neq i}\}$ and let

$$\mathcal{S}_\zeta(d\boldsymbol{e}_i) := \{\boldsymbol{x} \in [0,\chi]^N \ : \ |x_i - d| < \zeta, |x_j| < \zeta \ \forall j \neq i\} \tag{A.30}$$

denote the *square* ζ-neighborhood centered at $d\boldsymbol{e}_i$, intersected with $[0,\chi]^N$. Then, for all $j \neq i$ and $\boldsymbol{x} \in \mathcal{S}_\zeta(d\boldsymbol{e}_i) \setminus \partial([0,\chi]^N)$,[4]

$$\left|g_j(x_j, \boldsymbol{x}_{-j}) - g_j(0,(d,\mathbf{0}_{N-2}))\right| < \varepsilon$$

Notice that by construction, $g_j(0,(d,\mathbf{0}_{N-2})) < \varepsilon$. Therefore, $g_j(x_j,\boldsymbol{x}_{-j}) \ \forall \boldsymbol{x} \in \mathcal{S}_\zeta(d\boldsymbol{e}_i) \setminus \partial([0,\chi]^N)$. Hence, by (1.7), it follows that $\dot{x}_j < 0 \ \forall j \neq i$ at every point in $\boldsymbol{x} \in \mathcal{S}_\zeta(d\boldsymbol{e}_i) \setminus \partial([0,\chi]^N)$. Therefore, the power of all players other than i strictly decays when $\boldsymbol{x}$ is away from the x_i-axis but sufficiently close to the dictatorial steady state.

Note that since $h(\cdot,\mathbf{0}_{N-1})$ intersects $D_1C(\delta,\cdot)$ from above at d and both are continuous, it follows that the gap between these two functions widens as one moves away from d (at least within a sufficiently small neighborhood of this point). Hence, there exists a $\varphi > 0$ such that

$$\frac{\partial}{\partial x_i}\left(h(x_i,\mathbf{0}_{N-1}) - D_1C(\delta,x_i)\right) < 0 \ \forall x_i \in (d, d+\varphi). \tag{A.31}$$

Now, let $\eta = \min\{\zeta, \tilde{\varepsilon}, \varphi\}$ and fix an $\alpha \in (0,\eta)$. By construction,

$$h(d+\alpha,\mathbf{0}_{N-1}) < D_1C(\delta, d+\alpha).$$

[4]Here, "$\partial(\cdot)$" denotes the *boundary operator*.

h is strictly increasing in x_j $(\forall j \neq i)$ at $\boldsymbol{x} = (d+\alpha)\boldsymbol{e}_i$. Moreover, notice that

$$h(d+\alpha, \boldsymbol{x}_{-i}) \geq h(d, \mathbf{0}_{N-1}) > D_1 C(\delta, d) > D_1 C(\delta, d+\alpha) \text{ if } \sum_{j \neq i} x_j \geq \alpha.$$

Therefore, $\exists \beta_\alpha \in (0, \alpha)$ such that the following hold:

$$h(d+\alpha, \boldsymbol{x}_{-i}) = D_1 C(\delta, d+\alpha) \text{ when } \sum_{j \neq i} x_i = \beta_\alpha \tag{A.32}$$

$$h(d+\alpha, \boldsymbol{x}_{-i}) > D_1 C(\delta, d+\alpha) \text{ when } \sum_{j \neq i} x_i > \beta_\alpha \tag{A.32$'$}$$

$$h(d+\alpha, \boldsymbol{x}_{-i}) < D_1 C(\delta, d+\alpha) \text{ when } \sum_{j \neq i} x_i < \beta_\alpha. \tag{A.32$''$}$$

The two inequalities above follow from the fact that h is increasing in $x_j \forall j \neq i$ at $d\boldsymbol{e}_i$. Notice that by construction, β_α is strictly increasing in α.[5] Let $\omega = \min\left\{\eta, \beta_\eta\right\}$ and let $\mathcal{S}_\omega(d\boldsymbol{e}_i)$ denote the *square* ω–neighborhood about $d\boldsymbol{e}_i$, intersected by the unit hypercube. Since β_α is monotone in α, it follows that $\mathcal{S}_\omega(d\boldsymbol{e}_i)$ is *bisected* by the hypersurface $x_i = \mu(\boldsymbol{x}_{-i})$. μ is increasing in the sense that $\frac{\partial}{\partial x_j}\mu(\boldsymbol{x}_{-i}) > 0 \ \forall j \neq i$. Moreover, Note that $\mu(\mathbf{0}_{N-1}) = d$ and $\mu(\boldsymbol{x}_{-i}) = \omega$ whenever $\sum_{j \neq i} x_j = \omega$. Let

$$\mathcal{U} \equiv \mathcal{S}_\omega(d\boldsymbol{e}_i) \cap \{\boldsymbol{x} : x_i > \mu(\boldsymbol{x}_i)\}$$

and

$$\mathcal{B} \equiv \mathcal{S}_\omega(d\boldsymbol{e}_i) \cap \{\boldsymbol{x} : x_i < \mu(\boldsymbol{x}_i)\}.$$

In $\mathcal{U}$, $\dot{x}_i < 0$ and in $\mathcal{B}$, $\dot{x}_i > 0$. Recall that we have already established that $\dot{x}_j < 0$ for all $j \neq i$ throughout $\mathcal{S}_\omega(d\boldsymbol{e}_i)$. Therefore, any trajectory starting in $\mathcal{S}_\omega(d\boldsymbol{e}_i)$ will eventually intersect the x_i–axis. As we saw earlier, once $\boldsymbol{x}_t$ is on the x_i–axis (and in $\mathcal{S}_\omega(d\boldsymbol{e}_i)$), $\boldsymbol{x}_t$

[5] By (A.31), $h(d+\alpha, \mathbf{0}_{N-1}) - D_1 C(\delta, d+\alpha)$ becomes increasingly negative as α increases. Accordingly, an increasingly large β_α is needed in order to make $h(d+\alpha, \cdot) - D_1 C(\delta, d+\alpha) = 0$ again, as in (A.32).

will approach $d\boldsymbol{e}_i$ as $t \to \infty$. Therefore, the second part of the definition of stability is satisfied by $d\boldsymbol{e}_i$

Now, suppose that $h(\cdot, \mathbf{0}_{N-1})$ and $D_1C(\delta, \cdot)$ do not intersect at d. If $h(d, \mathbf{0}_{N-1}) > D_1C(\delta, d)$, then $\dot{x}_i$ evolves according to the last line of (1.7) at $\boldsymbol{x} = d\boldsymbol{e}_i$ since $d < \chi$. Hence $\dot{x}_i = (D_1C)^{-1}(h(d, \mathbf{0}_{N-1}), d) - \delta$ at $\boldsymbol{x} = d\boldsymbol{e}_i$. Since

$$h(d, \mathbf{0}_{N-1}) > D_1C(\delta, d) \Rightarrow (D_1C)^{-1}(h(d, \mathbf{0}_{N-1}), d) - \delta > 0$$

thereby violating the first part of Definition 1.

Now suppose that $h(d, \mathbf{0}_{N-1}) < D_1C(\delta, d)$. Since C is assumed to be strictly increasing and strictly convex in its first argument, this permits

$$h(d, \mathbf{0}_{N-1}) < D_1C(0, d)$$

to hold. In this case (1.7) implies that $\dot{x}_i = -\delta < 0$ at $\boldsymbol{x} = d\boldsymbol{e}_i$, thereby violating the first part of Definition 1. I now show that this part of the definition is still violated if $h(d, \mathbf{0}_{N-1}) \in (D_1C(0, d), D_1C(\delta, d))$. Since $d > 0$, $\dot{x}_i = (D_1C)^{-1}(h(d, \mathbf{0}_{N-1}), d) - \delta$ at $\boldsymbol{x} = d\boldsymbol{e}_i$ by the last line of (1.7). Since

$$h(d, \mathbf{0}_{N-1}) < D_1C(\delta, d) \Rightarrow (D_1C)^{-1}(h(d, \mathbf{0}_{N-1}), d) - \delta < 0$$

it follows that $\dot{x}_i < 0$ at $\boldsymbol{x} = d\boldsymbol{e}_i$.

Finally, suppose that $h(0, (d, \mathbf{0}_{N-2})) < D_1C(0, 0)$, and suppose that $h(\cdot, \mathbf{0}_{N-1})$ intersects $D_1C(\delta, \cdot)$ from *below*. This implies that for some $\varepsilon > 0$, $h(x_i, \mathbf{0}_{N-1}) < D_1C(\delta, x_i)$ $\forall x_i \in (d - \varepsilon, d)$ and $h(x_i, \mathbf{0}_{N-1}) > D_1C(\delta, x_i)$ $\forall x_i \in (d, d + \varepsilon)$. Let

$$\mathbb{L} = \{\boldsymbol{x} \in [0, \chi]^N \, : \, x_i \in (d - \varepsilon, d + \varepsilon), \; x_j = 0 \; \forall j \neq i\} \subseteq B_\varepsilon(d\boldsymbol{e}_i).$$

Clearly $\dot{x}_j = 0 \; \forall \boldsymbol{x} \in \mathbb{L}$. Arbitrarily fix $\rho \in (0, \varepsilon)$, and suppose the initial power structure is $\boldsymbol{x}_0 = (d - \rho)\boldsymbol{e}_i$. Since

$$h(d - \rho, \mathbf{0}_{N-1}) < D_1C(\delta, d - \rho) \Rightarrow (D_1C)^{-1}(h(d - \rho, \mathbf{0}_{N-1}), d - \rho) - \delta < 0$$

It follows from (1.7) that $\dot{x}_i$ at $\boldsymbol{x}_0$. Since $\rho \in (d - \varepsilon, d)$ was arbitrary, it follows that $\dot{x}_i < 0$ at all $\boldsymbol{x} \in \{\boldsymbol{x} \in \mathbb{L} \, : \, x_i < d\}$. It then follows that for any $\rho \in (d - \varepsilon, d)$, if $\boldsymbol{x}_0 = (d - \rho)\boldsymbol{e}_i$, $\exists t > 0$ s.t. $\boldsymbol{x}_t \notin B_\varepsilon(d\boldsymbol{e}_i)$, thereby violating Definition 1.

Proof of Part 5 Arbitrarily fix $i \in \{1, ..., N\}$. Here, I show that Condition V

$$h\big(\chi, (0,..., 0)\big) > D_1C(\delta, \chi) \text{ and } h\big(0, (\chi, 0,..., 0)\big) < D_1C(0, 0)$$

is necessary and sufficient for $\bar{\boldsymbol{x}} = \chi\boldsymbol{e}_i$ to be stable. For ease of exposition, the first and second parts of Condition V are respectively referred to as "Condition V.1" and "Condition V.2," below.

First suppose that Condition V holds. It follows from Condition V.1 and equation 2 of (1.7) that $\dot{x}_i = 0$ at this $\chi\boldsymbol{e}_i$. By Condition V.2 and equation 1 of (1.7), it follows that $\dot{x}_j = 0 \; \forall j \neq i$ at $\bar{\boldsymbol{x}}$. This establishes that if Condition V holds, then $(\dot{x}_1, \dots, \dot{x}_N) = \mathbf{0}_N$ at $\boldsymbol{x} = \bar{\boldsymbol{x}}$. Therefore, the first part of Definition 1 is satisfied. As before, I use Lyapunov's method to show that $\bar{\boldsymbol{x}}$ also satisfies the second part of Definition 1 if Condition V holds. Since h is continuous and $D_1C(I, \cdot)$ is continuous for any fixed $I \geq 0$, it follows that for some $\varepsilon > 0$

$$h(x_i, \boldsymbol{x}_{-i}) > D_1C(\delta, x_i) \tag{A.33}$$

$$h(x_j, \boldsymbol{x}_{-j}) < D_1C(0, x_j) \; \forall j \neq i \tag{A.34}$$

both hold for all $\boldsymbol{x} \in B_\varepsilon(\bar{\boldsymbol{x}})$. It then follows that at every $\boldsymbol{x} \in B_\varepsilon(\bar{\boldsymbol{x}}) \setminus \{\bar{\boldsymbol{x}}\}$, $\dot{x}_i > 0$ if $x_i < \chi$ (by equation 3 of (1.7)), $\dot{x}_i = 0$ if $x_i = \chi$ (by equation 2 of (1.7)), and that for every $j \neq i$, $\dot{x}_j < 0$ if $x_j > 0$ and $\dot{x}_j = 0$ otherwise (by equation 1 of (1.7)). Like before, I use the Lyapunov function from (A.9) evaluated at $\bar{\boldsymbol{x}} = \boldsymbol{e}_i$:

$$\Lambda(\boldsymbol{x}) = -[(\chi - x_i) - \sum_{j \neq i} x_j] \tag{A.35}$$

whose time derivative is

$$\dot{\Lambda}(\boldsymbol{x}) = -(\chi - x_i)\dot{x}_i + \sum_{j \neq i} x_j \dot{x}_j. \tag{A.36}$$

Fix some $\boldsymbol{x} \in B_\varepsilon(\bar{\boldsymbol{x}}) \cap [0, \chi]^N =: \mathscr{V}$. Notice that the first term of (A.36) is strictly negative if $x_i < \chi$ and zero if $x_i = \chi$. The second term of (A.36) is equal to zero if and only if $x_j = 0 \; \forall j \neq i$ and strictly negative otherwise. It then follows that $\dot{\Lambda}(\boldsymbol{x}) < 0 \; \forall \boldsymbol{x} \in \mathscr{V} \setminus \{\bar{\boldsymbol{x}}\}$. This establishes that if Condition V holds, then $\bar{\boldsymbol{x}}$ is stable.

I now show that if Condition V is violated, then stable strong dictatorial power structures cannot exist. Let $\bar{\boldsymbol{x}} = \chi \boldsymbol{e}_i$ and ε be defined as above. Suppose that only V.1 is violated, so that

$$h(\chi, \mathbf{0}_{N-1}) \leq D_1 C(\delta, \chi) \tag{A.37}$$

$$h(0, (\chi, \mathbf{0}_{N-2})) < D_1 C(0, 0) \tag{A.38}$$

if the inequality in (A.37) strictly holds, then by equation 1 of (1.7) $\dot{x}_i < 0$. Therefore, the first part of the definition of stability is violated. ■

Proof of Part 6 The previous parts of this proposition established conditions on model primitives under which stable power structures in

$$\{0,\chi\}^N \sqcup \left\{d\boldsymbol{e}_i \,:\, d \in (0,\chi], i \in \{1,...,N\}\right\}$$

exist. I now show that no $\bar{\boldsymbol{x}}$ outside this set is ever stable under Assumptions 1 and 2. To prove this, it is useful to first define

$$\mathcal{K}_z(\bar{\boldsymbol{x}}) \equiv \{i \in \{1,...,N\} \,:\, \bar{\boldsymbol{x}}_i = z\}\,(z = 0,\chi); \quad \mathcal{K}_{int}(\bar{\boldsymbol{x}}) \equiv \{i \in \{1,...,N\} \,:\, \bar{\boldsymbol{x}}_i \in (0,\chi)\}.$$

Moreover, let $k_z(\bar{\boldsymbol{x}}) \equiv \#(\mathcal{K}_z(\bar{\boldsymbol{x}}))$ for each $z \in \{0,\chi\}$ and $k_{int}(\bar{\boldsymbol{x}}) \equiv \#(\mathcal{K}_{int}(\bar{\boldsymbol{x}}))$, where $\#(\cdot)$ outputs the cardinality of its input. Note that when there is little risk of confusion, $\mathcal{K}_z$, k_z, $\mathcal{K}_{int}$, and k_{int} may have their inputs suppressed.

First, let us consider an arbitrarily fixed $\bar{\boldsymbol{x}}$ where at least two players have interior levels of power (i.e. $k_{int}(\bar{\boldsymbol{x}}) \geq 2$). I now suppose that $\bar{\boldsymbol{x}}$ is stable and proceed to demonstrate that this yields a contradiction. If this supposition is true, then at $\boldsymbol{x} = \bar{\boldsymbol{x}}$ we have $\dot{x}_i = 0\ \forall\, i$ (by the first part of Definition 1). The following are then immediately implied by (1.7):

$$h(\bar{x}_i, \bar{\boldsymbol{x}}_{-i}) = D_1 C(\dot{x}_i + \delta, \bar{x}_i)\big|_{\dot{x}_i = 0}\ \forall i \in \mathcal{K}_{int}(\bar{\boldsymbol{x}}), \tag{A.39}$$

$$h(\bar{x}_i, \bar{\boldsymbol{x}}_{-i}) > D_1 C(\delta, \bar{x}_i)\ \forall i \in \mathcal{K}_{\chi}(\bar{\boldsymbol{x}}), \tag{A.40}$$

$$h(\bar{x}_i, \bar{\boldsymbol{x}}_{-i}) < D_1 C(0, \bar{x}_i)\ \forall i \in \mathcal{K}_{0}(\bar{\boldsymbol{x}}). \tag{A.41}$$

First consider the case where for two players $j, j' \in \mathcal{K}_{int}(\bar{\boldsymbol{x}})$ we have $0 < \bar{x}_j < \bar{x}_{j'} < \chi$. It then follows from Assumption 1 that $D_1 C(\delta, \bar{x}_{j'}) \leq D_1 C(\delta, \bar{x}_j)$. It is also straightforward to

verify that $h(\bar{x}_{j'}, \bar{\boldsymbol{x}}_{-j'}) > h(\bar{x}_j, \bar{\boldsymbol{x}}_{-j})$: let $a = e^{\lambda \bar{x}_j}$, $b = e^{\lambda \bar{x}_{j'}}$, and $y = \sum_{\ell \in \{i\}_1^N \setminus \{j,j'\}} e^{\lambda \bar{x}_\ell}$. Then,

$$h(\bar{x}_j, \bar{\boldsymbol{x}}_{-j}) = \frac{\lambda a(y+b)}{(a+b+y)^2}, \tag{A.42a}$$

$$h(\bar{x}_{j'}, \bar{\boldsymbol{x}}_{-j'}) = \frac{\lambda b(y+a)}{(a+b+y)^2}. \tag{A.42b}$$

Elementary algebra verifies that $h(\bar{x}_{j'}, \bar{\boldsymbol{x}}_{-j'}) > h(\bar{x}_j, \bar{\boldsymbol{x}}_{-j})$.

The inequalities $D_1 C(\delta, \bar{x}_{j'}) \leq D_1 C(\delta, \bar{x}_j)$ and $h(\bar{x}_{j'}, \bar{\boldsymbol{x}}_{-j'}) > h(\bar{x}_j, \bar{\boldsymbol{x}}_{-j})$, yield a contradiction in light of (A.39):

$$h(\bar{x}_{j'}, \bar{\boldsymbol{x}}_{-j'}) = D_1 C(\delta, \bar{x}_{j'}) \leq D_1 C(\delta, \bar{x}_j) = h(\bar{x}_j, \bar{\boldsymbol{x}}_{-j}) < h(\bar{x}_{j'}, \bar{\boldsymbol{x}}_{-j'}). \; \lightning$$

Therefore, there exist no stable $\bar{\boldsymbol{x}}$ in the set

$$\left\{ \boldsymbol{x} \in [0, \chi]^N \; : \; \exists j \in \{1, ..., N\}, \; j' \in \{1, ..., N\} \setminus \{j\} \text{ s.t. } 0 < x_j < x_{j'} < \chi \right\},$$

under Assumptions 1 and 2.

Now suppose that $\bar{x}_j = \alpha \; \forall j, j' \in \mathcal{K}_{int}(\bar{\boldsymbol{x}})$ for some $\alpha \in (0, \chi)$. It is possible to choose a sufficiently small ε so that the inequalities in (A.40) and (A.41) hold in the following subset of an ε ball of $\bar{\boldsymbol{x}}$:

$$\mathscr{A} \equiv \{ \boldsymbol{x} \in [0, \chi]^N \; : \; x_i = \bar{x}_i \; \forall i \in \mathcal{K}_0(\bar{\boldsymbol{x}}) \cup \mathcal{K}_\chi(\bar{\boldsymbol{x}}), |\boldsymbol{x} - \bar{\boldsymbol{x}}| < \varepsilon \}$$

This follows from the continuity of h and the fact that within this neighborhood only interior components of $\bar{\boldsymbol{x}}$ vary in $\mathscr{A}$. Now consider $\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \rho \sum_{i \in \mathcal{K}_{int}(\bar{\boldsymbol{x}})} \boldsymbol{e}_i$, where $\rho > 0$ is chosen so that $\hat{\boldsymbol{x}} \in \mathscr{A}$. Then, we have for each $i \in \mathcal{K}_{int}(\bar{\boldsymbol{x}})$

$$h(\hat{x}_i, \hat{\boldsymbol{x}}_{-i}) > h(\bar{x}_i, \bar{\boldsymbol{x}}_{-i}) = D_1 C(\delta, \bar{x}_i) \geq D_1 C(\delta, \hat{x}_i),$$

where the first inequality follows from the fact that

$$\frac{\partial}{\partial \alpha} h(\alpha, (\alpha \mathbf{1}_{k_{int}(\bar{\boldsymbol{x}})-1}, \chi \mathbf{1}_{k_\chi(\bar{\boldsymbol{x}})}, \mathbf{0}_{k_0(\bar{\boldsymbol{x}})})) > 0,$$

while the last inequality follows from Assumption 1 that $D_1 C(I, \cdot)$ is weakly decreasing for every fixed $I \geq 0$. Therefore if one perturbs $\boldsymbol{x}_t$ from $\bar{\boldsymbol{x}}$ to $\hat{\boldsymbol{x}}$, it follows from (1.7) that $\dot{x}_j = \dot{x}_{j'} \geq 0 \ \forall j, j' \in \mathcal{K}_{int}(\bar{\boldsymbol{x}})$ at the perturbed point. Note that by construction, $\dot{x}_{i\tau} = 0 \ \forall i \in \mathcal{K}_0(\bar{\boldsymbol{x}}) \sqcup \mathcal{K}_\chi(\bar{\boldsymbol{x}})$ at every $\tau \in [t, t']$. This implies that $\boldsymbol{x}_{t'}$ will leave the ε-ball of $\bar{\boldsymbol{x}}$ at some time $t > t'$, thereby ruling out stability.

The case that remains to be considered are the $\bar{\boldsymbol{x}} \in [0, \chi]^N$ such that $k_{int}(\bar{\boldsymbol{x}}) = 1$ and $k_0(\bar{\boldsymbol{x}}) < N - 1$.[6] By symmetry, I can focus on the case where $\bar{\boldsymbol{x}} = (\alpha, \chi \mathbf{1}_{k_\chi(\bar{\boldsymbol{x}})}, \mathbf{0}_{k_0(\bar{\boldsymbol{x}})})$ (for an arbitrarily fixed $\alpha \in (0, \chi)$) without loss; as before, rearranging the components of $\bar{\boldsymbol{x}}$ does not affect the proof (besides notation).

Suppose that $\dot{x}_i = 0 \ \forall i \in \{1, \dots, N\}$ at $\bar{\boldsymbol{x}}$ (if this were not true, then the first part of Definition 1 has already been violated). I will now show that the second part of Definition 1 is violated. By continuity, (A.40) and (A.41) respectively hold $\forall i \in \mathcal{K}_\chi(\bar{\boldsymbol{x}})$ and $\forall i \in \mathcal{K}_0(\bar{\boldsymbol{x}})$ when $\boldsymbol{x}$ is inside some sufficiently small ε-ball of $\bar{\boldsymbol{x}}$. Since $\frac{\partial}{\partial \alpha} h(\alpha, (\chi \mathbf{1}_{k_\chi(\bar{\boldsymbol{x}})}, \mathbf{0}_{k_0(\bar{\boldsymbol{x}})})) > 0$ and $D_1 C(\delta, \cdot)$ is weakly decreasing, it follows that at $\boldsymbol{x} = (\alpha + \rho, \chi \mathbf{1}_{k_\chi(\bar{\boldsymbol{x}})}, \mathbf{0}_{k_0(\bar{\boldsymbol{x}})})$, we have $\dot{x}_1 > 0 \ \forall \rho \in (0, \varepsilon)$. Define the following subset of the ε-ball of $\bar{\boldsymbol{x}}$:

$$\mathbb{L} = \{\boldsymbol{x} \in B_\varepsilon(\bar{\boldsymbol{x}}) \ : \ x_1 \in (\alpha, \alpha + \varepsilon), \ x_i = 0 \ \forall i \in \mathcal{K}_0(\bar{\boldsymbol{x}}), \ x_j = \chi \ \forall i \in \mathcal{K}_\chi(\bar{\boldsymbol{x}})\}$$

Given the above, it has been established that at every $\boldsymbol{x} \in \mathbb{L}$ $\dot{x}_1 > 0$ and $\dot{x}_j = 0 \ \forall j \in \{2, \dots, N\}$. It then follows that for any $\rho \in (0, \varepsilon)$, if the initial power structure is $\boldsymbol{x}_0 = (\alpha + \rho, \chi \mathbf{1}_{k_\chi(\bar{\boldsymbol{x}})}, \mathbf{0}_{k_0(\bar{\boldsymbol{x}})}) \in \mathbb{L}$, then $\boldsymbol{x}_t \in B_\varepsilon(\bar{\boldsymbol{x}})$ at some $t > 0$, thus violating the second part of

[6] Hence $k_\chi(\bar{\boldsymbol{x}}) \geq 1$; note that if $k_\chi(\bar{\boldsymbol{x}}) = 0$, this would correspond to the weak dictatorial case, which was covered in Part 4 of this proof.

Definition 1. ■

Proof of Proposition 4

Recall that by the proof of part 1 of Proposition 3, the $(\chi, \dots, \chi)$ is stable if and only if

$$h(\boldsymbol{x}, (\chi,\dots, \chi); N) > D_1 C(\delta, \chi) \tag{A.43}$$

which is equivalent to

$$\frac{(N-1)\lambda}{N^2} > D_1 C(\delta, \chi). \tag{A.44}$$

Rearranging the above yields the quadratic inequality

$$0 > D_1 C(\delta, \chi) N^2 - \lambda N + \lambda \tag{A.45}$$

When $\frac{\lambda}{4} \leq D_1 C(\delta, \chi)$, the escalated inclusive power structure is not stable for any N; this quickly follows from Condition I.[7] Otherwise, solving the above quadratic inequality for N yields

$$\frac{\lambda - \sqrt{\left(\lambda - 4D_1 C(\delta, \chi)\right)\lambda}}{2D_1 C(\delta, \chi)} < N < \frac{\lambda + \sqrt{\left(\lambda - 4D_1 C(\delta, \chi)\right)\lambda}}{2D_1 C(\delta, \chi)}. \tag{A.46}$$

It is easily verified that the left-most term is always less than two:

$$\frac{\lambda - \sqrt{\left(\lambda - 4D_1 C(\delta, \chi)\right)\lambda}}{2D_1 C(\delta, \chi)} - 2 = \frac{\left[\sqrt{\lambda - 4D_1 C(\delta, \chi)} - \sqrt{\lambda}\right]\sqrt{\lambda - 4D_1 C(\delta, \chi)}}{2D_1 C(\delta, \chi)} \leq 0.$$

Noting that N is a natural number implies that no escalated inclusive steady state exist

[7] Notice that if $h(\chi, \chi; 2) = \lambda/4 < D_1 C(\delta, \chi)$, then Condition I ($h(\boldsymbol{x}, (\chi,\dots, \chi); N) \geq D_1 C(\delta, \chi)$) fails at all $N \geq 2$ since $h(\boldsymbol{x}, (\chi,\dots, \chi); N) = \frac{(N-1)\lambda}{N^2}$ is decreasing in N on $\{2, 3, \dots\}$.

for group sizes larger than

$$\bar{N}_{\chi}^{I} \equiv \left\lceil \frac{\lambda + \sqrt{\Big(\lambda - 4D_1C(\delta,\chi)\Big)\lambda}}{2D_1C(\delta,\chi)} \right\rceil. \tag{A.47}$$

■

Proof of Corollary 1

Suppose that $D_1C(\delta,\chi) = q$ for some $q > 0$ and that $\lambda \geq 4q$. Note that

$$\frac{\partial}{\partial\lambda}\left[\frac{\lambda + \sqrt{\lambda\cdot(\lambda-4q)}}{2q}\right] = \frac{1}{2q}\left[\frac{(\lambda-2q) + \sqrt{\lambda\cdot(\lambda-4q)}}{\sqrt{\lambda\cdot(\lambda-4q)}}\right], \tag{A.48}$$

and that

$$\frac{\partial}{\partial q}\left[\frac{\lambda + \sqrt{\lambda\cdot(\lambda-4q)}}{2q}\right] = \frac{-\sqrt{\lambda}}{2q^2\sqrt{\lambda-4q}}\left[(\lambda-2q) + \sqrt{\lambda\cdot(\lambda-4q)}\right]. \tag{A.49}$$

Observe that equations (A.48) and (A.49) are respectively positive and negative if $(\lambda-2q)+\sqrt{\lambda\cdot(\lambda-4q)}$ is positive, which is always the case under the aforementioned supposition:

$$(\lambda-2q) + \sqrt{\lambda\cdot(\lambda-4q)} > \underbrace{(\lambda-4q)}_{\geq 0} + \underbrace{\sqrt{\lambda}}_{>0}\underbrace{\sqrt{\lambda-4q}}_{\geq 0} \geq 0$$

■

Proof of Proposition 5

Arbitrarily fix $\varepsilon > 0$; suppose that $D_1C(\delta,\cdot)$ is bounded by ε from below. By equation (A.46) in Proposition 4, the escalated inclusive power structure is stable only in

group sizes smaller than

$$\bar{N}_{\chi}^{I} \equiv \left\lfloor \frac{\lambda + \sqrt{(\lambda - 4D_1C(\delta,\chi))\lambda}}{2D_1C(\delta,\chi)} \right\rfloor$$

when $\lambda > 4D_1C(\delta,\chi)$, and is otherwise not stable in any group size permitted in this model ($N \in \{2,3,...\}$). If $\lambda \leq 4\varepsilon$, Proposition 5.1 trivially follows because of Assumption 1.2. Otherwise, $\bar{N}_{\chi}^{I}$ clearly remains finite as $\chi \to \infty$ given the aforementioned supposition.

Proposition 11 in Appendix section A.1.2 showed in equation (A.55) that for each $k \in \{2,3,...\}$, k-archic power structures are stable only in groups smaller than

$$\bar{N}_{\chi}^{O_k} \equiv \left\lfloor k + e^{\lambda\chi}\left(\frac{\lambda + \sqrt{(\lambda - D_1C(\delta,\chi))\lambda}}{D_1C(\delta,\chi)} - k\right)\right\rfloor$$

when $\lambda > D_1C(\delta,\chi)$ and not stable at any N otherwise. Similarly to before, if $\lambda < \varepsilon$ then Proposition 5.2 trivially follows because of Assumption 1.2. In the remaining case where $\exists z > 0$ s.t. $\lambda > D_1C(\delta,\chi)\ \forall \chi > z$, notice that the limit of $\bar{N}_{\chi}^{O_k}$ as $\chi \to \infty$ only depends on the sign of

$$\lim_{\chi\to\infty} \frac{\lambda + \sqrt{(\lambda - D_1C(\delta,\chi))\lambda}}{D_1C(\delta,\chi)} - k.$$

$\bar{N}_{\chi}^{O_k} \to \infty$ as $\chi \to \infty$ if the above is strictly negative and $\lim_{\chi\to\infty} \bar{N}_{\chi}^{O_k} \leq 0$ otherwise.

Proposition 12 in Appendix A.1.2 showed in equation (A.58) that dictatorships are stable only in groups smaller than

$$\bar{N}_{\chi}^{D} \equiv \left\lfloor e^{\lambda\chi} \cdot \left[\frac{\lambda - 2D_1C(\delta,\chi) + \sqrt{\lambda}\sqrt{\lambda - 4D_1C(\delta,\chi)}}{2D_1C(\delta,\chi)}\right]\right\rfloor$$

which becomes arbitrarily large as $\chi \to \infty$ given the aforementioned supposition.

■

Proof of Proposition 6

As in the above proofs, let $\boldsymbol{e}_i$ denote the i^{th} standard basis vector, and for each $n \in \mathbb{N}$ let $\mathbf{0}_n \in \mathbb{R}^n$ denote the vector of zeros. Recall that by Proposition 3, a dictatorial power structure where the strongest player has $d \in (0, \chi)$ units of power is stable if and only if

$$h\big(\cdot, (0,\ldots,0); N\big) \text{ intersects } D_1C(\delta, \cdot) \text{ from above at } d, \text{ and} \qquad \text{(Condition IV)}$$
$$h\big(0, (d, 0,\ldots,0); N\big) < D_1C(0, 0),$$

holds, and that strong dictatorial power structures are stable if and only if

$$h\big(\chi, (0,\ldots,0); N\big) > D_1C(\delta, \chi) \text{ and } h\big(0, (\chi, 0,\ldots,0; N)\big) < D_1C(0, 0). \qquad \text{(Condition V)}$$

holds. Arbitrarily fix $N \in \{2, 3, \ldots\}$ and all other model primitives (χ, δ, λ, and C) such that for each $M\{N, N+1\}$, either (1) Condition IV holds at exactly one $d_M \in (0, \chi)$ and condition V fails or (2) Condition IV fails at all $d \in (0, \chi)$ and Condition V holds. The result of the proof is immediate in case where Condition V holds when the group size is $N + 1$.

Now suppose that for each $M \in \{N, N+1\}$, Condition IV holds at exactly one $d_M \in (0, \chi)$ and condition V fails. Note that for all $n \in \{2, 3, \ldots\}$ and $\ell \in \{0, 1, 2, \ldots\}$,

$$h\left(x_i - \frac{1}{\lambda}\ln\left(\frac{n+\ell-1}{n-1}\right), \mathbf{0}_{n-1}; n\right) = h\left(x_i, \mathbf{0}_{n+\ell-1}; n+\ell\right). \qquad \text{(A.50)}$$

That is, given an initial group size of n, adding ℓ more players is equivalent to translating $h(\cdot, \mathbf{0}_{n-1}; n)$ rightward by $\frac{1}{\lambda}\ln\left(\frac{n+\ell-1}{n-1}\right)$. Moreover, observe that $h(\cdot, \mathbf{0}_{N-1}; N)$ can only intersect $D_1C(\delta, \cdot)$ from above after the former attains its global maximum at $x_i = \frac{\ln(N-1)}{\lambda}$

as it is assumed that both functions are continuous and $D_1C(I,x)$ is weakly decreasing in its second argument. It then follows that $d_M \in \left(\frac{\ln(M-1)}{\lambda}, \chi\right) \forall M \in \{N, N+1\}$. If $d_N \in \left(\frac{\ln(N-1)}{\lambda}, \frac{\ln(N)}{\lambda}\right)$, then $d_N < d_{N+1}$ follows from the fact that $d_{N+1} > \frac{ln(N)}{\lambda}$. If instead $d_N \in \left(\frac{\ln(N)}{\lambda}, \chi\right)$, note that $h(\cdot, \mathbf{0}_{N-1}; N)$ and $h(\cdot, \mathbf{0}_N; N+1)$ are strictly decreasing on $\left(\frac{\ln(N)}{\lambda}, \chi\right)$. Since the latter is a rightward translation of the former, and since $D_1C(\delta, \cdot)$ is decreasing, it follows that $d_N < d_{N+1}$. Note that when restricting attention to the interval $\left(\frac{\ln(N)}{\lambda}, \chi\right]$, translating $h(\cdot, \mathbf{0}_{N-1}; N)$ rightward is equivalent to translating it upward; this immediately yields a contradiction upon supposing that Condition V holds for N and Condition IV holds for $N+1$ at exactly one $d_{N+1} \in (0, \chi)$. Noting that $h(0, (d, \mathbf{0}_{N-2}); N)$ is decreasing in N and d $\forall (N, d, \lambda) \in \{2, 3, ...\} \times (0, \infty)^2$, the proof is complete. ■

Proof of Proposition 7

I consider without loss of generality case where player 1 is a (weak) dictator: $\hat{\boldsymbol{x}} = (d, \mathbf{0}_N)$, where $d \in (0, \chi)$ is as in the first part of Condition IV, which is reproduced and discussed in the proof of Proposition 6, found immediately above. Assume that $C(\cdot, \cdot)$ complies with Assumption 1. Fix some small $\varepsilon > 0$ and choose some $\tilde{C}$ such that $D_1\tilde{C}(\cdot, \cdot) = D_1C(\cdot, \cdot) + \varepsilon$. Since $h(\cdot, \mathbf{0}_N)$ intersects $D_1C(\delta, \cdot)$ from above at $d < \chi$, it follows from Assumptions 1 and 2 that $h(\cdot, \mathbf{0}_N) - D_1C(\delta, \cdot)$ is locally decreasing around d. Hence, for sufficiently small but positive ε, $h(\cdot, \mathbf{0}_N)$ intersects $D_1\tilde{C}(\delta, \cdot)$ at $\tilde{d} < d$. This completes the proof for part (i) of this proposition. Note that since C is assumed convex in its first argument, part (ii) immediately follows.

Turning to part (iii), we consider the effect of an increase in λ on d. Note that

$$\frac{\partial}{\partial \lambda} h(x_i, \mathbf{0}_{N-1}) = \underbrace{-\frac{(N-1)e^{\lambda x_i}}{\left[(N-1) + e^{\lambda x_i}\right]^3}}_{<0 \ \because N \geq 2} (1 - e^{\lambda x_i} - N + \lambda x + \lambda x e^{\lambda x_i} - \lambda N x) \tag{A.51}$$

Setting the second term to less than zero and rearranging yields the inequality in part (iii) of this proposition, thereby completing its proof. ■

A.1.2 Auxiliary Results

Proposition 11. *Let $k \in \{2, \dots, N-1\}$. k-archies are never stable past group size*

$$\bar{N}_{\chi}^{O_k} = \left\lceil k + e^{\lambda\chi}\left(\frac{\lambda + \sqrt{(\lambda - D_1C(\delta,\chi))\lambda}}{D_1C(\delta,\chi)} - k\right)\right\rceil \tag{A.52}$$

Proof. To simplify notation, let q_χ denote $D_1C(\delta,\chi)$ Recall that in the proof of Part 3 of Proposition 3 (found in Appendix A.1.1), it was established that

$$h(\chi, (\chi \mathbf{1}_{k-1}, \mathbf{0}_{N-k}); N) > q_\chi \tag{A.53}$$

is necessary for the stability of each element in

$$\left\{\boldsymbol{x} \in \{0,\chi\}^N \ : \ \sum_{i=1}^{N} x_i = k\chi\right\}.$$

Hence if $h(\chi, (\chi \mathbf{1}_{k-1}, \mathbf{0}_{N-k})) \leq q_\chi$, no element of the above set is stable at any value of N permitted in this model. Hence, assume that $h(\chi, (\chi \mathbf{1}_{k-1}, \mathbf{0}_{N-k}); N) > q_\chi$ for the remainder of this proof.

Note that (A.53) is equivalent to

$$\frac{\lambda\left[k - 1 + (N-k)e^{-\lambda\chi}\right]}{\left[k + (N-k)e^{-\lambda\chi}\right]^2} > q_\chi \tag{A.54}$$

Which yields the following quadratic inequality in N. Letting $\alpha = e^{-\lambda\chi}$ and $\beta = \frac{\lambda}{q_\chi}$, this

is as follows:

$$\alpha^2 N^2 + \alpha[2(1-\alpha)k - \beta]N + \left\{\left[(1-\alpha)k - \frac{\beta}{2}\right]^2 + \left(1 - \frac{\beta}{4}\right)\beta\right\} < 0$$

Note that the coefficient of N^2 is positive; by the formula for the vertex of a parabola, it follows that no N satisfies this inequality if $\beta \cdot \left(1 - \frac{\beta}{4}\right) > 0$ ($\Leftrightarrow \lambda < 4q_\chi$). Otherwise, solving the above quadratic inequality yields the following:

$$k + e^{\lambda\chi}\left[\frac{\lambda - \sqrt{(\lambda - D_1 C(\delta,\chi))\lambda}}{D_1 C(\delta,\chi)} - k\right] < N < k + e^{\lambda\chi}\left[\frac{\lambda + \sqrt{(\lambda - D_1 C(\delta,\chi))\lambda}}{D_1 C(\delta,\chi)} - k\right] \tag{A.55}$$

Therefore, k-archies are never stable if N is greater than or equal to

$$\bar{N}_\chi^{O_k} = \left\lceil k + e^{\lambda\chi}\left(\frac{\lambda + \sqrt{(\lambda - D_1 C(\delta,\chi))\lambda}}{D_1 C(\delta,\chi)} - k\right)\right\rceil$$

■

Proposition 12. *Suppose Condition IV holds for some N and χ, then there exist finite $\underline{N}_\chi^{D_W}, \bar{N}_\chi^{D_W}, \bar{N}_\chi^{D_S}$ such that*

1. *Weak dictatorships are stable if $\underline{N}_\chi^{D_W} \leq N < \bar{N}_\chi^{D_W}$.*
2. *Only strong dictatorships are stable if $\bar{N}_\chi^{D_W} \leq N \leq \bar{N}_\chi^{D_S}$*
3. *Weak and strong dictatorships are unstable if $N > \bar{N}_\chi^{D_S}$.*

Proof. Recall that the marginal benefit of investment for player i when her power is $x_i \in [0, \chi]$ and all other players have zero power is given by

$$h(x_i, \mathbf{0}_{N-1}; N) \equiv \frac{\lambda(N-1)e^{-\lambda x_i}}{(1 + (N-1)e^{-\lambda x_i})^2} \quad (\lambda > 0). \tag{A.56}$$

Recall that by (A.50) given an initial group size of N, adding K more players shifts marginal benefit $h(\cdot, \mathbf{0}_{N-1}; N)$ rightward.

Suppose that $d_N \boldsymbol{e}_i$ ($d_N \in (0,1)$) is stable when group size is $N \in \{2,3,...\}$. This is only possible if $h(\cdot, \mathbf{0}_{N-1}; N)$ intersects $D_1C(\delta, \cdot)$ from above at d_N. I demonstrate this via proof by contrapositive. If $h(\cdot, \mathbf{0}_{N-1}; N)$ is strictly greater than (strictly less than) $D_1C(\delta, \cdot)$ at d_N, then the player i's marginal benefit of investment is strictly greater than (strictly less than) her marginal cost when $\boldsymbol{x} = d_N \boldsymbol{e}_i$, hence $\dot{x}_i > 0$ ($\dot{x}_i < 0$) at this point. Therefore if $h(d_N, \mathbf{0}_{N-1}; N) \neq D_1C(\delta, d_N)$ then $d_N \boldsymbol{e}_i$ is not a steady state. If the intersection is from below, then $d_N \boldsymbol{e}_i$ is not stable. Let $\varepsilon > 0$. Perturbing x_i to $d_N + \varepsilon$ ($d_N - \varepsilon$) causes the marginal benefit of investment to become strictly greater than (strictly less than) the marginal cost for player i, thereby inducing $\dot{x}_i > 0$ ($\dot{x}_i < 0$) at this perturbed point. $d_N \boldsymbol{e}_i$ is not stable if $h(\cdot, \mathbf{0}_{N-1}; N)$ is tangent to $D_1C(\delta, \cdot)$ at d_N. This is shown through similar reasoning. Finally, we consider the case where $h(x_i, \mathbf{0}_{N-1}; N) = D_1C(\delta, x_i) \; \forall \; x_i \in (d_N - \varepsilon, d_N + \varepsilon)$ for some $\varepsilon > 0$. (That is, $h(\cdot, \mathbf{0}_{N-1}; N)$ and $D_1C(\delta, \cdot)$ overlap in some ε-neighborhood of $x_i = d_N$.) Note that if $d_N \boldsymbol{e}_i$ is a steady state, we must have that $h(0, (d_N, \mathbf{0}_{N-2}); N) < D_1C(\delta, 0)$ Otherwise $\dot{x}_j > 0 \; \forall j \neq i$ at this point. By the continuity of h and D_1C, there must be some η-neighborhood of $d_N \boldsymbol{e}_i$ throughout which this strict inequality holds. Consider the perturbation to $\boldsymbol{x}' = (d_N + \nu)\, \boldsymbol{e}_i$, where $0 < \nu < \min\{\varepsilon, \eta\}$. By construction $h(d_N + \nu, \mathbf{0}_{N-1}; N) = D_1C(\delta, d_N + \nu)$, so $\dot{x}_i = 0$ at this point. Similarly, $h(0, (d_N + \nu, \mathbf{0}_N); N) < D_1C(\delta, 0)$, so $\dot{x}_j = 0 \; \forall j \neq i$. Therefore a trajectory that begins at $\boldsymbol{x}_0 = \boldsymbol{x}'$ does *not* approach $d_N \boldsymbol{e}_i$ in the limit, thereby ruling out its stability. Note that $\max_{x_i \in \mathbb{R}} h(x_i, 0; 2) = \frac{\lambda}{4}$; this global maximum is attained at $x_i = 0$. Since (A.50) implies that $h(\cdot, \mathbf{0}_{N-1}; N)$ is a rightward horizontal translation of $h(\cdot, 0; 2)$ ($N = 2, 3, ...$), it follows that $\max_{x_i \in \mathbb{R}} h(x_i, \mathbf{0}_{N-1}; N) = \frac{\lambda}{4}$ for every such N. Recall that $h(x_i, \mathbf{0}_{N-1}; N)$ attains its global maximum (about which it is unimodal) at $x_i = \frac{\ln(N-1)}{\lambda}$. Notice that this is monotonically increasing in N when $N \geq 2$. Since $D_1(I, x)$ is weakly decreasing in its second argument, it follows that $\min_{x_i \in [0, \chi]} D_1C(\delta, x_i) = D_1C(\delta, \chi)$.

Finally, notice that $\lim_{x_i \to \infty} h(x_i, \mathbf{0}_N; N) = 0$. The desired result is immediate.

If $\lambda < 4D_1C(\delta, \chi)$ then $\bar{N} = 2$. Now assume $\lambda > 4D_1C(\delta, \chi)$ throughout the remaining duration of this proof. Since $h(x, \mathbf{0}_{N-1}; N)$ is unimodal about $\frac{\ln(N-1)}{\lambda}$ it follows that there exist exactly two values of N that solve $h(\chi, \mathbf{0}_{N-1}; N) = D_1C(\delta, \chi)$. These are

$$\widehat{N}_1 = 1 + \left(\frac{e^{\lambda\chi}}{2D_1C(\delta, \chi)}\right)\left(\lambda - 2D_1C(\delta, \chi) - \sqrt{\lambda}\sqrt{\lambda - 4D_1C(\delta, \chi)}\right) \tag{A.57}$$

and

$$\widehat{N}_2 = 1 + \left(\frac{e^{\lambda\chi}}{2D_1C(\delta, \chi)}\right)\left(\lambda - 2D_1C(\delta, \chi) + \sqrt{\lambda}\sqrt{\lambda - 4D_1C(\delta, \chi)}\right). \tag{A.58}$$

Let $\bar{N}_\chi^{D_W} = \left\lceil \widehat{N}_1 \right\rceil$. If $\widehat{N}_2 \in \mathbb{N}$, then let $\bar{N}_\chi^{D_S} = \widehat{N}_2 - 1$; otherwise let $\bar{N}_\chi^{D_S} = \left\lfloor \widehat{N}_2 \right\rfloor$. When N $= \bar{N}_\chi^{D_S}$, we know that $h(x_i, \mathbf{0}_{\bar{N}_\chi^{D_S}-1}; \bar{N}_\chi^{D_S}) > D_1C(\delta, x_i)\ \forall\ x_i \in (\chi - \varepsilon, \chi]$ for some $\varepsilon > 0$ and the reverse inequality holds in $[0, \chi - \varepsilon]$. Therefore the only stable dictatorships that exist are $\{\chi \boldsymbol{e}_i\}_1^{\bar{N}_\chi^{D_S}}$. It follows from (A.50) that for all $N > \bar{N}_\chi^{D_S}$, $h(x_i, \mathbf{0}_{N-1}; N) < D_1C(\delta, x_i)$ $\forall\ x_i \in [0, \chi]$. Therefore no dictatorial steady state can exist at any such N. By construction $h(x_i, \mathbf{0}_{\bar{N}_\chi^{D_W}-1}; \bar{N}_\chi^{D_W}) > D_1C(\delta, x_i)\ \forall\ x_i \in \left(\frac{\ln\left(\bar{N}_\chi^{D_W}-1\right)}{\lambda}, \chi\right]$. It then follows that the only dictatorial steady states that exist are $\{\chi \boldsymbol{e}_i\}_1^M$. It follows from (A.50) that the same is true for all $N \in \{\bar{N}_\chi^{D_W}, ..., \bar{N}_\chi^{D_S}\}$. ■

Remark 5. It is natural to expect that – given a *fixed* χ – dictatorships also become unfeasible once the group surpasses a certain size: powerless players, in sufficiently large numbers, overwhelm all dictators. This subject was considered in Proposition 5, which explores what happens when χ is made arbitrarily large. Note that a non-trivial *lower* bound $\underline{N}_\chi^W$ is possible; this follows from the fact mentioned earlier: powerless players always have a chance of winning conflicts.

A.1.3 Supplementary Figures

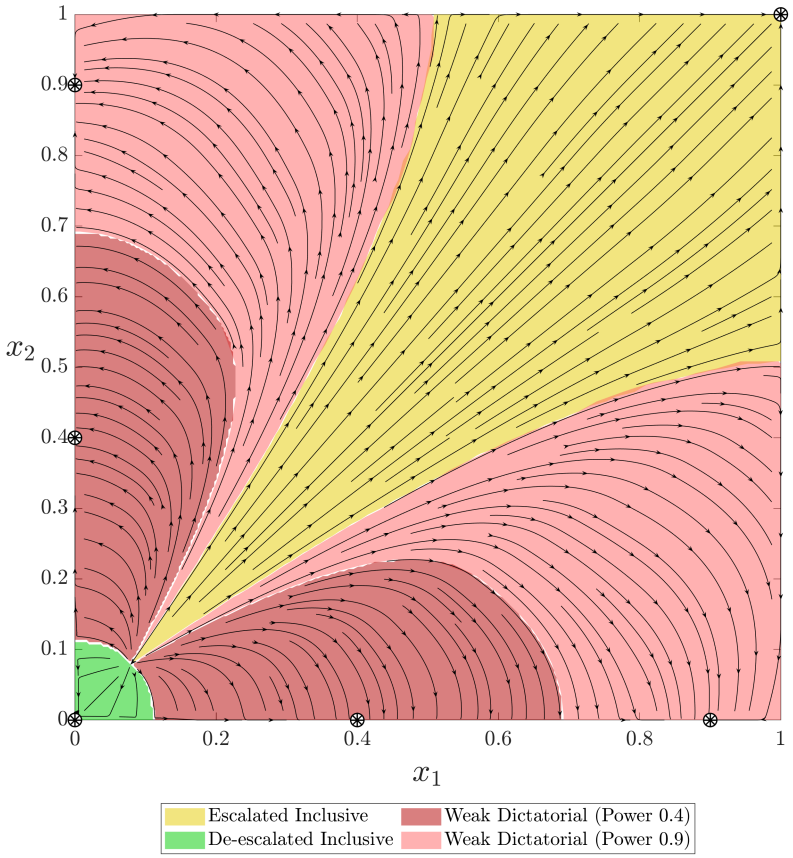


Figure A.1: Example of a case where two types of weak dictatorial power structures are stable. This simulation was generated using model primitives $\lambda = 3.5$, $\delta = 0.1$, $N = 2$, $\chi = 1$, and $C(I, x) = 0.77I^2 + \max\{0.8 - x, 0\}I$.

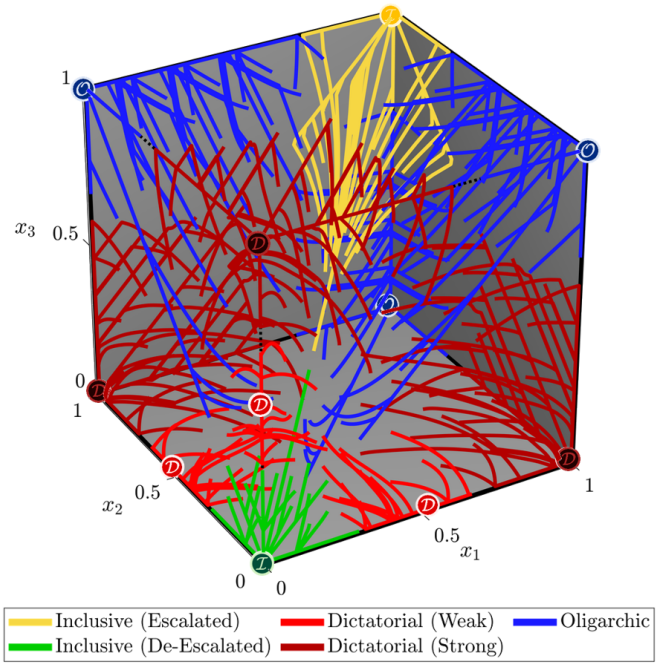


Figure A.2: Example of a three-player phase diagram where all possible classes (and subclasses) of stable power structures are featured. This simulation was generated using $N = 3$, $\lambda = 4$, $\delta = 0.1$, and the cost function, $C(I_{it}, x_{i,t-\Delta}) := 0.6I_{it}^2 + 1.37\max\{0.9 - x_{i,t-\Delta}, 0\}I_{it}$.

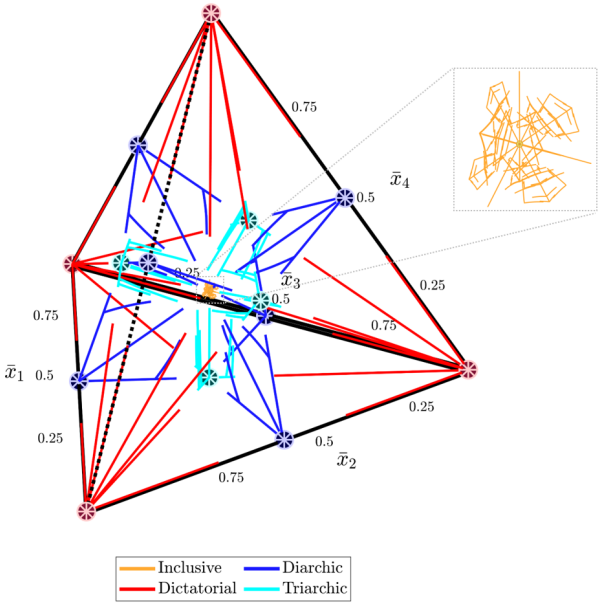
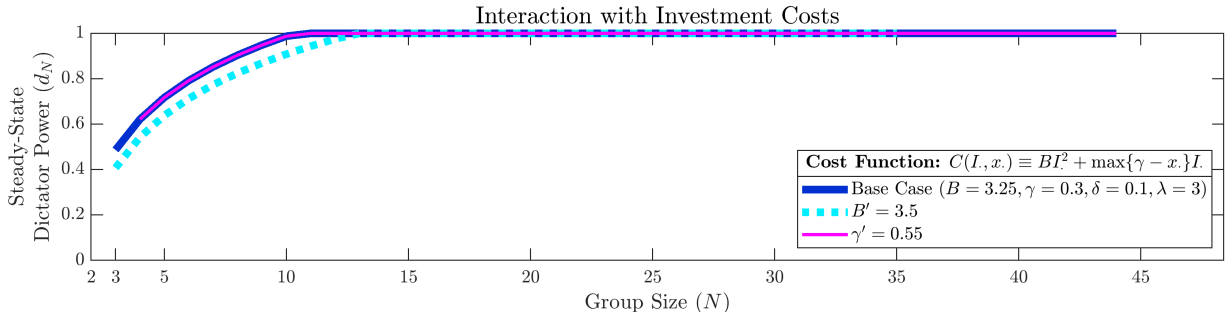
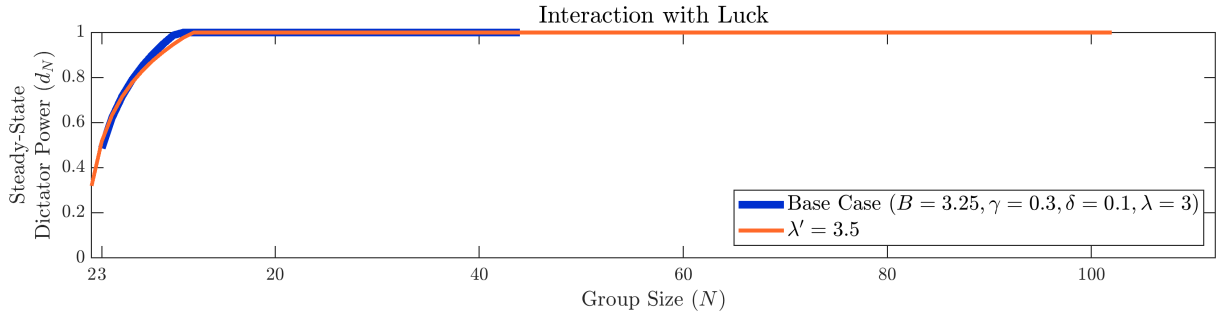


Figure A.3: Quaternary diagram depicting how the balance of power among four players evolves over time. This diagram was generated using the same parameter and function choices as in Figure 1.4. Like before, $\bar{x}_i := \frac{x_i}{x_1+x_2+x_3+x_4}$ denotes player i's share of the group's *aggregate* power.



(a) Increasing from B to B' increases the marginal cost of investment uniformly for all players; increasing γ to γ' increases marginal investment costs for players with $x_\cdot \in [\gamma, \gamma')$. Geometrically, increasing B (γ) shifts $D_1C(I, \cdot)$ upwards (rightwards) for every fixed I. Increasing B slows the growth rate of d_N and lowers $\bar{N}$, while increasing γ only has the effect of increasing $\underline{N}$.



(b) Increasing λ makes the outcome of conflict less noisy. This lowers $\underline{N}$ and raises both M and $\bar{N}$. Moreover, this increase has a non-monotonic effect on the path of d_N. This is to be expected, given part (iii) of Proposition 7.

Figure A.4: Comparative statics of $\{(N, d_N)\}_{N=2}^{\infty}$

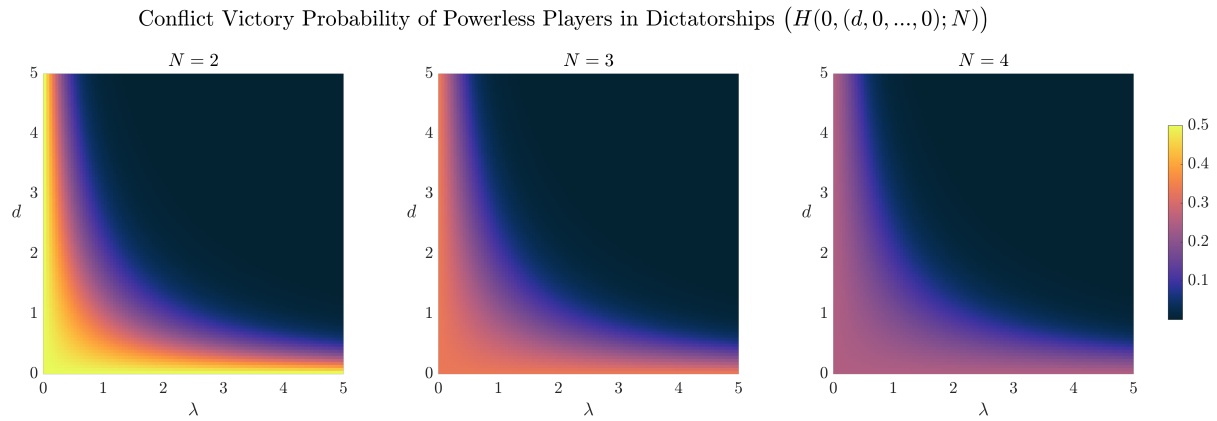


Figure A.5: Heatmaps of the probability $H(0, (d, 0 \ldots, 0); N)$ of a powerless player winning conflicts in a dictatorial power structure (where the dictator holds d units of power).

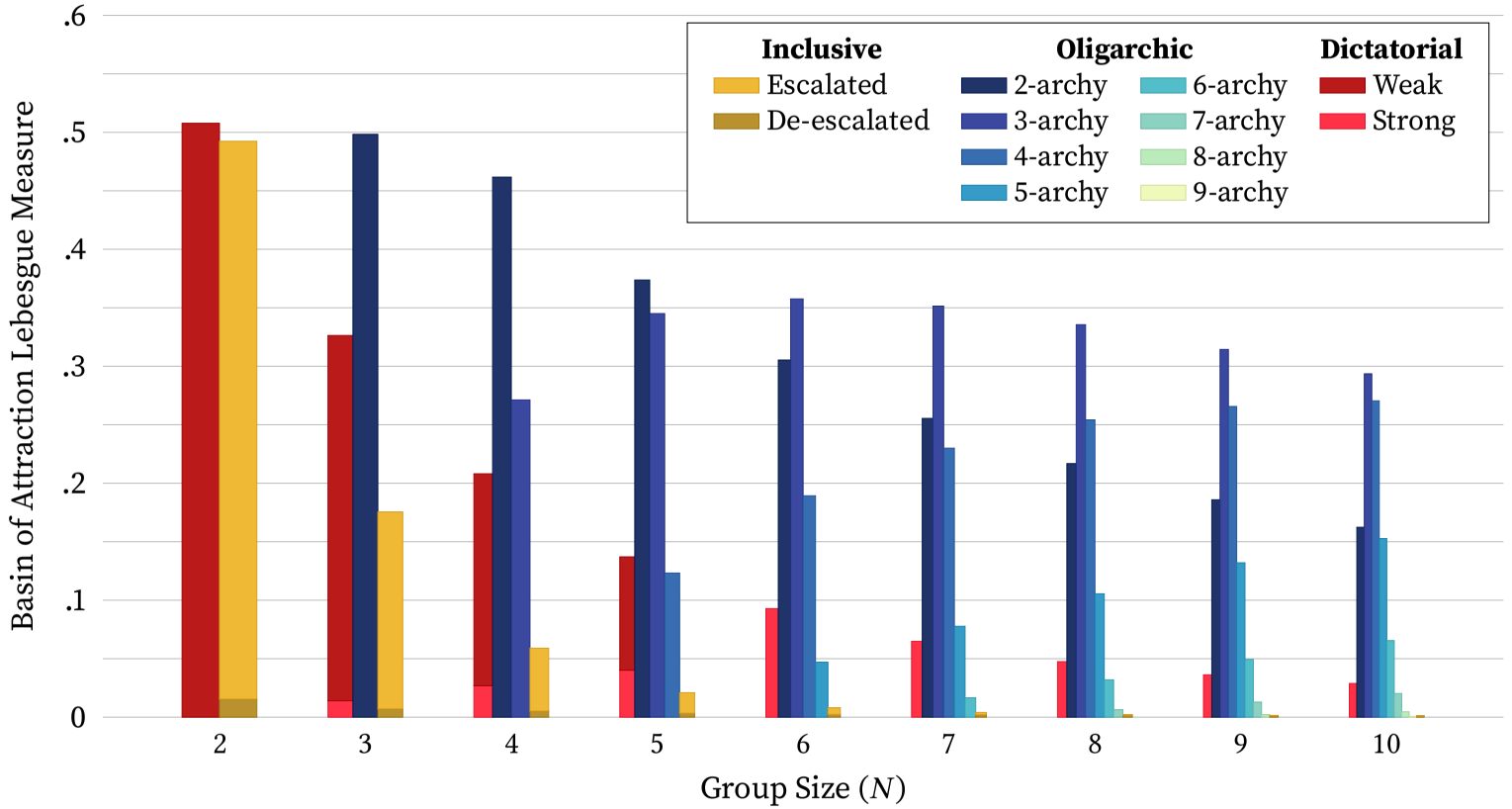


Figure A.6: Numerical approximation of the Lebesgue measure of the basins of attraction for each subclass of stable power structure for group sizes $N \in \{2, \dots, 10\}$ assuming $\lambda = 4$, $\delta = 0.1$ and investment cost $C(I_{it}, x_{i,t-\Delta}) = 1.5 \cdot I_{it}^2 + 1.5 \max\{0.5 - x_{i,t-\Delta}, 0\} I_{it}$.

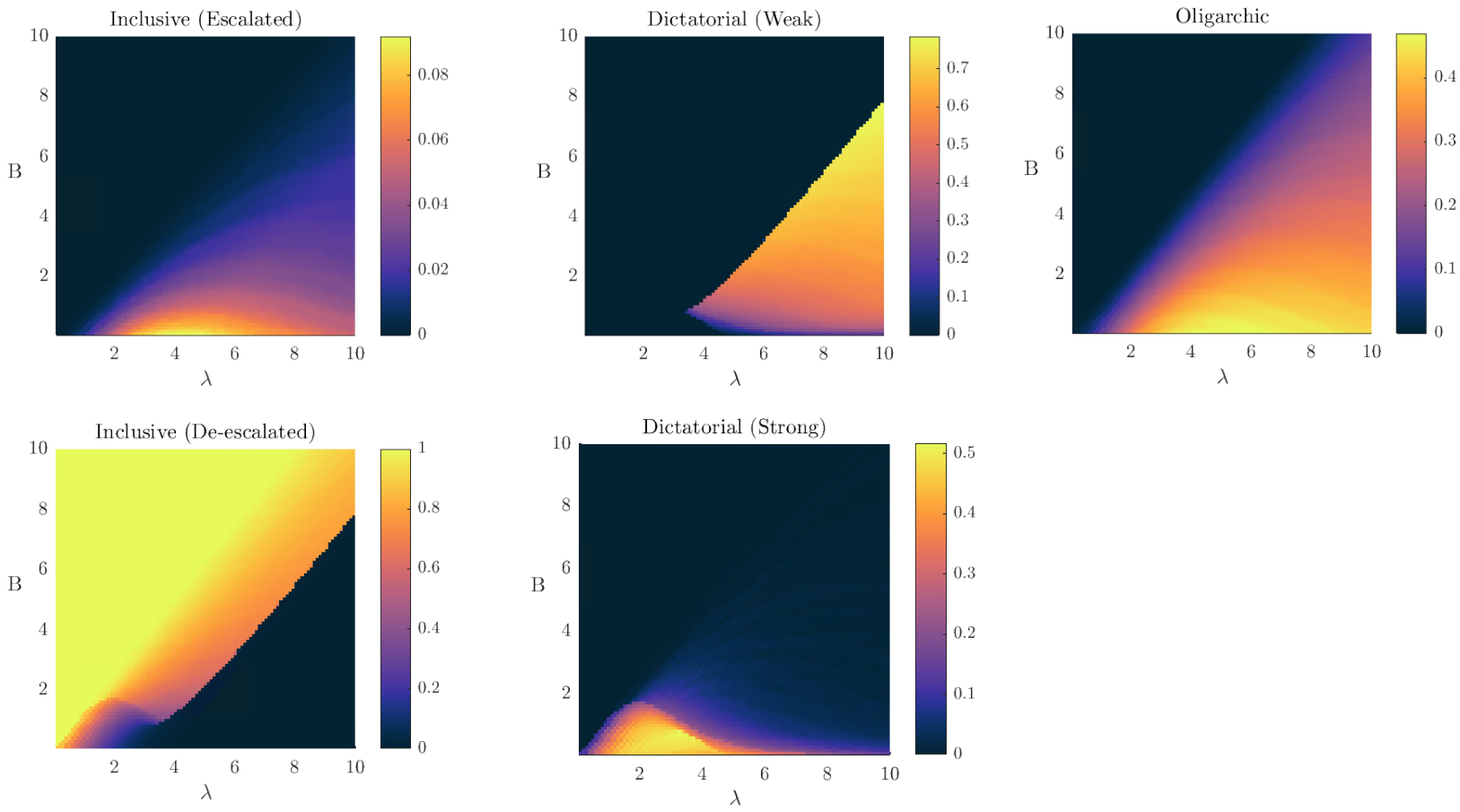


Figure A.7: Heatmaps visualizing the Lebesgue measure of each basin of attraction for various combinations of $(\lambda, B) \in [0.01, 10]^2$, where λ is the institutional constraint parameter and B is a parameter in the cost function, $C(I_{it}, x_{i,t-\Delta}) \equiv B \cdot I_{it}^2 + \max\{1 - x_{i,t-\Delta}, 0\} I_{it}$. These simulations were generated using $N = 3$, $\delta = 0.1$, $\chi = 1$, and the aforementioned cost function.

A.1.4 More on Contest Success Functions

In what follows, let $\mathcal{N} = \{1, \dots, N\}$ denote the set of lineages, where N is arbitrarily fixed. Moreover, throughout this section, all time subscripts will be suppressed and the player from lineage i will simply be referred to as "player i."

Role Played by Assumption 2

This section discusses the assumption (Assumption 2) made on $H(x_i, \boldsymbol{x}_{-i})$ is the *conditional* probability that player i wins an N-player conflict given that they hold power x_i and other players hold powers x_{-i}. Specifically, I assumed the functional form (1.3), reproduced below

$$H(x_i, \boldsymbol{x}_{-i}) \equiv \frac{e^{\lambda x_i}}{\sum_{j=1}^{N} e^{\lambda x_j}} = \frac{1}{1 + \sum\limits_{j \neq i} e^{-\lambda(x_i - x_j)}}, \quad (\lambda \geq 0). \tag{A.59}$$

This constitutes what is known as a Contest Success Function (CSF) in differences, since it only directly depends on power *differences*. While it would have been more general to directly assume that H is a continuous function of power differences, Skaperdas (1996, Theorem 3) shows that assuming the above functional form comes at very little additional loss of generality. Specifically, he shows that assuming that H satisfies the above functional form is equivalent to assuming that H is a continuous function of power differences that satisfies five additional axioms that – as we will now see – are all very mild.

In order to properly state these axioms, and Theorem 3 from Skaperdas (1996), a more general notion of contest success function is needed: specifically, one that accommodates "breakaway conflicts" where only players from a subset $\mathcal{M} \subseteq \mathcal{N}$ of lineages participate. Formally, let $p^i_{\mathcal{M}\mathcal{N}}(\boldsymbol{x})$ denote the probability that player i wins a

single-winner conflict given that the power structure is $\boldsymbol{x}$, the set of lineages is $\mathcal{N}$, and only players from lineages in $\mathcal{M} \subseteq \mathcal{N}$ participate.[8] The form of conflict considered in the present model has $\mathcal{M} = \mathcal{N}$ since players cannot be excluded from conflict.
The first three axioms are specifically in regards to $p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x})$, player i's probability of winning a conflict wherein all players participate when the power structure is $\boldsymbol{x}$. The first axiom states that $p^i_{\mathcal{N}\mathcal{N}}(\cdot)$ is a valid conditional probability distribution

Axiom 1. $\sum_{i=1}^{N} p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x}) = 1$ and $p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x}) \geq 0 \; \forall \; \boldsymbol{x}, i$

The next axiom states that a player's probability of winning a conflict (wherein all players participate) is increasing in how much power they hold, and decreasing in that held by each of their opponents

Axiom 2. $p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x})$ is increasing in x_j if $j = i$ and decreasing otherwise.

Axiom 3 is an anonymity condition that states that victory probabilities are independent of player identities

Axiom 3. Let $(\pi_1, \dots, \pi_N)$ be a permutation of $(1, \dots, N)$. If $x_j = \hat{x}_{\pi_i} \; \forall j$, then $p^i_{\mathcal{N}\mathcal{N}}(x_1, \dots, x_N) = p^{\pi_i}_{\mathcal{N}\mathcal{N}}(\hat{x}_{\pi_1}, \dots, \hat{x}_{\pi_N}) \; \forall i$.

The two remaining axioms relate to the victory of probability $p^{\cdot}_{\mathcal{M}\mathcal{N}}$ in conflicts wherein only a subset $\mathcal{M} \subset N$ of players participate.

Axiom 4. For each $\mathcal{M} \subseteq \mathcal{N}$ with at least two elements,

$$p^i_{\mathcal{M}\mathcal{N}}(\boldsymbol{x}) = \frac{p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x})}{\sum_{j \in \mathcal{M}} p^j_{\mathcal{N}\mathcal{N}}(\boldsymbol{x})} \; \forall i \in \mathcal{M} \text{ and } \forall \boldsymbol{x} \tag{A.60}$$

[8] Notice that $p^i_{\mathcal{M}\mathcal{N}}(\cdot)$ is formally a function of $\boldsymbol{x}$ – the vector of *all* player's powers – as opposed to the vector of powers held by players only in $\mathcal{M}$.

The left hand side of (A.60) is the conditional probability that player $i \in \mathcal{M}$ wins, given that only players in $\mathcal{M}$ participate, and that the power structure is $\boldsymbol{x}$. The right-hand side is the conditional probability that player $i \in \mathcal{M}$ wins, given that *all* players participate but that the winner is a player in $\mathcal{M}$ (and given that the power structure is $\boldsymbol{x}$).

The fifth and final axiom states that the probability of victory of those participating in a conflict is independent of power held by players excluded from participation.

Axiom 5. For each $\mathcal{M} \subseteq \mathcal{N}$ and $i \in \mathcal{N}$, $p^i_{\mathcal{M}\mathcal{N}}(\boldsymbol{x})$ is constant in $x_j \; \forall \; j \notin \mathcal{M}$.

Skaperdas (1996, Theorem 1) states that $p^i_{\mathcal{M}\mathcal{N}}(\boldsymbol{x})$ satisfies Axioms 1-5 if and only if there exists an increasing, positive function $f : \mathbb{R} \to \mathbb{R}$ such that

$$p^i_{\mathcal{M}\mathcal{N}}(\boldsymbol{x}) = \frac{f(x_i)}{\sum_{j \in \mathcal{M}} f(x_j)} \; \forall i \in \mathcal{M}, \forall \mathcal{M} \in \mathcal{N}, \; \forall \boldsymbol{x} \tag{A.61}$$

That is, stating that $p^i_{\mathcal{M}\mathcal{N}}(\boldsymbol{x})$ satisfies Axioms 1-5 is *equivalent* to stating that it takes the functional form in (A.61), for some f.

Finally, the following axiom formalizes the notion of a contest success function depending only on power differences

Axiom 6. $p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x}) = p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x} + (c, \dots, c)) \; \forall i$ and $\forall c \in \mathbb{R}$ s.t. $x_i + c \geq 0 \; \forall i$

We already know that if $p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x})$ satisfies Axioms 1-5, it must take the functional form in (A.61) for some f (Skaperdas, 1996, Theorem 1). (Skaperdas, 1996, Theorem 3) states that if one further assumes that $p^i_{\mathcal{N}\mathcal{N}}(\boldsymbol{x})$ satisfies Axiom 6 (which implies that it only depends on power differences) and that f is continuous, f must take the functional form

$$f(x_i) = e^{\lambda x_i} \; (\lambda > 0). \tag{A.62}$$

Remark 6. Corchón and Dahm (2010, p. 83) note that the above function is often referred to as an *effectivity function*, where $e^{\lambda x_i}$ corresponds to how *effectively* player i's power influences their victory probability.

A.2 Appendix of Chapter 2

A.2.1 Proof of Lemma 1

Genuine Players' ($i = 1, 2$) Problems

Given the vector of messages $\mathbf{m} \in \{L, H\}^3$ and types $(\theta_1, \theta_2) \in \{L, H\}^2$, (genuine) player $i \in \{1, 2\}$ has the following utility function:

$$u_i(x(\mathbf{m}), \theta_i) = -(x(\mathbf{m}) - \theta_i)^2, \tag{A.63}$$

where $x(\mathbf{m}) \equiv \mathcal{U}\{\text{mode}(\mathbf{m})\}$. Notice that under this simple setup, $x(\cdot)$ simplifies to

$$x(\mathbf{m}) \equiv \text{sgn}(m_1 + m_2 + m_3) \tag{A.64}$$

because of the facts that there are 3 voters and that $(L, H) = (-1, 1)$.
Let m_{k,θ_k} denote the message of player $k \in \{1, 2, 3\}$ who is of type $\theta_k \in \{L, H\} \cup \{T\}$. Then the expected utilities of players 1 and 2 are given by

$$\mathbb{E}_{\theta_2} u_1(m_{1\theta_1}, (m_{2\theta_2}, m_{3T}), \theta_1) = -p[x(m_{1\theta_1}, m_{2L}, m_{3T}) - \theta_1]^2 - (1-p)[x(m_{1\theta_1}, m_{2H}, m_{3T}) - \theta_1]^2$$

and

$$\mathbb{E}_{\theta_1} u_2(m_{2\theta_2}, (m_{1\theta_1}, m_{3T}), \theta_2) = -p[x(m_{1L}, m_{2\theta_2}, m_{3T}) - \theta_2]^2 - (1-p)[x(m_{1H}, m_{2\theta_2}, m_{3T}) - \theta_2]^2,$$

respectively.

To facilitate performing the calculations below, let

$\text{EU}_{1\theta_1}(m_{1\theta_1}, (m_{2L}, m_{2H}; m_{3T})) := \mathbb{E}_{\theta_2} u_1(m_{1\theta_1}, (m_{2\theta_2}, m_{3T}), \theta_1)$ and define

$\text{EU}_{1\theta_1}(m_{1\theta_1}, (m_{2L}, m_{2H}; m_{3T}))$ similarly.

Let $i = 1$ and $\theta_1 = L = -1$. To characterize player 1's optimal strategy, I consider the following 8 cases, summarized in the table, below:

	1	2	3	4	5	6	7	8
m_{2L}	L	L	H	H	L	L	H	H
m_{2H}	H	H	L	L	L	L	H	H
m_{3T}	L	H	L	H	L	H	L	H

Going through all of these cases exhausts all the possible (m_{2L}, m_{2H}, m_{3T}).

Case 1: $(m_{2L}, m_{2H}; m_{3T}) = (L, H; L)$

$EU_{1L}(-1, (-1, 1; -1)) = 0 > -4(1-p) = EU_{1L}(1, (-1, 1; -1)) \Rightarrow m^*_{1L}(-1, 1; -1) = -1.$

Case 2: $(m_{2L}, m_{2H}; m_{3T}) = (L, H; H)$

$EU_{1L}(-1, (-1, 1; 1)) = -4(1-p) > -4 = EU_{1L}(1, (-1, 1; 1)) \Rightarrow m^*_{1L}(-1, 1; 1) = -1.$

Case 3: $(m_{2L}, m_{2H}; m_{3T}) = (H, L; L)$

$EU_{1L}(-1, (1, -1; -1)) = 0 > -4 = EU_{1L}(1, (1, -1; -1)) \Rightarrow m^*_{1L}(1, -1; -1) = -1.$

Case 4: $(m_{2L}, m_{2H}; m_{3T}) = (H, L; H)$

$EU_{1L}(-1, (1, -1; 1)) = -4p > -4 = EU_{1L}(1, (1, -1; 1)) \Rightarrow m^*_{1L}(1, -1; 1) = -1.$

Case 5: $(m_{2L}, m_{2H}; m_{3T}) = (L, L; L)$

$EU_{1L}(-1, (-1, -1; -1)) = 0 = EU_{1L}(1, (-1, -1; -1)) \Rightarrow m^*_{1L}(-1, -1; -1) = \{-1, 1\}.$

Case 6: $(m_{2L}, m_{2H}; m_{3T}) = (L, L, H)$

$EU_{1L}(-1,(-1,-1;1)) = 0 > -4 = EU_{1L}(1,(-1,-1;1)) \Rightarrow m^*_{1L}(-1,-1;1) = -1.$

Case 7: $(m_{2L}, m_{2H}; m_{3T}) = (H, H; L)$

$EU_{1L}(-1,(1,1;-1)) = 0 > -4 = EU_{1L}(1,(1,1;-1)) \Rightarrow m^*_{1L}(1,1;-1) = -1.$

Case 8: $(m_{2L}, m_{2H}; m_{3T}) = (H, H; H)$

$EU_{1L}(-1,(1,1;1)) = 0 = EU_{1L}(1,(1,1;1)) \Rightarrow m^*_{1L}(1,1;1) = \{-1, 1\}.$

Summarizing the above,

$$m^*_{1L}(m_{2L}, m_{2H}; m_{3T}) = \begin{cases} \{-1,1\} & \text{, if } m_{2L} = m_{2H} = m_{3T} \\ -1 & \text{, otherwise} \end{cases}$$

The derivations for $m^*_{2L}(\cdot)$, $m^*_{1H}(\cdot)$, and $m^*_{2H}(\cdot)$ are very similar and are omitted. Let $i, j \in \{1,2\}$ $(i \neq j)$ and $\theta_i \in \{L, H\}$. Then,

$$m^*_{i\theta_i}(m_{jL}, m_{jH}, m_{3T}) = \begin{cases} \{-1,1\} & \text{, if } m_{2L} = m_{2H} = m_{3T} \\ \theta_i & \text{, otherwise.} \end{cases}$$

For simplicity, we impose that when

$$\mathrm{EU}_{i\theta_i}(m_{i\theta_i}, (m_{jL}, m_{jH}; m_{3T})) = \mathrm{EU}_{i\theta_i}(-m_{i\theta_i}, (m_{jL}, m_{jH}; m_{3T})),$$

player i (of type θ_i) will vote θ_i. Imposing this "indifference-breaking condition" yields

$$m^*_{i\theta_i}(m_{jL}, m_{jH}, m_{3T}) \equiv \theta_i \quad \text{(sincere voting by genuine types)}$$

Saboteur's Problem (Player 3)

Given θ_{-3} and $\mathbf{m}_{-3,\theta_{-3}}$,

$$u_3(m_{3T}, \mathbf{m}_{-3,\theta_{-3}}, \theta_{-3}) = -\sum_{j \neq 3} u_j(m_{j\theta_j}, \mathbf{m}_{-j,\theta_{-j}}, \theta_j)$$

Given the (common) prior, player 3's expected utility is as follows:

$$\mathbb{E}_{\theta_{-3}}[u_3(m_{3T}, \mathbf{m}_{-3,\theta_{-3}}, \theta_{-3})] \quad = \quad \sum_{\tau \in \{L,H\}^2} \mathbb{P}\{\theta_{-3} = \tau\} \cdot u_3(m_{3T}, \mathbf{m}_{-3,\tau}, \tau)$$

which can be explicitly written as

$$\begin{aligned}
\mathbb{E}_{\theta_{-3}}[u_3(m_{3T}, \mathbf{m}_{-3,\theta_{-3}}, \theta_{-3})] = \quad & p^2\big[-u_1(m_{1L}, (m_{2L}, m_{3T}), L) - u_2(m_{2L}, (m_{1L}, m_{3T}), L)\big] \\
& + p(1-p)\big[-u_1(m_{1L}, (m_{2H}, m_{3T}), L) - u_2(m_{2H}, (m_{1L}, m_{3T}), H)\big] \\
& + (1-p)(p)\big[-u_1(m_{1H}, (m_{2L}, m_{3T}), H) - u_2(m_{2L}, (m_{1H}, m_{3T}), L)\big] \\
& + (1-p)^2\big[-u_1(m_{1H}, (m_{2H}, m_{3T}), H) - u_2(m_{2H}, (m_{1H}, m_{3T}), H)\big].
\end{aligned}$$

Now, note the following:

(i) If $\theta_1 = \theta = \theta_2$, then player 3 is *not* pivotal, so

$$u_i(\theta, (\theta, m_{3T}), \theta) = 0 \;\; \forall\, m_{3T} \;\; (i = 1, 2;\; \theta = -1, 1)$$

(ii) If $\theta_i = L,\; \theta_j = H \;\; (i, j \in \{1, 2\},\; i \neq j)$, then player 3 *is* pivotal, so

$$u_i(-1, (1, m_{3T}), -1) = \begin{cases} 0 & \text{, if } m_{3T} = -1 \\ -4 & \text{, if } m_{3T} = 1 \end{cases}, \quad u_i(-1, (1, m_{3T}), 1) = \begin{cases} -4 & \text{, if } m_{3T} = -1 \\ 0 & \text{, if } m_{3T} = 1 \end{cases}.$$

It then follows that

$$\mathbb{E}_{\theta_{-3}}[u_3(m_{3T}, \mathbf{m}^*_{-3,\theta_{-3}}, \theta_{-3})] = 8p(1-p) \;\forall m_{3T} \in \{-1, 1\}$$

Hence, player 3 is indifferent between choosing $m_{3T} = -1$ and $m_{3T} = 1$ for all $p \in (0, 1)$. ■

A.2.2 Proof of Lemma 2

The structure of this derivation is *very* similar to that of the previous section. We first solve the genuine players' ($i = 1, 2$) problems using an exhaustive method. Afterward, we compare $\mathbb{E}_{\theta_{-3}}[u_3(-1, \mathbf{m}^*_{-3,\theta_{-3}}, \theta_{-3})]$ and $\mathbb{E}_{\theta_{-3}}[u_3(1, \mathbf{m}^*_{-3,\theta_{-3}}, \theta_{-3})]$ to derive $m^*_{3T}(\cdot)$. The only difference in this derivation is that now $x(\cdot)$ is given by (2.3) instead of (2.1).

Genuine Players' ($i = 1, 2$) Problems

Case 1: $EU_{1L}((-1,(-1,1;-1)) - EU_{1L}((1,(-1,1;-1)) = \frac{4}{3} - \frac{8p}{9} > 0$
$\Rightarrow m^*_{1L}((1,(-1,1;-1))) = -1,$

Case 2: $EU_{1L}((-1,(-1,1;1)) - EU_{1L}((1,(-1,1;1)) = \frac{20}{9} - \frac{8p}{9} > 0$
$\Rightarrow m^*_{1L}((1,(-1,1;1))) = -1,$

Case 3: $EU_{1L}((-1,(1,-1;-1)) - EU_{1L}((1,(1,-1;-1)) = \frac{8p}{9} + \frac{4}{9} > 0$
$\Rightarrow m^*_{1L}((1,(1,-1;-1))) = -1,$

Case 4: $EU_{1L}((-1,(1,-1;1)) - EU_{1L}((1,(1,-1;1)) = \frac{8p}{9} + \frac{4}{3} > 0$
$\Rightarrow m^*_{1L}((1,(1,-1;1))) = -1,$

Case 4: $EU_{1L}((-1,(1,-1;1)) - EU_{1L}((1,(1,-1;1)) = \frac{8p}{9} + \frac{4}{3} > 0$
$\Rightarrow m^*_{1L}((1,(1,-1;1))) = -1,$

Case 5: $EU_{1L}((-1,(-1,-1;-1)) - EU_{1L}((1,(-1,-1;-1)) = \frac{4}{9} > 0$
$\Rightarrow m^*_{1L}((1,(-1,-1;-1))) = -1,$

Case 6: $EU_{1L}((-1,(-1,-1;1)) - EU_{1L}((1,(-1,-1;1)) = \frac{4}{3} > 0$
$\Rightarrow m^*_{1L}((1,(-1,-1;1))) = -1,$

Case 7: $EU_{1L}((-1,(1,1;-1)) - EU_{1L}((1,(1,1;-1)) = \frac{4}{3} > 0$
$\Rightarrow m^*_{1L}((1,(1,1;-1))) = -1,$

Case 8: $EU_{1L}((-1,(1,1;1)) - EU_{1L}((1,(1,1;1)) = \frac{20}{9} > 0$
$\Rightarrow m^*_{1L}((1,(1,1;1))) = -1,$

Hence, player 1 (of type L) votes sincerely (i.e., $m^*_{1L}(\cdot) \equiv -1$. As before, the derivations for player 2 and for high types are omitted since they are *very* similar. Players 1 and 2 of either type vote sincerely (i.e., $m^*_{i\theta_i}(\cdot) \equiv \theta_i,\ \forall i \in \{1,2\}, \forall \theta_i \in \{-1,1\}$).

Saboteur's Problem (Player 3)

Solving player 3's problem is much simpler this time:

$$\mathbb{E}_{\theta_{-3}}[u_3(-1, \mathbf{m}^*_{-3,\theta_{-3}}, \theta_{-3})] - \mathbb{E}_{\theta_{-3}}[u_3(1, \mathbf{m}^*_{-3,\theta_{-3}}, \theta_{-3})] = \frac{8\,(p-1)^2}{9} - \frac{8\,p^2}{9} = \frac{8}{9} - \frac{16\,p}{9}$$

which is positive if and only if $p > \frac{1}{2}$. Hence,

$$m^*_{3T} \equiv \begin{cases} -1 & \text{, if } p < \frac{1}{2} \\ \{-1,1\} & \text{, if } p = \frac{1}{2} \\ 1 & \text{, if } p > \frac{1}{2} \end{cases}$$

■

A.2.3 Trolls' Behavior Under "Majority Rule"

When there are N agents, 2 types (γ_1 and γ_2), T trolls, and the voting mechanism is "Majority Rule", trolls will always vote for the less likely type.

Let $n_i(\theta) \in \{0, 1, ..., N\}$ denote the number of agents that are type γ_i ($i = 1, 2$), given the realization $\theta \in \{\gamma_1, \gamma_2\}^N$. Since there are two types, $n_2(\theta) \equiv N - n_1(\theta)$.[9] Let

$$\varphi(n_1) := \binom{N}{n_1} p^{n_1}(1-p)^{(N-n_1)} \tag{A.65}$$

denote the probability that n_1 players are of the first type.

[9]The input for n_1 will be suppressed throughout.

Notice that trolls are pivotal if and only if $|n_1 - n_2| \leq T$. This is equivalent to

$$\frac{N-T}{2} \leq n_1 \leq \frac{T+N}{2}.$$

For simplicity, assume that $\frac{N-T}{2}, \frac{N+T}{2} \notin \mathbb{N}$ and that $T < N$.[10] Trolls have two strategies to compare: vote for $\theta = \gamma_1$ or for $\theta = \gamma_2$. We will show that trolls optimally vote for γ_2 iff $p > \frac{1}{2}$.

Denote the trolls' message by m_T and let $EU_T(m_T)$ denote their expected utility from voting m_T.[11] Then, the trolls will optimally pick $m_T^* = \gamma_2$ iff $EU_T(\gamma_1) - EU_T(\gamma_2) < 0$. After some (omitted) algebraic simplification, we get

$$\begin{aligned}
EU_T(\gamma_1) - EU_T(\gamma_2) &= \sum_{n_1=\lfloor\frac{N-T}{2}\rfloor+1}^{\lfloor\frac{N+T}{2}\rfloor} \varphi(n_1)\left[(N-n_1)(\gamma_1-\gamma_2)^2 - n_1(\gamma_1-\gamma_2)^2\right] \\
&= \sum_{n_1=\lfloor\frac{N-T}{2}\rfloor+1}^{\lfloor\frac{N+T}{2}\rfloor} \varphi(n_1)(N-2n_1)(\gamma_1-\gamma_2)^2 \\
&= \sum_{n_1=\lfloor\frac{N-T}{2}\rfloor+1}^{\lfloor\frac{N}{2}\rfloor} \left(\varphi(n_1) - \varphi(N-n_1)\right)(N-2n_1)(\gamma_1-\gamma_2)^2,
\end{aligned}$$

where the last equality is due to the symmetry of $N - 2n_1$ around $n_1 = \frac{N}{2}$. Next, note

[10] These assumptions are by no means necessary. In a more exhaustive proof, where $N, T \in \mathbb{N}$, we would need to use some tie-breaking rule for when the $n_1 = \frac{N-T}{2}$ and $n_1 = \frac{N+T}{2}$. The assumption $T < N$ is just so that we focus on the less trivial part of the complete proof. When $T \geq N$, trolls are always pivotal.

[11] Here, trolls will always cast the same votes as one another, and hence can be treated as one player in this proof.

that due to the symmetry of binomial coefficients, we have

$$\varphi(n_1) - \varphi(N - n_1) = \binom{N}{n_1} \left(p^{n_1}(1-p)^{N-n_1} - p^{N-n_1}(1-p)^{n_1} \right),$$

which is negative for any $n_1 < \frac{N}{2}$ iff $p > \frac{1}{2}$. Thus, if $p > \frac{1}{2}$, we have $EU_T(\gamma_1) - EU_T(\gamma_2) < 0$, and the trolls' optimal strategy is to vote for the less likely type, $m = \gamma_2$.

"■"

A.2.4 Proof of Lemma 4

Proof. The difference between the designer's expected utility when trolls imitate γ_1 and γ_2 can be shown to satisfy

$$\begin{aligned}
\Delta EU = &- \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} \left(k \left(\frac{N-k}{N+T} \right)^2 + (N-k) \left(\frac{T+k}{N+T} \right)^2 \right) (\gamma_1 - \gamma_2)^2 \\
&+ \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} \left(k \left(\frac{T+N-k}{N+T} \right)^2 + (N-k) \left(\frac{k}{N+T} \right)^2 \right) (\gamma_1 - \gamma_2)^2 \\
\sim &\sum_{k=0}^{N} \binom{N}{k} \left(k \left(\frac{N-k}{N+T} \right)^2 + (N-k) \left(\frac{T+k}{N+T} \right)^2 \right) \left((1-p)^k p^{N-k} - p^k (1-p)^{N-k} \right)
\end{aligned}$$

The last line is due to us removing $(\gamma_1 - \gamma_2)^2$ to simplify the expression (it is a common positive term). Assuming N is odd[12], we can represent the sum above as follows:

$$\Delta EU \sim \sum_{k=0}^{\frac{N-1}{2}} \binom{N}{k} \left((1-p)^k p^{N-k} - p^k (1-p)^{N-k}\right) \left[k \left(\frac{N-k}{N+T}\right)^2 + (N-k)\left(\frac{T+k}{N+T}\right)^2 - \right.$$
$$\left. - (N-k)\left(\frac{k}{N+T}\right)^2 - k\left(\frac{T+N-k}{N+T}\right)^2 \right].$$

Note that for $k < \frac{N}{2}$, we have $(1-p)^k p^{N-k} - p^k (1-p)^{N-k} > 0$, since $p > \frac{1}{2}$. Also note:

$$k \left(\frac{N-k}{N+T}\right)^2 + (N-k)\left(\frac{T+k}{N+T}\right)^2 - (N-k)\left(\frac{k}{N+T}\right)^2 - k\left(\frac{T+N-k}{N+T}\right)^2 =$$
$$= (N-2k)\frac{T^2}{(N+T)^2} + 0 > 0,$$

since $k < \frac{N}{2}$.

Therefore, the difference ΔEU is strictly positive term by term. This implies that the designer achieves higher utility when trolls report more likely type ($\gamma = 2$ since we have $p > \frac{1}{2}$). Hence, trolls will optimize by picking the less likely type ($\gamma = 1$) to report. ■

A.2.5 Proof of Lemma 5

Proof. It can be shown that the expected utility of the designer (ex-ante) is given by

$$\mathbb{E}\left[\sum_{i=1}^{N} u_i(g(\theta, T), \theta_i)\right] = -\sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} \left(k\left(\frac{T+N-k}{N+T}\right)^2 + (N-k)\left(\frac{k}{N+T}\right)^2\right)$$
$$= -\frac{Np\left(N^2(1-p) + N(1-p)(2T-1) + T(2p+T-2)\right)}{(N+T)^2}$$

[12]The argument follows very similarly if N is even.

This sum is hard to interpret, but we can check comparative statics of it with respect to T and p. Straightforward algebraic calculations show that

$$\frac{\partial}{\partial T}\mathbb{E}\left[\sum_{i=1}^{N} u_i(g(\theta,T),\theta_i)\right] = -\frac{2Np(NT+p^2-3p+2)}{(N+T)^3} < 0,$$

which implies that having more trolls will strictly reduce the designer's welfare. This is expected, since more trolls will be able to bias the result of the mechanism in a more drastic way.

For the blind mechanism, the designer's expected utility under is given by

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{N} u_i(p,\theta_i)\right] &= -\sum_{k=0}^{N}\binom{N}{k}p^k(1-p)^{N-k}\left(k\,(1-p)^2+(N-k)\,(2-p)^2\right)\\ &= -Np(1-p).\end{aligned}$$

The blind mechanism performs better than the average-of-votes mechanism if and only if

$$-Np(1-p) > -Np\frac{N^2(1-p)+N(1-p)(2T-1)+T(2p+T-2)}{(N+T)^2}.$$

Rearranging and simplifying:

$$\begin{aligned}-(1-p) &> -\frac{N^2(1-p)+N(1-p)(2T-1)+T(2p+T-2)}{(N+T)^2}\\ p-1 &> -\frac{N^2(1-p)+N(1-p)(2T-1)+T(2p+T-2)}{(N+T)^2}\\ p &> \frac{pN^2+2pNT+(1-p)N+2(1-p)T}{(N+T)^2}\\ (N+T)^2p &> (N^2+2NT-N-2T)p+N+2T\\ p &> \frac{N+2T}{N+2T+T^2}.\end{aligned}$$

■

A.2.6 Proof of Proposition 1

Proof. Suppose that the observed number of votes for type γ_1 is $k \in \{0, 1, \dots, N+T\}$. The lowest number of trolls that can be in this number is 0 (if all of them voted for type γ_2), and the highest number is T (if all of them voted for type γ_1). Therefore, the highest possible number of genuine agents with type γ_1 is k and the lowest possible number is $\max\{k-T, 0\}$. The true distribution of genuine types is anything in the range from $(\max\{k-T,0\}, \min\{N-k+T, N\})$ to $(k, N-k)$.

Note that the optimal outcome is decreasing in the number of genuine agents of type γ_1. Hence, it follows that the largest outcome that could be optimal is

$$\underline{b} = \frac{\max\{k-T,0\}}{N}\gamma_1 + \frac{\min\{N-k+T,N\}}{N}\gamma_2, \text{ (if number of } \gamma_1 \text{ types is lowest)}$$

and the smallest outcome that could be optimal is

$$\overline{b} = \frac{k}{N}\gamma_1 + \frac{N-k}{N}\gamma_2. \text{ (if number of } \gamma_1 \text{ types is highest)}$$

Let $g(k)$ be the mechanism's outcome under the mechanism g. If $g(k) < \underline{b}$, the expected welfare under the mechanism can be improved if we set $g(k) = \underline{b}$. Similarly, if $g(k) > \overline{b}$, the expected welfare can be improved if we set $g(k) = \overline{b}$.

■

A.2.7 Proof of Proposition 2

Proof. Define a mechanism as an outcome rule $g : \{0, 1, \dots, N+T\} \to [\gamma_1, \gamma_2]$, where the argument is the number of votes for γ_1. The designer chooses this rule to maximize the ex-ante utility subject to the trolls' best response.

Suppose that for a given mechanism g' the trolls are not indifferent between $m = \gamma_1$

and $m = \gamma_2$. For concreteness, assume that they prefer $m = \gamma_1$. Fixing the trolls' action, the designer's ex-ante utility is continuous in $g(0)$, $g(1)$, ..., $g(N+T)$. Therefore, in a neighborhood[13] of g' the trolls' strategy can be treated as a constant in that it does not change when the designer slightly adjusts $g'(0)$, $g'(1)$, ..., $g'(N+T)$. There are two possibilities: either g' is a local (and global[14]) maximum, or the designer can improve upon it. We will prove that the former option is not possible.

If we fix the trolls' strategy at $m = \gamma_1$, the designer's best reply is to essentially "subtract" the trolls' votes from the total. That is, if the designer observes k votes for γ_1, she then knows that $k - T$ genuine agents have this type and $N + T - k$ genuine agents have the other type. Then the designer's optimal mechanism is g_1, where

$$\max_{g_1(k)} -(k-T)\left(g_1(k)-\gamma_1\right)^2 - (N+T-k)\left(g_1(k)-\gamma_2\right)^2$$
$$\implies g_1(k) = \frac{k-T}{N}\gamma_1 + \frac{N+T-k}{N}\gamma_2.$$

Note that g_1 completely neutralizes the trolls' influence and achieves the same utility level as under perfect information. This, however, cannot be the equilibrium, since the trolls can benefit by switching some of their votes to $m = \gamma_2$. In that case, the mechanism will not take optimal action given any distribution of votes, and the designer's ex-ante utility will be lower. Hence, the trolls would prefer to deviate from $m = \gamma_1$.

This implies that the designer's optimal mechanism cannot be g_1, or any mechanism for which the trolls strictly prefer message $m = \gamma_1$. A similar proof can be done for the mechanisms for which the trolls prefer $m = \gamma_2$. Hence, the optimal mechanism must make the trolls indifferent between the messages. ■

[13] I.e. a set of mechanisms g such that $\left\|\left(g(0), g(1), \dots, g(N+T)\right) - \left(g'(0), g'(1), \dots, g'(N+T)\right)\right\| < \epsilon$ for some $\epsilon > 0$, where $\|\cdot\|$ is the Euclidean norm.

[14] This is due to the concavity of the designer's utility function.

A.2.8 Proof of Proposition 3

Proof. Assume $N+T$ is odd. The proof below can be easily adapted to the case where $N+T$ is even.

We will show this result by proving that changing the majority rule g_{mr} to a supermajority rule $g_{smr}^{\hat{\alpha},\gamma_1}$ with $\hat{\alpha} = \frac{1}{2} + \frac{1}{N+T}$ is always strictly beneficial for the designer. This is true regardless of how the trolls respond to the change of the mechanism.

First, suppose that the trolls' strategy remains the same. Recall that in the majority rule, the trolls strictly prefer to vote for $\theta = \gamma_2$. Thus, the mechanism $g_{smr}^{\alpha,\gamma_1}$ will differ in its outcome from g_{mr} in only one instance: when there are $\hat{k} = \left[\frac{N+T}{2}\right] + 1$ votes for γ_2. The majority rule has $g_{mr}\left(\hat{k}\right) = \gamma_2$, but the supermajority rule has $g_{smr}^{\hat{\alpha},\gamma_1}\left(\hat{k}\right) = \gamma_1$. This is a beneficial change because in this instance there are more voters with type $\theta = \gamma_1$ than $\theta = \gamma_2$. Hence, the change to $g_{smr}^{\hat{\alpha},\gamma_1}$ leads to a higher expected welfare.

We will now verify that this improvement remains if the trolls switch their strategy from voting for $\theta = \gamma_2$ to voting for $\theta = \gamma_1$. In this case, $g_{smr}^{\hat{\alpha},\gamma_1}$ will have the same outcomes (and the same expected welfare) as g_{mr} if the trolls voted for $\theta = \gamma_1$ instead of $\theta = \gamma_2$. The expected welfare under g_{mr} is strictly higher if the trolls vote for $\theta = \gamma_1$ than if they vote for $\theta = \gamma_2$. This implies that the change to $g_{smr}^{\hat{\alpha},\gamma_1}$ leads to a higher expected welfare. ■

A.2.9 Proof of Proposition 4

Proof. Recall that for $\beta = 1$, trolls prefer voting for $\theta = \gamma_2$, since $p > \frac{1}{2}$. For β close to 1, this will remain true due to the expected welfare's continuity in β (see below). The

weighted-average-of-votes mechanism's expected welfare is

$$V(g^{\beta}_{am}) = -\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}\Bigg(i\left(\frac{N+T-i}{i\beta+(N+T-i)}\right)^2 + \\ + (N-i)\left(\frac{i\beta}{i\beta+(N+T-i)}\right)^2\Bigg)(\gamma_1-\gamma_2)^2.$$

To show the result, we will take the derivative with respect to β and show that it is negative at $\beta = 1$. We will drop the (positive) term $(\gamma_1 - \gamma_2)^2$ to simplify the calculations.

$$\begin{aligned}\frac{\partial}{\partial\beta}V(g^{\beta}_{am}) &\sim -\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}\Bigg(-\frac{2i^2(N+T-i)^2}{(i\beta+(N+T-i))^3}+ \\ &\qquad + (N-i)\frac{2i^2\beta(i\beta+(N+T-i))-2i^3\beta^2}{(i\beta+(N+T-i))^3}\Bigg) \\ &= -\sum_{i=0}^{N}\binom{N}{i}p^i(1-p)^{N-i}2i^2(N+T-i)\frac{(N+T-i)+(N-i)\beta}{(i\beta+(N+T-i))^3}.\end{aligned}$$

All terms are positive except for the negative sign at the front, so $\frac{\partial}{\partial\beta}V(g^{\beta}_{am}) < 0$ at $\beta = 1$. Due to continuity of the expected utility, the trolls' strategy will remain the same for a range $\beta \in (\bar{\beta}, 1)$ for some $\bar{\beta}$. Hence, the designer can improve upon the mechanism's expected welfare by reducing β. ■

A.2.10 Proof of Proposition 5

Proof of Proposition 5. We know that g is continuous, i.e.

$$\forall\epsilon > 0\ \exists\delta > 0 \text{ s.t. } \forall p, p' \in \Delta\Gamma,\ |p - p'| < \delta \ \Rightarrow\ |g(p) - g(p')| < \epsilon,$$

where $|\cdot|$ denotes the standard Euclidean norm.

Define $t = \min_{x \in g(\Delta\Gamma)} \mathbb{E}\left[\sum_{i=1}^{N} u_i(x, \theta_i)\right]$ and $p_t = \arg\min_{x \in g(\Delta\Gamma)} \mathbb{E}\left[\sum_{i=1}^{N} u_i(x, \theta_i)\right]$. Due to continuity of g, there exists a neighborhood of p_t where the mechanism's outcomes are close to t, the trolls' ideal outcome. We want to prove that as $T \to \infty$, the trolls will be able to get into that neighborhood no matter the distribution of other agents' types.

For T trolls, let $p(T)$ be a distribution of their votes (excluding normal agents) that is closest to p_t. Formally, let $F(T) = \{p \in \Delta\Gamma \mid \forall \gamma_i, p(\gamma_i) = \frac{k}{T} \text{ for some } k \in \mathbb{N}\}$, and define $p(T)$ as

$$p(T) = \min_{p \in F(T)} |p - p_t|.$$

Note that $F(T)$ for different $T \in \mathbb{N}$ can comprise points with arbitrary rational coordinates. Since rational numbers are dense in $\mathbb{R}$, it follows that we can always find sufficiently large T so that $|p(T) - p_t|$ is arbitrarily close to 0. Formally, for any $\delta > 0$ there exists $\hat{T}_1 \in \mathbb{N}$ such that

$$T > \hat{T}_1 \;\Rightarrow\; |p(T) - p_t| < \frac{\delta}{2}.$$

This will be useful later on.

$p(T)$ is the distribution of votes that trolls generate on their own. Now consider possible variations that can be introduced to $p(T)$ due to normal agents' (sincere) votes. Let $\theta = (\theta_1, \dots, \theta_N)$ be vector of normal agents' types, and let $p(\theta, T)$ be the distribution of votes with normal agents and trolls included. That means

$$p(\theta, T)(\gamma_i) = \frac{T \cdot p(T)(\gamma_i) + \sum_{j=1}^{N} \mathbb{1}\{\theta_j = \gamma_i\}}{N + T}.$$

The distance between total distribution of votes and trolls' distribution of votes is

given by

$$|p(T) - p(\theta, T)| = \sqrt{\sum_{\gamma_i \in \Gamma} \Big(p(T)(\gamma_i) - p(\theta, T)(\gamma_i) \Big)^2}$$
$$= \sqrt{\sum_{\gamma_i \in \Gamma} \left(\frac{1}{N+T} \cdot \left(N \cdot p(T)(\gamma_i) - \sum_{j=1}^{N} \{\theta_j = \gamma_i\} \right) \right)^2}.$$

Due to quadratic nature of the norm, it will achieve its maximum when all normal agents are of the same type. Thus, the distance will take form

$$|p(T) - p(\theta, T)| = \sqrt{\left(\frac{N}{N+T} \cdot (p(T)(\gamma) - 1) \right)^2 + \sum_{\gamma_i \neq \gamma} \left(\frac{N}{N+T} p(T)\gamma_i \right)^2}$$
$$= \frac{N}{N+T} \cdot \sqrt{\Big(p(T)(\gamma) - 1 \Big)^2 + \sum_{\gamma_i \neq \gamma} \Big(p(T)\gamma_i \Big)^2}$$

for some $\gamma \in \Gamma$. Clearly, $\lim_{T \to \infty} |p(T) - p(\theta, T)| = 0$, since the expression under the square root is bounded by $|\Gamma|$, which is finite. Thus, for any $\delta > 0$ there exists $\hat{T}_2 \in \mathbb{N}$ such that

$$T > \hat{T}_2 \;\Rightarrow\; |p(T) - p(\theta, T)| < \frac{\delta}{2} \text{ for any } \theta.$$

Now we can take the maximum of $\hat{T}_1$ and $\hat{T}_2$ that will ensure

$$T > \hat{T} = \max\{\hat{T}_1, \hat{T}_2\} \;\Rightarrow\; |p(T) - p_t| < \frac{\delta}{2} \text{ and } |p(T) - p(\theta, T)| < \frac{\delta}{2}.$$

Finally, note that

$$|p(\theta, T) - p_t| \leq |p(\theta, T) - p(T)| + |p(T) - p_t|.$$

Therefore, we can conclude that

$$T > \hat{T} \;\Rightarrow\; |p(\theta, T) - p_t| < \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

Hence,

$$\forall \epsilon > 0 \; \exists \hat{T} \in \mathbb{N} \text{ such that } T > \hat{T} \;\; |g(p(\theta, T)) - g(p_t)| < \epsilon.$$

If T is large enough, trolls can guarantee that the outcome of mechanism g is arbitrarily close to the ex-ante worst-case outcome $g(p_t)$. This finishes the proof.

■

A.3 Appendix of Chapter 3

A.3.1 Formulæ for the Receiver's posterior beliefs over the state

This subsection provides explicit formulæfor the Receiver's posterior belief over the state ω formed using Bayes' rule after observing message $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$ on the equilibrium path. In the interest of making notation more compact, let

$$\sigma(m_{z_1,z_2}|X_1, X_2, R) \equiv \tilde{\sigma}^{z_1 z_2}_{X_1 X_2 R} \quad (R \in \{1,2\};\ (X_1, X_2) \in \{L, H\}^2; (z_1, z_2) \in \{\emptyset, L, H\}^2).$$

If m_{HL} is sent on the equilibrium path (i.e. if $\max\{\tilde{\sigma}^{HL}_{HL1}, \tilde{\sigma}^{HL}_{HL2}\} > 0$) then

$$\mu^R(m_{HL}) = \frac{\gamma_1(1-\nu_2)\tilde{\sigma}^{HL}_{HL1}\mu}{\gamma_1(1-\nu_2)\tilde{\sigma}^{HL}_{HL1}\mu + \gamma_2\nu_1\tilde{\sigma}^{HL}_{HL2}(1-\mu)}. \tag{A.66}$$

If m_{LH} is sent on the equilibrium path (i.e. if $\max\{\tilde{\sigma}^{LH}_{LH1}, \tilde{\sigma}^{LH}_{LH2}\} > 0$) then

$$\mu^R(m_{LH}) = \frac{\gamma_2(1-\nu_1)\tilde{\sigma}^{LH}_{LH2}\mu}{\gamma_2(1-\nu_1)\tilde{\sigma}^{LH}_{LH2}\mu + \gamma_1\nu_2\tilde{\sigma}^{LH}_{LH1}(1-\mu)}. \tag{A.67}$$

If $m_{H\varnothing}$ is sent on the equilibrium path (i.e. if $\max\{\tilde{\sigma}^{H\varnothing}_{X_1X_2R} : X_1 = H\} > 0$) then

$$\mu^R(m_{H\varnothing}) = \frac{\left[\gamma_1(1\text{-}\nu_2)\tilde{\sigma}^{H\varnothing}_{HL1} + \gamma_1\nu_2\tilde{\sigma}^{H\varnothing}_{HH1} + \gamma_2\nu_1\tilde{\sigma}^{H\varnothing}_{HH2}\right]\mu}{\left[\gamma_1(1\text{-}\nu_2)\tilde{\sigma}^{H\varnothing}_{HL1} + \gamma_1\nu_2\tilde{\sigma}^{H\varnothing}_{HH1} + \gamma_2\nu_1\tilde{\sigma}^{H\varnothing}_{HH2}\right]\mu + \gamma_2\nu_1\tilde{\sigma}^{H\varnothing}_{HL2}(1\text{-}\mu)}. \tag{A.68}$$

If $m_{\varnothing H}$ is sent on the equilibrium path (i.e. if $\max\{\tilde{\sigma}^{\varnothing H}_{X_1X_2R} : X_2 = H\} > 0$) then

$$\mu^R(m_{\varnothing H}) = \frac{\left[\gamma_1\nu_2\tilde{\sigma}^{\varnothing H}_{HH1} + \gamma_2(1-\nu_1)\tilde{\sigma}^{\varnothing H}_{LH2}\gamma_2\nu_1\tilde{\sigma}^{\varnothing H}_{HH2}\right]\mu}{\left[\gamma_1\nu_2\tilde{\sigma}^{\varnothing H}_{HH1} + \gamma_2(1-\nu_1)\tilde{\sigma}^{\varnothing H}_{LH2}\gamma_2\nu_1\tilde{\sigma}^{\varnothing H}_{HH2}\right]\mu + \gamma_1\nu_2\tilde{\sigma}^{\varnothing H}_{LH1}(1\text{-}\mu)}. \tag{A.69}$$

If $m_{L\varnothing}$ is sent on the equilibrium path (i.e. if $\max\{\tilde{\sigma}^{L\varnothing}_{X_1X_2R} : X_1 = L\} > 0$) then

$$\mu^R(m_{L\varnothing}) = \frac{\gamma_2(1-\nu_1)\tilde{\sigma}^{L\varnothing}_{LH2}\mu}{\gamma_2(1-\nu_1)\tilde{\sigma}^{L\varnothing}_{LH2}\mu + \left[\gamma_1(1-\nu_2)\tilde{\sigma}^{L\varnothing}_{LL1} + \gamma_1\nu_2\tilde{\sigma}^{L\varnothing}_{LH1} + \gamma_2(1-\nu_1)\tilde{\sigma}^{L\varnothing}_{LL2}\right](1\text{-}\mu)}. \tag{A.70}$$

If $m_{\varnothing L}$ is sent on the equilibrium path (i.e. if $\max\{\tilde{\sigma}^{\varnothing L}_{X_1X_2R} : X_1 = L\} > 0$) then

$$\mu^R(m_{\varnothing L}) = \frac{\gamma_1(1-\nu_2)\tilde{\sigma}^{\varnothing L}_{HL1}\mu}{\gamma_1(1-\nu_2)\tilde{\sigma}^{\varnothing L}_{HL1}\mu + \left[\gamma_1(1-\nu_2)\tilde{\sigma}^{\varnothing L}_{LL1} + \gamma_2(1-\nu_1)\tilde{\sigma}^{\varnothing L}_{LL2} + \gamma_2\nu_1\tilde{\sigma}^{\varnothing L}_{HL2}\right](1\text{-}\mu)}. \tag{A.71}$$

Table A.1: $\mu^R(m_{HL}) \equiv \text{Prob}\{\omega = H | m = m_{HL}\}$

$\sigma(m_{HL}\|X_1,X_2,C)$		
(H,L,1)	(H,L,2)	$\mu^R(m_{HL})$
0	0	any value in $[0,1]$
1	0	1
0	1	0
1	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2\nu_1(1-\mu)}$

Table A.2: $\mu^R(m_{LH}) \equiv \text{Prob}\{\omega = H | m = m_{LH}\}$

$\sigma(m_{LH}\|X_1,X_2,C)$		
(L,H,1)	(L,H,2)	$\mu^R(m_{LH})$
0	0	any value in $[0,1]$
1	0	0
0	1	1
1	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+\gamma_1\nu_2(1-\mu)}$

Table A.3: $\mu^R(m_{H\varnothing}) \equiv \text{Prob}\{\omega = H | m = m_{H\varnothing}\}$

$\sigma(m_{H\varnothing}\|X_1, X_2, C)$				
(H,L,1)	(H,H,1)	(H,H,2)	(H,L,2)	$\mu^R(m_{H\varnothing})$
0	0	0	0	any value in $[0,1]$
at least one non-zero			0	1
0	0	0	1	0
1	0	0	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2\nu_1(1-\mu)}$
0	1	0	1	$\frac{\gamma_1\nu_2\mu}{\gamma_1\nu_2\mu+\gamma_2\nu_1(1-\mu)}$
0	0	1	1	μ
1	1	0	1	$\frac{\gamma_1\mu}{\gamma_1\mu+\nu_1\gamma_2(1-\mu)}$
1	0	1	1	$\frac{[(1-\nu_2)\gamma_1+\nu_1\gamma_2]\mu}{[(1-\nu_2)\gamma_1+\nu_1\gamma_2]\mu+\nu_1\gamma_2(1-\mu)}$
0	1	1	1	$\frac{[\gamma_1\nu_2+\gamma_2\nu_1]\mu}{[\gamma_1\nu_2+\gamma_2\nu_1]\mu+\nu_1\gamma_2(1-\mu)}$
1	1	1	1	$\frac{(1+\gamma_2\nu_1)\mu}{(1+\gamma_2\nu_1)\mu+\gamma_2\nu_1(1-\mu)}$

Table A.4: $\mu^R(m_{\varnothing H}) \equiv \text{Prob}\{\omega = H | m = m_{\varnothing H}\}$

$\sigma(m_{\varnothing H}\|X_1, X_2, C)$				
(L,H,2)	(H,H,1)	(H,H,2)	(L,H,1)	$\mu^R(m_{\varnothing H})$
0	0	0	0	any value in $[0,1]$
at least one non-zero			0	1
0	0	0	1	0
1	0	0	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+\gamma_1\nu_2(1-\mu)}$
0	1	0	1	μ
0	0	1	1	$\frac{\gamma_2\nu_1\mu}{\gamma_2\nu_1\mu+\gamma_1\nu_2(1-\mu)}$
1	1	0	1	$\frac{[\gamma_2(1-\nu_1)+\gamma_1\nu_2]\mu}{[\gamma_2(1-\nu_1)+\gamma_1\nu_2]\mu+\gamma_1\nu_2(1-\mu)}$
1	0	1	1	$\frac{\gamma_2\mu}{\gamma_2\mu+\gamma_1\nu_2(1-\mu)}$
0	1	1	1	$\frac{[\gamma_1\nu_2+\gamma_2\nu_1]\mu}{[\gamma_1\nu_2+\gamma_2\nu_1]\mu+\gamma_1\nu_2(1-\mu)}$
1	1	1	1	$\frac{(\gamma_1\nu_2+\gamma_2)\mu}{(\gamma_1\nu_2+\gamma_2)\mu+\gamma_1\nu_2(1-\mu)}$

Table A.5: $\mu^R(m_{L\varnothing}) \equiv \text{Prob}\{\omega = H | m = m_{L\varnothing}\}$

$\sigma(m_{L\varnothing}\|X_1, X_2, C)$				
(L,L,1)	(L,L,2)	(L,H,1)	(L,H,2)	$\mu^R(m_{L\varnothing})$
0	0	0	0	any value in $[0,1]$
at least one non-zero			0	0
0	0	0	1	1
1	0	0	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+\gamma_1(1-\nu_2)(1-\mu)}$
0	1	0	1	μ
0	0	1	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+\gamma_1\nu_2(1-\mu)}$
1	1	0	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+[\gamma_1(1-\nu_2)+\gamma_2(1-\nu_1)](1-\mu)}$
1	0	1	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+\gamma_1(1-\mu)}$
0	1	1	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+[\gamma_1\nu_2+\gamma_2(1-\nu_1)](1-\mu)}$
1	1	1	1	$\frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu+[\gamma_1+\gamma_2(1-\nu_1)](1-\mu)}$

Table A.6: $\mu^R(m_{\varnothing L}) \equiv \text{Prob}\{\omega = H | m = m_{\varnothing L}\}$

$\sigma(m_{\varnothing L}\|X_1, X_2, C)$				
(L,L,1)	(L,L,2)	(H,L,2)	(H,L,1)	$\mu^R(m_{\varnothing L})$
0	0	0	0	any value in $[0,1]$
at least one non-zero			0	0
0	0	0	1	1
1	0	0	1	μ
0	1	0	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2(1-\nu_1)(1-\mu)}$
0	0	1	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2\nu_1(1-\mu)}$
1	1	0	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+[\gamma_1(1-\nu_2)+\gamma_2(1-\nu_1)](1-\mu)}$
1	0	1	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2\nu_1(1-\mu)}$
0	1	1	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+\gamma_2(1-\mu)}$
1	1	1	1	$\frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu+[\gamma_1(1-\nu_2)+\gamma_2](1-\mu)}$

A.3.2 Equivalence with sequential equilibrium

This appendix section elaborates on the claim made in Remark 4 of Chapter 3; this claim is formally proven below in Lemma 7. Before this, a few definitions are required.

A system of beliefs $b^* \equiv \{\pi^R(\cdot), \pi^S(\cdot), \pi^R(\cdot|\cdot), \pi^S(\cdot|\cdot)\}$ consists of (1) each agent's prior belief $\pi^R(\cdot) = \pi^S(\cdot) = \pi(\cdot)$ over (ω, C, X_1, X_2) (where $\pi(\cdot)$ is given in (3.3)), (2) S's (degenerate) posterior belief $\pi^S(\cdot|\cdot)$ over (ω, C, X_1, X_2) formed using Bayes' rule after observing their type $\theta = (x_1, x_2, c)$, which is given in (3.4), and (3) R's posterior belief $\pi^R(\cdot|\cdot)$ over (ω, C, X_1, X_2) after observing S's message m.

Definition 3 (Consistent Systems of Beliefs (Kreps and Wilson, 1982)). A system of beliefs $b^* \equiv \left(\pi^R(\cdot), \pi^S(\cdot), \pi^R(\cdot|\cdot), \pi^S(\cdot|\cdot)\right)$ is *consistent* with a strategy profile (α^*, σ^*) if there exists a sequence of strategy profiles and systems of beliefs $\{(\phi_n, b_n)\}_{n=1}^{\infty}$ such that

1. (α_n, σ_n) is a *fully mixed* strategy profile $(n = 1, 2, ...)$.
2. Each belief in $b_n \equiv \left(\pi_n^R(\cdot), \pi_n^S(\cdot), \pi_n^R(\cdot|\cdot), \pi_n^S(\cdot|\cdot)\right)$ is calculated using (α_n, σ_n) and Bayes' rule. $(n = 1, 2, ...)$
3. $(\alpha_n, \sigma_n, b_n) \to (\alpha^*, \sigma^*, b^*)$ as $n \to \infty$.

Notice that, given the game considered in Chapter 3, the only non-trivial parts of this definition involve the Sender's disclosure strategy and the Receiver's posterior belief. This is because (1) each agent's prior and posterior belief is constant in the Receiver's action strategy, (2) agents have a common prior over (ω, C, X_1, X_2), and (3) the Sender's posterior belief over (ω, C, X_1, X_2) can always be formed using Bayes' rule, which results in the degenerate belief given in (3.4). Formally put: for any choice of α^* and $\{\alpha_n\}_{n=1}^{\infty}$, such that $\alpha_n \to \alpha^*$ as $n \to \infty$, $\pi_n^R(\cdot) = \pi^R(\cdot) = \pi(\cdot)$ and $\pi_n^S(\cdot) = \pi^S(\cdot) = \pi(\cdot)$ $\forall n \in \mathbb{N}$ $(\therefore \pi_n^j(\cdot) \to \pi^j(\cdot)$ as $n \to \infty$ $\forall j \in \{R, S\})$ and $\pi_n^S(\cdot|\cdot) \equiv \pi^S(\cdot|\cdot)$ $\forall n \in \mathbb{N}$ $(\therefore \pi_n^S(\cdot|\cdot) \to \pi^S(\cdot|\cdot)$ as $n \to \infty)$.

Therefore, we can focus on the Sender's strategies and the Receiver's posterior beliefs (i.e. $\{\sigma^n\}_{n=1}^{\infty}$, σ^*, $\{\pi_n^R(\cdot|\cdot)\}_{n=1}^{\infty}$, and $\pi^R(\cdot|\cdot)$) when applying Definition 3.

Recall that when it is possible for R to use Bayes' rule to form $\pi^R(\cdot|m)$ – which is always the case when S is using a fully mixed strategy – R's posterior belief was given in equation (3.10) as follows:

$$\pi^R(\omega, C, X_1, X_2|m) = \frac{\sigma(m|x_1, x_2, c)\pi(\omega, C, X_1, X_2)}{\sum_{\tilde{\omega}\in\{L,R\}}\sum_{\tilde{c}\in\{1,2\}}\sum_{\tilde{x_1}\in\{L,R\}}\sum_{\tilde{x_2}\in\{L,R\}}\sigma(m|\tilde{x_1}, \tilde{x_2}, \tilde{c})\pi(\tilde{\omega}, \tilde{c}, \tilde{x_1}, \tilde{x_2})}$$

Recall that the Receiver's belief that $\omega = H$ under posterior $\pi^R(\cdot|m)$ is given by

$$\mu^R(m) \equiv \sum_{\tilde{c}=1}^{2}\sum_{\tilde{x_1}\in\{L,R\}}\sum_{\tilde{x_2}\in\{L,R\}}\pi^R(\omega = H, C = \tilde{c}, X_1 = \tilde{x_1}, X_2 = \tilde{x_2}|m)$$

For readers' convenience, appendix section A.3.1 contains explicit formulæ for $\mu^R(m)$ when m is on the equilibrium path (i.e. $\pi^R(\cdot|\cdot)$ is formed using Bayes' rule) along with probability tables for $\mu^R(m)$ when the Sender uses a pure disclosure strategy.

Lemma 7.

1. *If a system of beliefs b is consistent with a strategy profile (α, σ) s.t. $\sigma(m_{LL}|\theta) = 0$ $\forall\theta \in \{L, H\}^2 \times \{1, 2\}$, then $\mu^R(m_{LL}) = 0$.*

2. *If a system of beliefs b is consistent with a strategy profile (α, σ) s.t. $\sigma(m_{HH}|\theta) = 0$ $\forall\theta \in \{L, H\}^2 \times \{1, 2\}$, then $\mu^R(m_{HH}) = 1$.*

3. *Let (α, σ) be a strategy profile with $\sigma(m|\theta) = 0$ $\forall\theta \in \{L, H\}^2 \times \{1, 2\}$ for some $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$. Then for any $\psi \in [0, 1]$, there exists a system of beliefs b_ψ that is consistent with (α, σ) that has $\mu^R(m) = \psi$.*

Proof. I will now show that for each $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$, if m is observed by R off the equilibrium path, then any $\mu^R(m) \in [0,1]$ can be achieved while complying with the definition of Consistency stated above. Henceforth let

$$\mu_n^R(m) \equiv \sum_{\tilde{c}=1}^{2} \sum_{\tilde{x_1} \in \{L,R\}} \sum_{\tilde{x_2} \in \{L,R\}} \pi_n^R(\omega = H, C = \tilde{c}, X_1 = \tilde{x}_1, X_2 = \tilde{x}_2 | m)$$

where $\pi_n^R(\cdot|\cdot)$ is as defined in Definition 3.

I first show this for message $m = m_{HL}$ (the logic for other $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$ is very similar). If the Sender uses a fully mixed strategy, then, as stated in equation (A.67),

$$\begin{aligned}\mu^R(m_{HL}) &= \frac{\gamma_1(1-\nu_2)\sigma(m_{HL}|H,L,1)\mu}{\gamma_1(1-\nu_2)\sigma(m_{HL}|H,L,1)\mu + \gamma_2\nu_1\sigma(m_{HL}|H,L,2)(1\text{-}\mu)} \\ &= \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1\frac{\sigma(m_{HL}|H,L,2)}{\sigma(m_{HL}|H,L,1)}(1\text{-}\mu)}\end{aligned}$$

Suppose that σ has $\sigma(m_{HL}|H,L,1) = 0 = \sigma(m_{HL}|H,L,2)$, so that m_{HL} is *off* the equilibrium path. First suppose that for each $n \in \mathbb{N}$,

$$\sigma_n(m_{HL}|H,L,1) = \frac{1}{n^2},\ \sigma_n(m_{HL}|H,L,2) = \frac{1}{n}.$$

It then follows that

$$\begin{aligned}\mu_n^R(m_{HL}) &= \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1-\mu)\frac{1/n}{1/(n^2)}} \\ &= \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1-\mu)n} \to 1 \text{ as } n \to \infty\end{aligned}$$

Now suppose instead that for each $n \in \mathbb{N}$,

$$\sigma_n(m_{HL}|H,L,1) = \frac{1}{n},\ \sigma_n(m_{HL}|H,L,2) = \frac{1}{n^2}.$$

In this case, it follows that

$$\mu_n^R(m_{HL}) = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1-\mu)\frac{1/(n^2)}{1/n}}$$
$$= \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1-\mu)\frac{1}{n}} \to 0 \text{ as } n \to \infty$$

Finally, suppose that for each $n \in \mathbb{N}$,

$$\sigma_n(m_{HL}|H, L, 1) = \frac{1}{n},\ \sigma_n(m_{HL}|H, L, 2) = \frac{\eta}{n}.$$

for some $\eta \neq 0$. Given this, it follows that

$$\mu_n^R(m_{HL}) = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1-\mu)\frac{\eta/n}{1/n}} = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1-\mu)\eta},$$

which is constant in n (hence, $\mu_n^R(m_{HL})$ trivially approaches the value after the second equality, above, as $n \to \infty$). Note that for any fixed $\psi \in (0, 1)$, setting

$$\eta = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_2\nu_1(1-\mu)} \cdot \frac{1-\psi}{\psi}$$

implies that $\mu_n^R(m_{HL}) = \psi\ \forall n \in \mathbb{N}$ so that $\mu_n^R(m_{HL}) \to \psi$ as $n \to \infty$. Therefore, it can be concluded that any $\mu^R(m_{HL}) \in [0, 1]$ can be achieved while complying with the definition of Consistency, mentioned above.

As mentioned above, the logic for other $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$ is very similar. Notice that formulæ in equations (A.67)-(A.71) all take a similar form. Namely, for any arbitrarily fixed $m \in \mathcal{M} \setminus \{m_{LL}, m_{HH}\}$, if m is observed on the equilibrium path, then $\mu^R(m)$ takes the form

$$\mu^R(m) = \frac{\sum_{j=1}^{k} \beta_j \sigma(m|\boldsymbol{\tau}_j)}{\sum_{j=1}^{k} \beta_j \sigma(m|\boldsymbol{\tau}_j) + \sum_{j=1}^{\tilde{k}} \tilde{\beta}_j \sigma(m|\tilde{\boldsymbol{\tau}}_j)}$$

for some $(k, \tilde{k}) \in \{1,2,3\}^2$, $(\beta_j)_{j=1}^k \in (0,1)^k$, $(\tilde{\beta}_j)_{j=1}^{\tilde{k}} \in (0,1)^{\tilde{k}}$, and collection of *distinct*[15]types $\{\boldsymbol{\tau}\}_{j=1}^k \sqcup \{\tilde{\boldsymbol{\tau}}_j\}_{j=1}^{\tilde{k}} \subset \{L,H\}^2 \times \{1,2\}$. Following the same logic as before, first suppose that for each $n \in \mathbb{N}$

$$\sigma_n(m|\boldsymbol{\tau}_j) = \frac{1}{n} \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^k, \;\; \sigma_n(m|\tilde{\boldsymbol{\tau}}_j) = \frac{1}{n^2} \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^{\tilde{k}}.$$

Then, we have that

$$\mu_n^R(m) = \frac{\sum_{j=1}^k \frac{1}{n}\beta_j}{\sum_{j=1}^k \frac{1}{n}\beta_j + \sum_{j=1}^{\tilde{k}} \frac{1}{n^2}\tilde{\beta}_j} = \frac{\sum_{j=1}^k \beta_j}{\sum_{j=1}^k \beta_j + \frac{1}{n}\sum_{j=1}^{\tilde{k}} \tilde{\beta}_j} \to 1 \text{ as } n \to \infty.$$

If we instead suppose that for each $n \in \mathbb{N}$

$$\sigma_n(m|\boldsymbol{\tau}_j) = \frac{1}{n^2} \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^k, \;\; \sigma_n(m|\tilde{\boldsymbol{\tau}}_j) = \frac{1}{n} \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^{\tilde{k}},$$

it then follows that

$$\mu_n^R(m) = \frac{\sum_{j=1}^k \frac{1}{n^2}\beta_j}{\sum_{j=1}^k \frac{1}{n^2}\beta_j + \sum_{j=1}^{\tilde{k}} \frac{1}{n}\tilde{\beta}_j} = \frac{\sum_{j=1}^k \beta_j}{\sum_{j=1}^k \beta_j + n\sum_{j=1}^{\tilde{k}} \tilde{\beta}_j} \to 0 \text{ as } n \to \infty.$$

Finally, if we arbitrarily fix some $\eta \neq 0$ and suppose that

$$\sigma_n(m|\boldsymbol{\tau}_j) = \frac{1}{n} \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^k, \;\; \sigma_n(m|\tilde{\boldsymbol{\tau}}_j) = \frac{\eta}{n} \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^{\tilde{k}},$$

it then follows that

$$\mu_n^R(m) = \frac{\sum_{j=1}^k \frac{1}{n}\beta_j}{\sum_{j=1}^k \frac{1}{n}\beta_j + \sum_{j=1}^{\tilde{k}} \frac{\eta}{n}\tilde{\beta}_j} = \frac{\sum_{j=1}^k \beta_j}{\sum_{j=1}^k \beta_j + \eta\sum_{j=1}^{\tilde{k}} \tilde{\beta}_j} \; \forall n \in \mathbb{N}.$$

[15]I.e. $\boldsymbol{\tau}_j \neq \boldsymbol{\tau}_\ell \; \forall j \neq \ell$, $\tilde{\boldsymbol{\tau}}_j \neq \tilde{\boldsymbol{\tau}}_\ell \; \forall j \neq \ell$, and $\boldsymbol{\tau}_j \neq \tilde{\boldsymbol{\tau}}_\ell \; \forall j \in \{\hat{i}\}_{\hat{i}=1}^k, \; \ell \in \{\hat{i}\}_{\hat{i}=1}^{\tilde{k}}$.

Notice that for any arbitrarily fixed $\psi \in (0,1)$, if one sets $\eta = \frac{1-\psi}{\psi} \cdot \frac{\sum_{j=1}^{k} \beta_j}{\sum_{j=1}^{\tilde{k}} \tilde{\beta}_j}$, it then follows that $\mu_n^R(m) = \psi \; \forall n \in \mathbb{N}$ so that $\mu_n^R(m)$ trivially approaches ψ as $n \to \infty$.

Finally, I show that when $m = m_{HH}$ or $m = m_{LL}$ is observed off the equilibrium path, Consistency requires that $\mu^R(m_{HH}) = 1$ and $\mu^R(m_{LL}) = 0$, respectively. Recall that by equation (3.5), m_{HH} can only be sent by types $\theta \in \{(H,H,1),(H,H,2)\}$ and m_{LL} can only be sent by types $\theta \in \{(L,L,1),(L,L,2)\}$. It then follows that

$$\mu^R(m_{HH}) = \frac{[\gamma_1 \cdot \sigma(m_{HH}|H,H,1) \cdot 1 \cdot \nu_2 + \gamma_2 \cdot \sigma(m_{HH}|H,H,2) \cdot 1 \cdot \nu_1]\,\mu}{[\gamma_1 \cdot \sigma(m_{HH}|H,H,1) \cdot 1 \cdot \nu_2 + \gamma_2 \cdot \sigma(m_{HH}|H,H,2) \cdot 1 \cdot \nu_1]\,\mu + (0)(1\text{-}\mu)} = 1$$

for any $(\sigma(m_{HH}|H,H,1), \sigma(m_{HH}|H,H,2)) \in (0,1)^2$. It also follows that

$$\mu^R(m_{LL}) = \frac{(0)\mu}{(0)\mu + [\gamma_1 \cdot \sigma(m_{LL}|L,L,1) \cdot 1 \cdot (1-\nu_2) + \gamma_2 \cdot \sigma(m_{LL}|L,L,2) \cdot 1 \cdot (1-\nu_1)]\,(1\text{-}\mu)} = 0$$

for any $(\sigma(m_{LL}|L,L,1), \sigma(m_{LL}|L,L,2)) \in (0,1)^2$. It then obviously follows that for any sequence $\{\sigma_n\}_{n\in\mathbb{N}}$ of fully mixed disclosure strategies of the Sender, $\mu_n^R(m_{HH}) = 1$ and $\mu_n^R(m_{LL}) = 0 \; \forall n$, which respectively approach 1 and 0 as $n \to \infty$. ■

A.3.3 Proof of Proposition 8

Suppose that type (x_1, x_2, c) sends message m_{x_1,x_2} with probability 1 $\forall x_1, x_2$. It then follows that $\mu(m_{LL}) = 0$. By construction $m_{L\varnothing}$ and $m_{\varnothing L}$ are both off the equilibrium path. Any choice of $\mu^R(m) \in [0,1]$ satisfies condition (iv) in Definition 2 $\forall m \in \{m_{L\varnothing}, m_{\varnothing L}\}$. To illustrate, consider off-path message $m_{L\varnothing}$ condition (iv) only requires that $\pi(\omega, C, X_1 = H, X_2 | m_{L\varnothing}) = 0$. Hence, a $\pi(\cdot|m_{L\varnothing})$ with a marginal distribution over C that places probability 1 on the event "$C = 1$" is permitted by condition (iv). This would imply that after observing $m = m_{L\varnothing}$ off the equilibrium path $\mu(m_{L\varnothing}) = 0$; otherwise, the assumption that the conditional independence structure of (ω, C, X_1, X_2) is common

knowledge would be violated. Condition (iv) also permits $\pi(\cdot|m_{L\varnothing})$ to have a marginal distribution over C that places probability 1 on the event "$C = 1$," and a marginal distribution over X_2 that places any probability $\beta \in [0, 1]$ on the event "$X_2 = H$." Since the conditional independence structure of (ω, C, X_1, X_2) is common knowledge, such a $\pi(\cdot|m_{L\varnothing})$ would have $\mu^R(m_{L\varnothing}) = \beta$.

Hence, one may suppose that $\mu^R(m_{L\varnothing})$=0=$\mu^R(m_{\varnothing L})$. Since $\mathcal{M}_{LL} = \{m_{LL}, m_{L\varnothing}, m_{\varnothing L}\}$, it then clearly follows that types $(L, L, 1)$ and $(L, L, 2)$ each do not have a profitable deviation. Each type $\theta \in \{(H, H, 1), (H, H, 2)\}$ sends m_{HH} with probability 1, so that $\mu^R(m_{HH}) = 1$. Therefore, each such type does not have a profitable deviation. Each type $(L, H, 1)$ and $(L, H, 2)$ sends m_{LH} with probability 1, so

$$\mu^R(m_{LH}) = \frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu + \gamma_1\nu_2(1\text{-}\mu)} \in (0, 1)$$

Moreover, since each type $\theta \in \{(H, L, 1), (H, L, 2)\}$ sends m_{LH} with probability 1, so

$$\mu^R(m_{HL}) = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1\text{-}\mu)} \in (0, 1)$$

The proof that condition (iv) of Definition 2 places no restriction on $\mu^R(m) \in [0, 1]$ $\forall m \in \{m_{\varnothing H}, m_{H\varnothing}\}$ is identical to the proof found in the first paragraph of this subsection. Therefore, arbitrarily fixing any

$$\mu^R(m_{\varnothing H}) \leq \frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu + \gamma_1\nu_2(1\text{-}\mu)}$$

$$\mu^R(m_{H\varnothing}) \leq \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + \gamma_2\nu_1(1\text{-}\mu)}$$

is in compliance with condition (iv). Given this, it follows that no type in

$$\{(H, L, 1), (H, L, 2), (L, H, 1), (L, H, 2)\}$$

has a profitable deviation. ■

A.3.4 Proof of Proposition 9

Before proceeding with the proof, it is useful to recall equation (3.14), which implies that in any equilibrium, the message $m^*_{x_1x_2c} \in \mathcal{M}_{x_1x_2}$ sent by type (x_1, x_2, c) of Sender with probability 1 must be an element of $\arg\max_{m\in\mathcal{M}_{x_1x_2}} \{\mu^R(m)\}\ \forall(x_1, x_2, c)$. In this proof, let $s_{x_1x_2c}$ denote the message sent by type (x_1, x_2, c) of Sender with probability 1 in a given candidate equilibrium. Finally, probability tables for $\mu^R(\cdot)$ are provided in Tables A.1-A.6 in Appendix section A.3.1 for readers' convenience.

I first show that there cannot exist any equilibria where type $(L, H, 1)$ sends message $m \in \mathcal{M}_{LH}$ with probability 1 while type $(L, H, 2)$ sends a different message $m' \in \mathcal{M}_{LH} \setminus \{m\}$ with probability 1. Recall that the set of messages available to each of this types is given by

$$\mathcal{M}_{LH} \equiv \{m_{L\varnothing}, m_{\varnothing H} m_{LH}\}.$$

First suppose that $(s_{LH1}, s_{LH2}) = (m_{L\varnothing}, m_{\varnothing H})$. Then $\mu^R\left(m_{L\varnothing}\right) = 0$ and $\mu^R\left(m_{\varnothing H}\right) = 1$ (see the second row of Table A.5 and Table A.4, respectively). Therefore type $(L, H, 1)$ has a profitable deviation to message $m_{\varnothing H}$. If $(s_{LH1}, s_{LH2}) = (m_{L\varnothing}, m_{LH})$, then $\mu^R\left(m_{L\varnothing}\right) = 0$ (see the second row of Table A.5) and $\mu^R\left(m_{LH}\right) = 1$ (see the third row of Table A.2). Therefore type $(L, H, 1)$ has a profitable deviation to message m_{LH}.

Now suppose that $(s_{LH1}, s_{LH2}) = (m_{\varnothing H}, m_{L\varnothing})$. It is expositionally most clear to proceed by breaking this case down into two sub-cases: first, suppose that $\exists r' \in \{1, 2\}$ such that $s_{HHr} = m_{\varnothing H}$. Given Table A.4 it then follows from the assumption that $\mu, \gamma_1, \nu_1, \nu_2 \in (0, 1)$ that $\mu^R(m_{\varnothing H}) < 1$. Recalling that condition (iv) of Definition 2 implies that $\mu^R(m_{HH})$ on *and off* the equilibrium path, it follows that type (H, H, r') of Sender has a profitable deviation to message m_{HH}. If we then suppose that $s_{HHr'} \neq m_{\varnothing H}\ \forall r' \in \{1, 2\}$. Recall that

in the case under consideration, $s_{LH2} \neq m_{L\varnothing}$. It then follows from Table A.4 that $\mu^R(m_{EH}) = 0$. Observing Table A.5, it follows that $\mu^R(m_{L\varnothing}) > 0$ since $\mu, \gamma_1, \nu_1, \nu_2 \in (0,1)$. Therefore type $(L, H, 1)$ of Sender has a profitable deviation to message $m_{L\varnothing}$.

If $(s_{LH1}, s_{LH2}) = (m_{\varnothing H}, m_{LH})$, then $\mu^R\left(m_{\varnothing H}\right) < 1$ and $\mu^R\left(m_{LH}\right) = 1$. The latter equality follows from row 3 of Table A.2. The former inequality mechanically follows from the fact that $s_{LH1} = m_{\varnothing H} \neq m_{LH}$ and the assumption that $\mu, \gamma_1, \nu_1, \nu_2 \in (0,1)$. Intuitively, $\mu^R\left(m_{\varnothing H}\right) < 1$ because message $m_{\varnothing H}$ is sent by type $(L, H, 1)$ with strictly positive probability and source 1 has a strictly positive probability of being relevant (i.e. $\gamma_1 > 0$). Since $\omega = L$ with probability 1 if the relevant source's signal is drawn as L, it then follows that $\pi^R_\omega(\cdot|m_{\varnothing H})$ gives strictly positive weight to the event that ω is L. Since $\mu^R\left(m_{\varnothing H}\right) < 1 = \mu^R\left(m_{LH}\right)$, it then follows that type $(L, H, 1)$ has a profitable deviation to message m_{LH}.

Now suppose that $(s_{LH1}, s_{LH2}) = (m_{LH}, m_{L\varnothing})$. Then by the second row of Table A.2, $\mu^R\left(m_{LH}\right) = 0$. In this case, $\mu^R(m_{L\varnothing}) > 0$. Mechanically, this follows from the assumption that $\mu, \gamma_1, \nu_1, \nu_2 \in (0,1)$, as is made apparent in row 3 onward of Table A.5. Intuitively, it follows from the fact that $m_{L\varnothing}$ is sent with strictly positive probability by type $(L, H, 2)$, whose signal from the relevant source ($c = 2$) is $x_2 = H$. Since the second source is relevant with strictly positive probability (i.e. $\gamma_2 = 1 - \gamma_1 > 0$), and the state ω is H with probability if the relevant source signal is drawn as H, it follows that $\pi^R_\omega(\cdot|m_{L\varnothing})$ must place strictly positive weight on the event that $\omega = H$. Since $\mu^R(m_{LH}) = 0 < \mu^R(m_{L\varnothing})$, it follows that type $(L, H, 1)$ has a profitable deviation to message $m_{L\varnothing}$.

Finally, let us suppose that $(s_{LH1}, s_{LH2}) = (m_{LH}, m_{\varnothing H})$. It then follows that $\mu^R\left(m_{LH}\right) = 0$ and $\mu^R\left(m_{\varnothing H}\right) = 1$. The former equality holds for the same reason as before: since m_{LH} is sent only by type $(L, H, 1)$, the Receiver knows that the first source is relevant, and its signal was realized as $X_1 = L$. The latter equality follows from the fact that $m_{\varnothing H}$ is not sent by type $(L, H, 1)$, and the remaining types that can send $m_{\varnothing H}$ are $(L, H, 2)$, $(H, H, 1)$,

$(H, H, 2)$. Therefore, when the Receiver observes $m_{\varnothing H}$, they believe the Sender holds an H draw of a relevant source with probability 1, and hence can infer that the state is H with probability 1. Since $\mu^R(m_{LH}) = 0 < 1 = \mu^R(m_{\varnothing H})$, type $(L, H, 1)$ has a profitable deviation to message $m_{\varnothing H}$.

Since all distinct pairs of messages in $\mathcal{M}_{LH}$ have now been exhausted, we can now conclude that there do not exist equilibria wherein types $(L, H, 1)$ and $(L, H, 2)$ send different messages, each with probability 1.

The proof that no equilibria exist where $(H, L, 1)$ and $(H, L, 2)$ send different messages, each with probability 1 is nearly identical to the proof presented above. For completeness, it is presented below. Recall that the set of messages available to these types of Senders is given by

$$\mathcal{M}_{HL} \equiv \{m_{H\varnothing}, m_{\varnothing L}, m_{HL}\}.$$

If $(s_{HL1}, s_{HL2}) = (m_{H\varnothing}, m_{\varnothing L})$, then $\mu^R(m_{H\varnothing}) = 1$ and $\mu^R(m_{\varnothing L}) = 0$. Therefore type $(H, L, 2)$ has a profitable deviation to message $m_{H\varnothing}$.

If $(s_{HL1}, s_{HL2}) = (m_{H\varnothing}, m_{HL})$, then $\mu^R(m_{H\varnothing}) = 1$ and $\mu^R(m_{HL}) = 0$. Therefore type $(H, L, 2)$ has a profitable deviation to message $m_{H\varnothing}$.

Now suppose that $(s_{HL1}, s_{HL2}) = (m_{\varnothing L}, m_{H\varnothing})$. Then, $\mu^R(m_{\varnothing L}) > 0$, which follows from Table A.6 and the assumption that γ_1, ν_1, ν_2, μ are each elements of $(0, 1)$. As in the third case of the previous part of this proof, it is convenient to break this case into two sub-cases. First suppose that $s_{HHr} \neq m_{H\varnothing}\ \forall r \in \{1, 2\}$. It then follows from Table A.3 that $\mu^R(m_{H\varnothing}) = 0$. Hence, type $(H, L, 2)$ has a profitable deviation to $m_{\varnothing L}$. If we instead suppose that $s_{HHr} = m_{H\varnothing}$ for some $r \in \{1, 2\}$, then $\mu^R(m_{H\varnothing}) \in (0, 1)$ by Table A.3 and the assumption that $(\gamma_1, \nu_1, \nu_2, \mu) \in (0, 1)^4$. Since condition (iv) of Definition 2 requires that $\mu^R(m_{HH}) = 1$ on and off the equilibrium path, it then follows that type (H, H, r) has a

profitable deviation to m_{HH}.

If $(s_{HL1}, s_{HL2}) = (m_{\varnothing L}, m_{HL})$, then $\mu^R(m_{HL}) > 1$ and $\mu^R(m_{HL}) = 0$. The former inequality follows from Table A.6 (row three onward) and the assumption that $(\gamma_1, \nu_1, \nu_2, \mu) \in (0,1)^4$; the latter equality follows from row three of Table A.1. Given this, it follows that type $(H, L, 2)$ has a profitable deviation to message $m_{\varnothing L}$.

Now suppose that $(s_{HL1}, s_{HL2}) = (m_{HL}, m_{H\varnothing})$. Then $\mu^R(m_{HL}) = 1$ (by row two of Table A.1) and $\mu^R(m_{H\varnothing}) < 1$ (which follows from row three onward of Table A.3 because of the assumption that $(\gamma_1, \nu_1, \nu_2, \mu) \in (0,1)^4$). It then follows that $(H, L, 2)$ has a profitable deviation to m_{HL}.

If $(s_{HL1}, s_{HL2}) = (m_{HL}, m_{\varnothing L})$, then $\mu^R(m_{HL}) = 1$ and $\mu^R(m_{\varnothing L}) < 1$, for the same reasons as mentioned previously. Therefore type $(H, L, 2)$ has a profitable deviation to message m_{HL}.

Since all distinct pairs of messages in $\mathcal{M}_{HL}$ have now been exhausted, we can now conclude that there do not exist equilibria wherein types $(H, L, 1)$ and $(H, L, 2)$ send different messages, each with probability 1. ■

A.3.5 Proof of Proposition 10

As in the Proof of Proposition 9 (found in appendix section A.3.4), I let $s_{x_1 x_2 c}$ denote the message sent by type (x_1, x_2, c) of Sender with probability 1 in a given candidate equilibrium. By construction, focus is placed on candidate equilibria where

$$s_{HL1} = s_{HL2} = m_{\varnothing L};\ s_{LH1} = s_{LH2} = m_{L\varnothing}. \tag{A.72}$$

As is evident from Table A.6, $\mu^R(m_{\varnothing L}) = 1$ is possible only if $m_{\varnothing L}$ is not sent on the equilibrium path or is sent only by type $(H, L, 1)$ due to the assumption that $\mu, \gamma_1, \nu_1, \nu_2 \in (0,1)$. Therefore, it follows that $\mu^R(m_{\varnothing L}) < 1$. Similarly, it is evident from

Table A.5 that $\mu^R(m_{L\varnothing})$ can be equal to one only if $m_{L\varnothing}$ is not sent on the equilibrium path or sent only by type $(L, H, 2)$. Therefore, we can also conclude that $\mu^R(m_{L\varnothing}) < 1$. Suppose that $\exists r \in \{1, 2\}$ such that type (H, H, r) sends message $s_{HHr} \in \{m_{H\varnothing}, m_{\varnothing H}\}$ with probability 1. If this were the case, then $\mu^R(m) = 1$. It would then follow that there exists a type in $\{(L, H, r'), (H, L, r')\}_{r'=1}^{2}$ that has a profitable deviation to message m. Hence any equilibrium wherein (A.72) holds must have $s_{HHr} = m_{HH}$ $\forall r \in \{1, 2\}$. What remains to be considered are the messages sent by types $(L, L, 1)$ and $(L, L, 2)$. I consider these cases in the order that parts (a)-(d) are presented in the statement of Proposition 10. Note that there cannot exist an equilibrium where $s_{LLr} = m_{LL}$ for some $r \in \{1, 2\}$. This is because $\mu^R(m_{LL}) = 0$ on and off the equilibrium path but $\mu^R(m) > 0$ $\forall m \in \{m_{L\varnothing}, m_{\varnothing L}\}$ because of (A.72) and the assumption that $\mu, \gamma_1, \nu_1, \nu_2 \in (0, 1)$. Hence if $s_{LLr} = m_{LL}$ for some $r \in \{1, 2\}$, type (L, L, r) would have a profitable deviation to some message $m \in \{m_{L\varnothing}, m_{\varnothing L}\}$.

Part (a) First suppose that $s_{LL1} = m_{L\varnothing}$ and $s_{LL2} = m_{L\varnothing}$. It then follows from the last row of Table A.5 that

$$\mu^R(m_{L\varnothing}) = \frac{\gamma_2(1\text{-}\nu_1)\mu}{\gamma_2(1\text{-}\nu_1)\mu + [\gamma_1 + \gamma_2(1\text{-}\nu_1)](1\text{-}\mu)} = \frac{1}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1 + \gamma_2(1\text{-}\nu_1)}{\gamma_2(1\text{-}\nu_1)}\right)}. \tag{A.73}$$

It follows from row 6 of Table A.6 that

$$\mu^R(m_{\varnothing L}) = \frac{\gamma_1(1\text{-}\nu_2)\mu}{\gamma_1(1\text{-}\nu_2)\mu + \gamma_2\nu_1(1\text{-}\mu)} = \frac{1}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2\nu_1}{\gamma_1(1\text{-}\nu_2)}\right)}. \tag{A.74}$$

In order for neither type $(L, L, 1)$ nor $(L, L, 2)$ to have a profitable deviation, the following must hold

$$\frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1+\gamma_2(1-\nu_1)}{\gamma_2(1-\nu_1)}\right)} \geq \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2\nu_1}{\gamma_1(1-\nu_2)}\right)} \Leftrightarrow \frac{\gamma_1+\gamma_2(1-\nu_1)}{\gamma_2(1-\nu_1)} \leq \frac{\gamma_2\nu_1}{\gamma_1(1-\nu_2)}. \tag{A.75}$$

In order for neither type $(L, H, 1)$ nor $(L, H, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{\varnothing H}, m_{LH}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{L\varnothing}) \; \forall m \in \{m_{\varnothing H}, m_{LH}\},$$

for the value of $\mu^R(m_{L\varnothing})$ given in (A.73). In order for neither type $(H, L, 1)$ nor $(H, L, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{H\varnothing}, m_{HL}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{\varnothing L}) \; \forall m \in \{m_{H\varnothing}, m_{HL}\},$$

for the value of $\mu^R(m_{\varnothing L})$ given in (A.74). Given the above, it clearly follows that an equilibrium that satisfies

$$s_{LL1} = m_{L\varnothing}, s_{LL2} = m_{L\varnothing};\; s_{LHr} = m_{L\varnothing} \; \forall r \in \{1,2\};\;\; s_{HLr} = m_{\varnothing L} \; \forall r \in \{1,2\};\; s_{HHr} = m_{HH} \; \forall r \in \{1,2\}$$

exists if and only if (A.75) holds.

Part (b) First suppose that $s_{LL1} = m_{\varnothing L}$ and $s_{LL2} = m_{\varnothing L}$. It then follows from row 6 of Table A.5 that

$$\mu^R(m_{L\varnothing}) = \frac{\gamma_2(1-\nu_1)\mu}{\gamma_2(1-\nu_1)\mu + \gamma_1\nu_2(1-\mu)} = \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1\nu_2}{\gamma_2(1-\nu_1)}\right)}. \tag{A.76}$$

It follows from the last row of Table A.6 that

$$\mu^R(m_{\varnothing L}) = \frac{\gamma_1(1-\nu_2)\mu}{\gamma_1(1-\nu_2)\mu + [\gamma_1(1-\nu_2)+\gamma_2](1-\mu)} = \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1(1-\nu_2)+\gamma_2}{\gamma_1(1-\nu_2)}\right)}. \tag{A.77}$$

In order for neither type $(L, L, 1)$ nor $(L, L, 2)$ to have a profitable deviation, the following must hold

$$\frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1\nu_2}{\gamma_2(1-\nu_1)}\right)} \leq \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1(1-\nu_2)+\gamma_2}{\gamma_1(1-\nu_2)}\right)} \Leftrightarrow \frac{\gamma_1\nu_2}{\gamma_2(1-\nu_1)} \geq \frac{\gamma_1(1-\nu_2)+\gamma_2}{\gamma_1(1-\nu_2)}. \tag{A.78}$$

In order for neither type $(L, H, 1)$ nor $(L, H, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{\varnothing H}, m_{LH}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{L\varnothing}) \; \forall m \in \{m_{\varnothing H}, m_{LH}\},$$

for the value of $\mu^R(m_{L\varnothing})$ given in (A.76). In order for neither type $(H, L, 1)$ nor $(H, L, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{H\varnothing}, m_{HL}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{\varnothing L}) \; \forall m \in \{m_{H\varnothing}, m_{HL}\},$$

for the value of $\mu^R(m_{\varnothing L})$ given in (A.77). Given the above, it clearly follows that an equilibrium that satisfies

$$s_{LL1} = m_{\varnothing L}, s_{LL2} = m_{\varnothing L};\; s_{LHr} = m_{L\varnothing} \; \forall r \in \{1,2\};\;\; s_{HLr} = m_{\varnothing L} \; \forall r \in \{1,2\};\; s_{HHr} = m_{HH} \; \forall r \in \{1,2\}$$

exists if and only if (A.78) holds.

Part (c) First suppose that $s_{LL1} = m_{\varnothing L}$ and $s_{LL2} = m_{L\varnothing}$. It then follows from penultimate row of Table A.5 that

$$\mu^R(m_{L\varnothing}) = \frac{\gamma_2(1\text{-}\nu_1)\mu}{\gamma_2(1\text{-}\nu_1)\mu + [\gamma_1\nu_2+\gamma_2(1\text{-}\nu_1)](1\text{-}\mu)} = \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1\nu_2+\gamma_2(1\text{-}\nu_1)}{\gamma_2(1\text{-}\nu_1)}\right)}. \tag{A.79}$$

It follows from third-to-last row of Table A.6 that

$$\mu^R(m_{\varnothing L}) = \frac{\gamma_1(1\text{-}\nu_2)\mu}{\gamma_1(1\text{-}\nu_2)\mu + \gamma_2\nu_1(1\text{-}\mu)} = \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2\nu_1}{\gamma_1(1\text{-}\nu_2)}\right)}. \tag{A.80}$$

In order for neither type $(L, L, 1)$ nor $(L, L, 2)$ to have a profitable deviation, the following must hold

$$\frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1\nu_2+\gamma_2(1\text{-}\nu_1)}{\gamma_2(1\text{-}\nu_1)}\right)} = \frac{1}{1+\left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2\nu_1}{\gamma_1(1\text{-}\nu_2)}\right)} \Leftrightarrow \frac{\gamma_1\nu_2+\gamma_2(1\text{-}\nu_1)}{\gamma_2(1\text{-}\nu_1)} = \frac{\gamma_2\nu_1}{\gamma_1(1\text{-}\nu_2)}. \tag{A.81}$$

In order for neither type $(L, H, 1)$ nor $(L, H, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{\varnothing H}, m_{LH}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{L\varnothing}) \; \forall m \in \{m_{\varnothing H}, m_{LH}\},$$

for the value of $\mu^R(m_{L\varnothing})$ given in (A.79). In order for neither type $(H, L, 1)$ nor $(H, L, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{H\varnothing}, m_{HL}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{\varnothing L}) \; \forall m \in \{m_{H\varnothing}, m_{HL}\},$$

for the value of $\mu^R(m_{\varnothing L})$ given in (A.80). Given the above, it clearly follows that an equilibrium that satisfies

$$s_{LL1} = m_{\varnothing L}, s_{LL2} = m_{L\varnothing};\ s_{LHr} = m_{L\varnothing}\ \forall r \in \{1,2\};\ \ s_{HLr} = m_{\varnothing L}\ \forall r \in \{1,2\};\ s_{HHr} = m_{HH}\ \forall r \in \{1,2\}$$

exists if and only if (A.81) holds. That is, any such equilibrium exists only in a knife-edge case.

Part (d) First suppose that $s_{LL1} = m_{L\varnothing}$ and $s_{LL2} = m_{\varnothing L}$. It then follows from the third-to-last row of Table A.5 that

$$\mu^R(m_{L\varnothing}) = \frac{\gamma_2(1\text{-}\nu_1)\mu}{\gamma_2(1\text{-}\nu_1)\mu + \gamma_1(1\text{-}\mu)} = \frac{1}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1}{\gamma_2(1\text{-}\nu_1)}\right)}. \tag{A.82}$$

It follows from the penultimate row of Table A.6 that

$$\mu^R(m_{\varnothing L}) = \frac{\gamma_1(1\text{-}\nu_2)\mu}{\gamma_1(1\text{-}\nu_2)\mu + \gamma_2(1\text{-}\mu)} = \frac{1}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2}{\gamma_1(1\text{-}\nu_2)}\right)}. \tag{A.83}$$

In order for neither type $(L, L, 1)$ nor $(L, L, 2)$ to have a profitable deviation, the following must hold

$$\frac{1}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_1}{\gamma_2(1\text{-}\nu_1)}\right)} = \frac{1}{1 + \left(\frac{1-\mu}{\mu}\right)\left(\frac{\gamma_2}{\gamma_1(1\text{-}\nu_2)}\right)} \Leftrightarrow \frac{\gamma_1}{\gamma_2(1\text{-}\nu_1)} = \frac{\gamma_2}{\gamma_1(1\text{-}\nu_2)}. \tag{A.84}$$

In order for neither type $(L, H, 1)$ nor $(L, H, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{\varnothing H}, m_{LH}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{L\varnothing})\ \forall m \in \{m_{\varnothing H}, m_{LH}\},$$

for the value of $\mu^R(m_{L\emptyset})$ given in (A.82). In order for neither type $(H, L, 1)$ nor $(H, L, 2)$ to have a profitable deviation, the Receiver's off path beliefs formed after message $m \in \{m_{H\emptyset}, m_{HL}\}$ must satisfy

$$\mu^R(m) \leq \mu^R(m_{\emptyset L}) \; \forall m \in \{m_{H\emptyset}, m_{HL}\},$$

for the value of $\mu^R(m_{\emptyset L})$ given in (A.83). Given the above, it clearly follows that an equilibrium that satisfies

$$s_{LL1} = m_{L\emptyset}, s_{LL2} = m_{\emptyset L};\; s_{LHr} = m_{L\emptyset} \; \forall r \in \{1, 2\};\;\; s_{HLr} = m_{\emptyset L} \; \forall r \in \{1, 2\};\; s_{HHr} = m_{HH} \; \forall r \in \{1, 2\}$$

exists if and only if (A.84) holds. Just as in the previous part, any such equilibrium exists only in a knife-edge case. ■

Bibliography

Acemoglu, D. (2005). Politics and economics in weak and strong states. *Journal of monetary Economics 52*(7), 1199–1226.

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2016). Fragility of asymptotic agreement under bayesian learning. *Theoretical Economics 11*(1), 187–225.

Acemoglu, D. and J. A. Robinson (2006). De facto political power and institutional persistence. *American economic review 96*(2), 325–330.

Acemoglu, D. and J. A. Robinson (2019). *The narrow corridor: States, societies, and the fate of liberty*. Penguin Press.

Acemoglu, D. and J. A. Robinson (2022a, Aug). Why Taiwan Matters. *Project Syndicate*.

Acemoglu, D. and J. A. Robinson (2022b). Weak, despotic, or inclusive? how state type emerges from state versus civil society competition. *American Political Science Review*, 1–14.

Aghion, P. (2005). *Competition and growth : reconciling theory and evidence*. Zeuthen lecture book series. Cambridge, Mass: MIT Press.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005). Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics 120*(2), 701–728.

Aziz, H., Y. Bachrach, E. Elkind, and M. Paterson (2011). False-name manipulations in weighted voting games. *Journal of Artificial Intelligence Research 40*, 57–93.

Bachrach, Y. and E. Elkind (2008). Divide and conquer: False-name manipulations in weighted voting games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pp. 975–982. Citeseer.

Becker, G. S. (1983). A theory of competition among pressure groups for political influence. *The Quarterly Journal of Economics 98*(3), 371–400.

Becker, G. S. (1985). Public policies, pressure groups, and dead weight costs. *Journal of Public Economics 28*(3), 329–347.

Bénabou, R., A. Falk, and J. Tirole (2019). Narratives, imperatives, and moral persuasion. Working paper.

Berry, S. K. (1993). Rent-seeking with multiple winners. *Public Choice 77*(2), 437–443.

Bottomore, T. (1964). *Elites and Society*. Penguin Books.

Bowen, R., I. Hwang, and S. Krasa (2022). Personal power dynamics in bargaining. *Journal of Economic Theory 205*, 105530.

Bowen, T. R. and Z. Zahran (2012). On dynamic compromise. *Games and Economic Behavior 76*(2), 391–419.

Brahma, S., L. Njilla, and S. Nan (2022). Optimal auction design with malicious sellers. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pp. 661–666.

British Broadcasting Corporation (2016, Mar). Four times when having an online poll was a bad idea. https://www.bbc.com/news/uk-35860830.

Bull, J. and J. Watson (2019). Statistical evidence and the problem of robust litigation. *RAND Journal of Economics 50*(4), 974–1003.

Callander, S., D. Foarta, and T. Sugaya (2021). Market competition and political influence: An integrated approach.

Cammack, P. (1990). A critical assessment of the new elite paradigm. *American Sociological Review 55*(3), 415–420.

Chiang, H.-D. and L. F. C. Alberto (2015). *Stability regions of nonlinear dynamical systems : theory, estimation, and applications*. Cambridge: Cambridge University Press.

Chorppath, A. K. and T. Alpcan (2011). Adversarial behavior in network mechanism design. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pp. 506–514.

Chowdhury, S. M. and O. Gürtler (2015). Sabotage in contests: a survey. *Public Choice 164*(1-2), 135–155.

Chowdhury, S. M. and S.-H. Kim (2014). A note on multi-winner contest mechanisms. *Economics Letters 125*(3), 357 – 359.

Citizens United v. Federal Election Commission (2010, Jan). 558 U.S. 310. Number No. 08–205.

Clark, D. J. and C. Riis (1996). A multi-winner nested rent-seeking contest. *Public Choice 87*(1/2), 177–184.

Collins, B. and B. Popken (2019, Jun). Trolls target online polls following first democratic presidential debate. *National Broadcasting Company News*. https://www.nbcnews.com/tech/tech-news/trolls-target-online-polls-following-first-democratic-presidential-debate-n1023406.

Conitzer, V. (2008). Anonymity-proof voting rules. In C. Papadimitriou and S. Zhang (Eds.), *Internet and Network Economics*, Berlin, Heidelberg, pp. 295–306. Springer Berlin Heidelberg.

Corchón, L. and M. Dahm (2010). Foundations for contest success functions. *Economic Theory 43*(1), 81–98.

Dahl, R. A. (1958). A critique of the ruling elite model. *The American Political Science Review 52*(2), 463–469.

Dahl, R. A. (1971). *Polyarchy : participation and opposition*. New Haven: Yale University Press.

Dalio, R. (2021). *Principles for Dealing with the Changing World Order: Why Nations Succeed Or Fail*. Avid Reader Press; Simon and Schuster.

Dechenaux, E., D. Kovenock, and R. M. Sheremeta (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics 18*(4), 609–669.

Desai, R. M. and A. Olofsgård (2011). The costs of political influence: Firm-level evidence from developing countries. *Quarterly Journal of Political Science 6*(2), 137–178.

Diefenbach, T. (2019). Why michels' 'iron law of oligarchy' is not an iron law – and how democratic organisations can stay 'oligarchy-free'. *Organization Studies 40*(4), 545–562.

Dixit, A. (2021, December). "somewhere in the middle you can survive": Review of the narrow corridor by daron acemoglu and james robinson. *Journal of Economic Literature 59*(4), 1361–75.

Drutman, L. (2015, 05). *The Business of America is Lobbying: How Corporations Became Politicized and Politics Became More Corporate*. Oxford University Press.

Eliaz, K. and R. Spiegler (2020, December). A model of competing narratives. *American Economic Review 110*(12), 3786–3816.

Elkind, E., P. Faliszewski, and A. Slinko (2011). Cloning in elections: Finding the possible winners. *Journal of Artificial Intelligence Research 42*, 529–573.

Ewerhart, C. (2021). A typology of military conflict based on the hirshleifer contest. *University of Zurich, Department of Economics, Working Paper* (400).

Fioravanti, F. and J. Massó (2022). False-name-proof and strategy-proof voting rules under separable preferences. *Available at SSRN 4175113*.

Francois, P. (2002). *Social Capital and Economic Development*. Routledge.

Freedom House (2022, February). Freedom in the world 2022: The global expansion of authoritarian rule. By Sarah Repucci and Amy Slipowitz.

Frenkel, S., K. Browning, and T. Lorenz (2020, Jun). Tiktok teens and k-pop stans say they sank trump rally. *New York Times*.

Gary-Bobo, R. J. and T. Jaaidane (2000). Polling mechanisms and the demand revelation problem. *Journal of Public Economics 76*(2), 203 – 238.

Gilens, M. and B. I. Page (2014). Testing theories of american politics: Elites, interest groups, and average citizens. *Perspectives on politics 12*(3), 564–581.

Glazer, J. and A. Rubinstein (2021). Story builders. *Journal of Economic Theory 193*, 105211.

Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics 24*(3), 461–483.

Harris, C. and J. Vickers (1987). Racing with uncertainty. *The Review of Economic Studies 54*(1), 1–21.

He, K. (2018). Mislearning from censored data: The gambler's fallacy in optimal-stopping problems. *arXiv preprint arXiv:1803.08170*.

Hendricks, K. and R. P. McAfee (2006). Feints. *Journal of Economics & Management Strategy 15*(2), 431–456.

Hillman, A. L. and J. G. Riley (1989). Politically contestable rents and transfers. *Economics & Politics 1*(1), 17–39.

Hirsch, A. V. and J. P. Kastellec (2022). A theory of policy sabotage. *Journal of Theoretical Politics 34*(2), 191–218.

Hirshleifer, J. (1989). Conflict and rent-seeking success functions: Ratio vs. difference models of relative success. *Public Choice 63*(2), 101–112.

Hirshleifer, J. (1991a). The paradox of power. *Economics & Politics 3*(3), 177–200.

Hirshleifer, J. (1991b). The technology of conflict as an economic activity. *The American Economic Review 81*(2), 130–134.

Huntington, S. (1996). *The Clash of Civilizations and the Remaking of World Order*. Penguin Group.

Huntington, S. P. (1991). Democracy's third wave. *Journal of democracy 2*(2), 12–34.

Hyde, S. D. (2020). Democracy's backsliding in the international environment. *Science 369*(6508), 1192–1196.

Invernizzi, G. M. (2020). Electoral competition and factional sabotage. *Unpublished Manuscript*.

Ishida, J. (2012). Dynamically sabotage-proof tournaments. *Journal of Labor Economics 30*(3), 627–655.

Jeon, J. S. and I. Hwang (2020). The emergence and persistence of oligarchy: A dynamic model of endogenous political power. *Available at SSRN 3177771*.

Jiang, Y., J. Kang, D. Niyato, X. Ge, Z. Xiong, C. Miao, Xuemin, and Shen (2022). Reliable distributed computing for metaverse: A hierarchical game-theoretic approach.

Jittorntrum, K. (1978). An implicit function theorem. *Journal of Optimization Theory and Applications 25*(4), 575–577.

Kennedy, C. (2020, Feb). Assessing the risks to online polls from bogus respondents.

Koçak, K. (2018). Sequential updating: A behavioral model of belief change. Working Paper.

Konrad, K. A. (2012). Dynamic contests and the discouragement effect. *Revue d'économie politique 122*(2), 233–256.

Kreps, D. M. and R. Wilson (1982). Sequential Equilibria. *50*(4), 863–894.

Krugman, P. (2014a, Apr). What the 1% Don't Want Us to Know. Interview by Bill Moyers.

Krugman, P. (2014b, March). Wealth over work. *The New York Times*.

Krugman, P. (2020, July). Why do the rich have so much power? *The New York Times*.

Kumagai, S. (1980). An implicit function theorem: Comment. *Journal of Optimization Theory and Applications 31*(2), 285–288.

La Salle, J. and S. Lefschetz (2012). *Stability by Liapunov's Direct Method with Applications by Joseph La Salle and Solomon Lefschetz*. Elsevier.

Lambert, N. and Y. Shoham (2008). Truthful surveys. In C. Papadimitriou and S. Zhang (Eds.), *Internet and Network Economics*, Berlin, Heidelberg, pp. 154–165. Springer Berlin Heidelberg.

Lanchester, F. W. (1916). *Aircraft in warfare: The dawn of the fourth arm*. Constable limited.

Lang, M. (2020). Mechanism design with narratives. *CESifo Working Paper*.

LaSalle, J. (1960, December). Some extensions of liapunov's second method. *IRE Transactions on Circuit Theory 7*(4), 520–527.

Leach, D. K. (2005). The iron law of what again? conceptualizing oligarchy across organizational forms. *Sociological Theory 23*(3), 312–337.

Leach, D. K. (2015). Oligarchy, iron law of. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (Second Edition ed.)., pp. 201–206. Oxford: Elsevier.

Liang, A. and X. Mu (2020). Complementary information and learning traps. *The Quarterly Journal of Economics 135*(1), 389–448.

Liang, A., X. Mu, and V. Syrgkanis (2022). Dynamically aggregating diverse information. *Econometrica*.

Liu, C., S. Wang, L. Ma, X. Cheng, R. Bie, and J. Yu (2017). Mechanism design games for thwarting malicious behavior in crowdsourcing applications. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1–9.

Lyapunov, A. M. (1892). The general problem of the stability of motion (in russian). *Mathematical Society of Kharkov II*(7), 1–250.

Lyapunov, A. M. (1992). The general problem of the stability of motion (english translation of original 1892 manuscript). *International journal of control 55*(3), 531–534.

Martini, G. (2022). Multidimensional disclosure. Working Paper.

Maskin, E. and J. Tirole (2001). Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory 100*(2), 191–219.

Maxwell, G. and P. Oliver (1993). *The Critical Mass in Collective Action: A Micro-social Theory*. Cambridge University Press.

Michels, R. (1915). *Political parties: A sociological study of the oligarchical tendencies of modern democracy*. Hearst's International Library Company.

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 380–391.

Mills, C. W. (1956). *The Power Elite*. Oxford University Press.

Mosca, G. (1939). *The ruling class = (Elementi di scienza politica)* (1st ed. ed.). New York ;: McGraw-Hill Book Company, Inc. translation by Hannah D. Kahn; edited and revised with an introduction by Arthur Livingston.

New Zealand Government (2015). Flag design gallery. https://web.archive.org/web/20150811023846/https://www.govt.nz/browse/engaging-with-government/the-nz-flag-your-chance-to-decide/gallery/?sort=random. Archive.org capture from August 11, 2015.

Ober, J. (2008). *Democracy and knowledge : innovation and learning in classical Athens*. Princeton: Princeton University Press.

OECD (2008). *Growing unequal?: Income distribution and poverty in OECD countries*.

OECD (2011). *Divided We Stand: Why Inequality Keeps Rising*. Paris, France.

OECD (2012, January). Income inequality and growth: The role of taxes and transfers. *OECD Economics Department Policy Notes No. 9*.

OECD (2015). *In It Together: Why Less Inequality Benefits All*.

OECD (2021). *Does Inequality Matter?*

Oniśko, A., M. J. Druzdzel, and H. Wasyluk (2001). Learning bayesian network parameters from small data sets: Application of noisy-or gates. *International Journal of Approximate Reasoning 27*(2), 165–182.

Page, B. I., L. M. Bartels, and J. Seawright (2013). Democracy and the policy preferences of wealthy americans. *Perspectives on Politics 11*(1), 51–73.

Pareto, V. (1935). *The mind and society*. London: [J. Cape]. translation by Aldous Huxley.

Pareto, V. (1991). *The Rise and Fall of Elites: An Application of Theoretical Sociology* (1 ed.). Somerset: Routledge. Hans L. Translation by Zetterberg. Original work published in 1901.

Pearl, J. (1985). Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pp. 15–17.

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *The American Political Science Review 94*(2), 251–267.

Piketty, T. (1995). Social mobility and redistributive politics. *The Quarterly journal of economics 110*(3), 551–584.

Piketty, T. (2013). *Capital in the Twenty-First Century*. Harvard University Press.

Piketty, T. (2015). Putting distribution back at the center of economics: Reflections on capital in the twenty-first century. *Journal of Economic Perspectives 29*(1), 67–88.

Piketty, T. (2018, 3). Rising inequality and the changing structure of political conflict. The Inaugural James M. and Cathleen D. Stone Lecture in Economic Inequality, John F. Kennedy Jr. Forum, Harvard Kennedy School. URL: https://iop.harvard.edu/forum/thomas-piketty. Accessed October 3, 2022.

Piketty, T. (2019). *Capital and Ideology*. Harvard University Press.

Qing, J. (2013). *A Confucian Constitutional Order: How China's Ancient Past Can Shape Its Political Future*. Princeton University Press.

Rastegari, B., A. Condon, and K. Leyton-Brown (2007). Revenue monotonicity in combinatorial auctions. *ACM SIGecom Exchanges 7*(1), 45–47.

Rausser, G. C., J. Swinnen, and P. Zusman (2011). *Political power and economic policy : theory, analysis, and empirical applications*. Cambridge ;: Cambridge University Press.

Repucci, S. (2020). A leaderless struggle for democracy.

Rustow, D. A. (1966). The study of elites: Who's who, when, and how. *World Politics 18*(4), 690–717.

Saez, E. and G. Zucman (2019, January). Alexandria Ocasio-Cortez's Tax Hike Idea Is Not About Soaking the Rich. *The New York Times*.

Schwartzstein, J. (2014). Selective attention and learning. *Journal of the European Economic Association 12*(6), 1423–1452.

Schwartzstein, J. and A. Sunderam (2021). Using models to persuade. *American Economic Review 111*(1), 276–323.

Shishkin, D. (2021). Evidence acquisition and voluntary disclosure. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 817–818.

Skaperdas, S. (1996, Jun). Contest success functions. *Economic Theory 7*(2), 283–290.

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy 128*(3), 907–947.

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290.

Spiegler, R. (2020). Behavioral implications of causal misperceptions. *Annual Review of Economics 12*, 81–106.

Spiegler, R. (2021). Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association 18*(2), 583–617.

Spiegler, R., K. Eliaz, and Y. Weiss (2021). Cheating with models. *American Economic Review: Insights*.

Spina, N., D. C. Shin, and D. Cha (2011). Confucianism and democracy: A review of the opposing conceptualizations. *Japanese Journal of Political Science 12*(1), 143–160.

Stiglitz, J. E. (2011, May). Of the 1%, by the 1%, for the 1%. *Vanity Fair*.

Stiglitz, J. E. (2016). 8. inequality and economic growth. *The Political Quarterly 86*(S1), 134–155.

The Economist (2020, May). No one knows how many people live in north macedonia. https://www.economist.com/europe/2020/05/14/no-one-knows-how-many-people-live-in-north-macedonia.

Titova, M. (2022). Persuasion with verifiable information. Working Paper.

Tullock, G. (1967). The welfare costs of tariffs, monopolies, and theft. *Economic Inquiry 5*(3), 224–232.

Tullock, G. (1980). Efficient rent-seeking. In J. M. Buchanan, R. D. Tollison, and G. Tullock (Eds.), *Toward a Theory of the Rent-Seeking Society*, Chapter 5, pp. 97–112. Texas A&M University Press.

United Nations Department of Economic and Social Affairs (2020). World social report 2020: Inequality in a rapidly changing world. Technical report, United Nations.

Weber, A. (1929). *Theory of the location of industries*. University of Chicago Press. (Translated by Carl Joachim Friedrich).

Weber, M. (1925). *Wirtschaft und gesellschaft*. JCB Mohr (P. Siebeck).

Winters, J. A. (2011). *Oligarchy*. Cambridge University Press.

Winters, J. A. and B. I. Page (2009). Oligarchy in the united states? *Perspectives on politics 7*(4), 731–751.

World Bank (2005). *World Development Report 2006: Equity and Development*. : World Bank; New York: Oxford University Press.

World Bank (2017). *World Development Report 2017: Governance and the Law*. : World Bank. doi:10.1596/978-1-4648-0950-7. License: Creative Commons Attribution CC BY 3.0 IGO.

Yang, G., S. He, and Z. Shi (2017). Leveraging crowdsourcing for efficient malicious users detection in large-scale social networks. *IEEE Internet of Things Journal 4*(2), 330–339.

Yokoo, M. (2003). Characterization of strategy/false-name proof combinatorial auction protocols: Price-oriented, rationing-free protocol. In *IJCAI*, pp. 733–742.

Yokoo, M. (2008). *False-Name-Proof Auction*, pp. 308–310. Boston, MA: Springer US.

Yokoo, M., T. Matsutani, and A. Iwasaki (2006). False-name-proof combinatorial auction protocol: Groves mechanism with submodular approximation. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pp. 1135–1142.

Yokoo, M., Y. Sakurai, and S. Matsubara (2001). Robust combinatorial auction protocol against false-name bids. *Artificial Intelligence 130*(2), 167–181.

Yokoo, M., Y. Sakurai, and S. Matsubara (2004). The effect of false-name bids in combinatorial auctions: new fraud in internet auctions. *Games and Economic Behavior 46*(1), 174–188.