Norms and the Evolution of Leaders Followership

BSE Working Paper 1381

January 2023 (Revised July 2024)

Antonio Cabrales, Esther Hauk

bse.eu/research

Norms and the evolution of leaders' followership*

Antonio Cabrales†, Esther Hauk‡

July 29, 2024

Abstract

In this paper, we model the interaction between leaders, their followers and crowd followers in a coordination game with local interaction. The steady states of a dynamic best-response process can feature a coexistence of Pareto-dominant and risk-dominant actions in the population. The existence of leaders and their followers, along with the local interaction, which leads to clustering, is crucial for the survival of the Pareto-dominant actions. The evolution of leader and crowd followership shows that leader followership can also be locally stable around Pareto-dominant leaders. The paper answers the questions of which leader should be removed and how to optimally place leaders in the network to enhance payoff-dominant play.

JEL Classification: C72, D85.

Keywords: Leadership, norms, local interaction, networks

*Hauk acknowledges financial support from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centers of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S) and from the Ministerio de Ciencia e Innovación through research project PID2021-126209OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF "A way of making Europe." Cabrales acknowledges support through the María de Maeztu Programme CEX2021-001181-M and PID2021-126209OB-I00 and through Comunidad de Madrid grant EPUC3M11 (V PRICIT). We are also grateful for comments from seminar audiences at NEAT (Padova/Venice), the Cambridge INET webinar series, CEPR ESSET and the Barcelona BSE forum. Declarations of interest: None.

†Department of Economics, Universidad Carlos III de Madrid; email: antonio.cabrales@uc3m. es

‡Instituto de Análisis Económico (IAE-CSIC) and Barcelona School of Economics, Campus UAB, 08193 Bellaterra (Barcelona); email: esther.hauk@iae.csic.es

1 Introduction

Many human activities are characterized by a coordination problem where multiple stable social situations can arise. Previous literature has emphasized social conventions as a common way of dealing with this coordination issue (Young 1993, 1998, Burke and Young 2011). How, however, do such conventions arise? In this paper, we examine the effect of leadership on the outcomes of coordination games. Humans as a species, similarly to many primates, tend to organize themselves into rather hierarchical groups.[1] Some individuals take an action mostly because they are following their leader. Others may take an action because they prefer to behave similarly to their peers.[2] Our main aim is precisely to analyze the interactions of these two ways of reaching a convention in an environment with two possible conventions that differ in the societal level of welfare that they yield.

An important aspect of our model is that we consider the impact of "local" leadership. Individuals interact mostly with those close to them, and their leaders are part of their communities. The local character of leadership turns out to be empirically important in identifying its effects in different contexts, such as technology adoption in developing countries (Yengoh et al. 2010, Dwivedi et al. 2022) and vaccine adoption (Dhallival et al. 2023, Vincenzo et al. 2023), but there is scant theoretical literature to provide a good framework for analysis.[3]

We model a game with N players located in a circle.[4] Each of them plays a coordination game with their k closest neighbors, choosing a single action, which is either payoff dominant or risk dominant. There are three types of players: leaders (L), leader followers (LF) and crowd followers (CF). Leaders always take the same action regardless of all the other players' choices. All

[1] This tendency has important implications for our psychology (Cummins 2005), social organization (Manner and Case 2016), and health (Gilbert 2001).

[2] Bicchieri and Chavez (2010) and Krupka and Weber (2013), for example, have measured the impact of others' expectations about the "correct action" on our own choices.

[3] Of course, global leaders who reach the whole population are also important. In an extension of our work here, we show that our qualitative results do not change when we consider leadership of global rather than local reach.

[4] In a later section, we study interactions in a lattice with 2 dimensions and derive implications for more complex networks.

the other players care about the material payoffs from the coordination games. Leader followers receive an additional payoff of α_L when following their closest leader, where α_L reflects the leader's charisma. Crowd followers (CF) care about choosing the same action as their neighbors. Therefore, they receive an extra payoff proportional to the fraction of their neighbors whose action their own action matches with weight α_C, a measure of this peer influence. Our L players have a fixed position. They do not try to "pander" to the opinions of LF players or anyone else. This modeling choice follows the spirit of citizen candidate models (Besley and Coate 1998) and models of information transmission by leaders in which high-quality incumbents do not need to shift their information to "pander" to crowds (Canes-Wrone, Herron, and Shotts 2001). LF players also have some innate preference for playing A or B (they obtain α_L if they choose their favorite action, and zero otherwise, in addition to the payoff from playing with their neighbors) and then locate close to an L player, who acts as a focal point for coordination of like-minded players. We model two types of agents, leader followers and crowd followers, motivated by the fact that sensitivity to leadership varies widely in the population (see, e.g., Smith et al. 2007). Nevertheless, given that the players are playing a coordination game, everyone can be affected by leaders.

The aim of the model is to capture coordination problems with two different norms competing for societal dominance: a good but weak norm (the Pareto-dominant but risk-dominated one) and a bad but strong one (the Pareto-dominated but risk-dominant one). Such coordination problems have numerous possible instantiations: the use of clean or polluting energy (Ang et al. 2020), the adoption of a farming technology (Müller et al. 2018), the choice of a software platform for developers (Fang et al. 2021), language adoption (Iriberri and Uriarte 2012), the spread of academic ideas (Sunstein 2000), and expression of opinions on controversial social topics (Buskens et al. 2008), among others. In many of these applications, the choices of leaders, their followers and crowd followers determine which option survives and how.

We analyze the game as a dynamic process in which agents adjust their actions over time. We first analyze the steady states of population choices when individuals start from an arbitrary action and best-respond to the population

choices in the previous period. Clearly, the stationary states are equilibria of the game. After play has reached a stationary steady state, we allow all agents except the leaders to switch not just their actions but also their types.

An important insight of this model is that "good but weak" (i.e., payoff-dominant but not risk-dominant) social norms need clustered groups of supporters and very charismatic leaders. These two features—clustering and charisma—are crucial for explaining some important social phenomena. Vaccine adoption and social isolation in an epidemic, for example, can be very beneficial but also quite expensive strategies. Importantly, for vaccines or isolation to yield benefits, a large fraction of a locally interacting population must adopt the strategy together. The alternative strategy is to wait for some kind of herd immunity to arise in the population. In other words, epidemic containment is a coordination game with local interaction. We know already that local leaders are important in such a context in both developed (Hallgren et al. 2021), and developing (Afolabi and Ilesanmi 2021) countries (see also Dhallival et al. 2023, Vincenzo et al. 2023). Our results on leader targeting and location can inform interventions in this context.

Another insight delivered by the model is the stark asymmetry in importance between the leader espousing *good* (payoff-dominant) norms and the leader espousing *bad* (risk-dominant) norms. The risk-dominant norm is guaranteed to survive in the population as long as one leader subscribes to the risk-dominant action even if nobody gains from following her. However, if a leader who chooses the payoff-dominant action is not sufficiently charismatic (has a low value of α_L), the payoff-dominant norm may disappear. This norm will take over in a cluster of the population only if α_L is sufficiently high and there are no risk-dominant leaders inside the cluster.

We explore different policies through which a social planner might improve welfare. One possibility is for the social planner to target "behavioral change" by removing one influencer/leader in a context where not all the leaders can be changed. This policy can work only if the payoff-dominant influencers are sufficiently charismatic. Since welfare is improved if the payoff-dominant norm spreads, clearly the target can only be a risk-dominant leader. It is usually best to remove a risk-dominant leader located between two payoff-dominant

leaders. Sometimes, however, it can be better to remove a risk-dominant leader who has only one payoff-dominant leader as a neighbor. This is true when the payoff-dominant leader gains a very large sphere of influence from this removal.

We also study how to optimally place a given number of leaders in the network to enhance payoff-dominant play. The optimal distribution is always to cluster leaders of the same type. While, for very charismatic leaders, clustering of risk-dominant leaders at a minimum distance limits their area of influence, for less charismatic leaders, payoff-dominant leaders should be clustered at a distance that optimally solves the tradeoff between a larger area of influence and the probability of inducing payoff-dominant play.

The remainder of the paper is organized as follows: The next section discusses the related literature. The model is laid out in Section 3. Section 4 derives the steady states with fixed types, and Section 5 discusses their stochastic stability. Section 6 analyzes the evolution of types according to the best response dynamic. Section 7 discusses the importance of leaders for the survival of risk- versus payoff-dominant norms. Section 8 suggests some policies to foster Pareto-dominant types. Section 9 analyzes what happens if the neighborhood is a 2-dimensional lattice. Section 10 concludes and suggests directions for future research.

2 Related literature

Several important strands of the literature connect to our work. Most obviously, Acemoglu and Jackson (2015, 2017) have explored the role of social norms and leadership in coordination games. Their work follows on the foundational contributions of Young (1993, 1998) and Binmore and Samuelson (1994).[5] Our contribution to this literature is twofold. On the one hand, we emphasize the local aspect of social norms enforcement and the possibility of multiple social norms arising in steady state through local clustering. On the other hand, we emphasize the importance of agents who follow leaders relative to that of those who simply follow crowds and the interaction of the two types.

[5]Later, expanded expositions can be found, e.g., in Burke and Young (2011) and Binmore (2010).

Methodologically, we borrow tools from models of learning and evolution with local interaction. Ellison (1993) extends the stochastic stability tools of Young (1993) and Kandori, Mailath, and Rob (1993) for obtaining uniqueness in the very long run for evolutionary games. He shows that, with local interaction, convergence times are significantly shorter. Eshel, Shaked and Samuelson (1998) show that when agents imitate the best-performing agents in their local surroundings, the population can converge to playing cooperatively in a prisoners' dilemma. Morris's (2000) paper is perhaps the most closely related in this group to ours. He characterizes the conditions under which a risk-dominant strategy can invade the whole population when individuals play a coordination game in a network. He shows that the survival of a risk-dominated strategy depends on the relative isolation of a specific group. The additional insight that we offer to this line of work is the importance of leaders and their followers in situations in which communities are not isolated.

Important extensions are provided by Alós-Ferrer and Weidenhozer (2008, 2016), whose model captures how the (local) interaction of agents is different from the information about past behavior. Agents imitate the behavior of the best-performing agent in the information neighborhood, which is larger than the interaction neighborhood. They find that the payoff-dominant equilibrium is the unique long-run equilibrium when the interactions are not too global under arbitrary network systems. Chen, Chow and Wu (2013) modify the imitation rule from Alós Ferrer and Weidenholzer (2008) so that players imitate the best-average strategy rather than the best player, and they find that the dynamics can converge to the risk-dominant equilibrium alone, or to some non-monomorphic absorbing states with payoff-dominant-strategy takers being the majority. As is the case for the literature in the previous paragraph, these papers do not take into account the importance of leadership in the long-run outcomes of the population.

Our paper is also related to the vast literature on social norms sustained through community enforcement. Elinor Ostrom proposed these as an explanation for collective solutions to social dilemmas (see, e.g., Ostrom 2000), and Cristina Bicchieri shows how they can be measured and the importance of empirical and normative expectations from contacts (Bicchieri 2005, 2016).

We add to this literature by underscoring the importance of leaders and their followers for the establishment and survival of norms.

A large literature examines coordination games in networks, starting with Jackson and Watts (2002) and Goyal and Vega Redondo (2005) and extending to Cui (2014), Khan (2014) and Bilancini and Boncinelli (2018). Ushchev and Zenou (2020) explicitly work with social norms, rather than coordination games, in a linear-in-means model for networks. We contribute to this literature by studying how leadership and social dynamics impact social norm adoption.

An important literature centered on leaders was initiated by Ballester, Calvó-Armengol and Zenou (2006). They study a game with a continuum of strategies. All agents have the same payoff structure, and there are synergies between their actions, which depend on their position in the network. Some individuals are more important because their centrality makes their action affect those of others in a stronger way. One of the authors' main results is to identify the agent whose removal would hurt collective output most strongly. A main difference between our approach and theirs is that leadership in their model is based purely on location and requires that the network be irregular. The kind of game is also very different in that it is mostly about the production rather than the adoption of a kind of social norm. There have been numerous extensions of their approach (for a thorough review, see Zenou 2016). In our context, it is interesting to note the work of Zhou and Chen (2015), who study a game similar to that in Ballester, Calvó-Armengol and Zenou (2006) but where players can move sequentially, and the authors study which player should move first to maximize output.

There is also a literature on targeting in networks. Some of it concentrates on how to design networks so that they are more resilient to external attacks (Vigier and Goyal 2014, Dziubiński, and Goyal 2017), with a good review in Dziubiński et al. (2016). Some of this literature concentrates on the optimal way to identify and potentially change the type of the agents with the most influence in the network (Galeotti and Goyal 2009, 2010, Golub, Galeotti and Goyal 2020). As with the pioneering work of Ballester, Calvó-Armengol and Zenou (2006), the leaders in this literature are influential mostly because of

their network position, and they would have no special power in a regular network.

An interesting paper from the perspective of the question we study here is Zimmerman and Eguiluz (2009). These authors study a model where individuals play a prisoners' dilemma in a network and choose their action by imitating the individual in their neighborhood with the highest payoff. The dynamics of such a game tend to converge to full defection or to states with a large amount of cooperation. The cooperative steady state depends heavily on "leaders" who cooperate and have a high payoff and many contacts. If leaders are removed by means of a random change to their state, there can be fluctuations between the two steady states.

Another literature in complex science deals with dynamics in the presence of "stubborn" agents, who can influence others through their actions but do not change their strategies. These papers often take as their starting point the "voter model," (Holley and Liggett 1975, Clifford and Sudbury 1973) in which agents start in a state and then change their state by randomly imitating neighbors. Yildiz et al. (2012) show that the presence of stubborn agents significantly changes the dynamics, which now do not have absorbing homogeneous states, and, similarly to us, the authors discuss the optimal placement of stubborn agents. Hunter and Zaman (2022) ask similar questions but start from the DeGroot (1974) model for reaching a consensus in a network.

There is also a literature in evolutionary biology (King, Johnson, Van Vugt 2009; Van Vugt, Hogan, Kaiser 2008) that considers leadership as an evolved means of solving coordination problems. However, these works do not consider the interaction of leader followers with crowd followers or take into account the local interaction aspect that we study.

A recent paper by Levine, Modica and Rustichini (2022) models leadership in societies with potential conflicts between groups as games between leaders. There are two classes of leaders: group leaders, who share their group's interest, and a common leader, who cares about both groups. Each leader makes a recommendation to her potential followers which strategy to play in a 2×2 game with the other group. Followers compare the proposed strategy of their group leader with the proposed strategy of the common leader and follow the

recommendation with the highest implied promised payoff. They will punish the leader they followed if the realized payoff falls short of the promised payoff. The paper shows that because of competition between the common leader and the group leader, the leaders can solve cooperation problems such as those arising in the prisoners' dilemma and the battle of the sexes and in coordination games, but only if the followers' capacity for punishment is sufficiently high. This approach is complementary to ours because it assumes that there is competition between leaders and that followers blindly adhere to their leader's recommendation. The leader followers in our paper choose the leader's preferred action only if it is in their best interest.

Our paper is also related to works modeling leadership and the dynamics of cultural norms (Hauk and Mueller 2015, Prummer and Siedlarek 2017, Verdier and Zenou 2018). In these papers, the leaders have an objective function that leads them to manipulate the diffusion of norms, and they care either only about the long-run steady state (Hauk and Mueller 2015, Prummer and Siedlarek 2017) or about both it and the cultural transmission path (Verdier and Zenou 2018). In Prummer and Siedlarek (2017), leaders are Benevolent—i.e., they care about the well-being of their followers. Hauk and Mueller (2015) examine cultural dynamics in contexts with leaders who try to maximize the number of people socialized to their trait (engage in proselytism) or with rent-seeking leaders who try to maximize the overall level of socialization effort exerted in their group and can manipulate cultural perceptions. The leaders in Verdier and Zenou (2018) manipulate via provision of a group-specific public good. These papers have active and optimizing leaders, while the leaders in our model are simply stubborn agents with no concern for their followers (either their number or their well-being). In addition, in these papers, everybody follows some leader, and there is no network structure.

3 Model

Society in our model consists of N players located in a circle and of three different types. Each player's type is leader, denoted by L, leader follower, denoted by LF, or crowd follower, denoted by CF. Each type plays a coordi-

nation game with her k (even number) nearest neighbors in k games using a single action $x \in \{A, B\}$, where k includes the neighbors on both sides of the player. In other words, each player plays with $k/2$ neighbors on her right and $k/2$ neighbors on her left. We denote by n_i the neighborhood of any player i. The baseline utility of player i from action $x \in \{A, B\}$ is given by

$$u_i(x) = \frac{1}{k} \sum_{j \in n_i} u(x, x_j), \tag{1}$$

namely, the average payoff from playing the same action with each player that belongs to her neighborhood.

Action A is payoff (Pareto) dominant and action B risk dominant in the coordination game, which has the following payoff matrix:

	A	B
A	d, d	e, f
B	f, e	b, b

where $d > f$, $b > e$, $d > b$, $d + e < b + f$.

The L player cares only about the strategy she supports, which is her dominant strategy.[6] LF and CF players receive the "baseline utility" from the coordination games (1) and an additional utility that depends on their type and the action of their neighbors. The LF player has utility $u_i(x) + \alpha_L I_L$ when taking action x. Here, I_L is an indicator function taking the value 1 if she uses the action of the leader closest to her and 0 otherwise, and $\alpha_L \geq 0$ reflects the charisma or influence of the leader L.

The CF player has a neighborhood of reference, comprising the k closest players on the left and right with whom she plays the coordination games. Her utility from action x is $u_i(x) + \alpha_C k_x/k$, where k_x is the number of her k closest neighbors taking action x and $\alpha_C \geq 0$ captures the relative weight given to

[6]Formally, we could assume that the leader obtains a payoff of 1 if she follows her supported strategy and of 0 otherwise and is unaffected by the baseline utility (1). Alternatively, we could assume that the leader cares so much about playing her preferred strategy that even if all her neighbors were using the opposite one, she would still obtain a higher payoff from using her favorite strategy.

conforming with the reference neighborhood of their peers, "the crowd."[7]

Each L player is surrounded by l_L LF players on her left and l_L LF players on her right, where $l_L > k/2$ is chosen randomly. All players who are not LF or L are CF. Define by l_C the number of CF between two groups of LF players. We assume that $l_C > k$ is a randomly chosen even number. The combined assumptions on l_C and l_L imply that the distance between two leaders is a random even number of at least $2k$.[8] The class (A or B) of the leader is also random.

We first analyze the stationary steady states with these fixed types. Once these stationary steady states are reached, we will allow for the possibility of types shifting over time between CF and LF if the payoff of one is higher than that of the other. Before doing so, however, we briefly discuss our main model assumptions.

3.1 Discussion of the main assumptions

One key feature of the model is that the interaction is local. The coordination games are played only with the k closest neighbors in the circle; the L players affect only LF players who are "nearby," and the CF are concerned with whether their actions are the same as those of their neighbors. Obviously, our modeling of leaders as having only local influence is a simplification. Some leaders do have local influence, and as discussed in the introduction, their influence has been extensively considered in the contexts of vaccination and epidemic containment, for example. However, there are indeed leaders with a wider influence. We will see in section 10 that our framework can handle the existence of global leaders without markedly changing our conclusions.

The other key assumption in our model is how we choose to model the different player types and their location. Our L players are stubborn agents who do not "choose" their strategy to entice more followers to opt for the

[7]Note that since the players are playing a coordination game, there is already a premium for conformity. We introduce α_C because it is important to understand the evolution of players' types in later sections.

[8]We assume that this distance is an even number to prevent players from having two closest leaders.

leaders' own path. It is simply that they are "born" with an opinion and they stick to it. They may be better or worse at being followed (i.e., they can have lower or higher "charisma"), but this is not a conscious choice. Thus, our L players are closer in spirit to citizen candidates (Besley and Coate 1998) than to professional politicians (Barro 1973). Canes-Wrone, Herron, and Shotts (2001) show that high-quality leaders, those with superior information, always signal truthfully even if doing so runs counter to voters' current opinions. There is also very good experimental evidence that leaders who change their positions to fit the situation are punished by followers (Tomz 2007).[9] We can conceive of leaders in the sense of the famous quote from Harry Truman: "To be able to lead others, a man must be willing to go forward alone."

Our modeling of two types of agents, leader followers and crowd followers, is motivated by the fact that sensitivity to leadership can vary considerably across the population. We will see that even crowd followers are affected by leadership because of their need to coordinate.[10]

The location and identity of leader followers LF can be rationalized as follows. Some players have an "innate preference" for action A or B. All else equal, they receive an extra payoff if they choose this preferred strategy, which we call α_L. Clearly a good leader L is capable of delivering a higher α_L, but this is also an "innate" parameter of leader L—her "charisma." Knowing that other A- or B-loving players are likely to place themselves close to leaders L espousing the corresponding strategy and that the former also care about their neighbors' choices, LF players accordingly choose locations close to their like-minded leaders. This assumption has empirical support: Connaughton and Daly (2004) and Meirovich and Ashita (2021) show a close association between identification with a leader, trust, and physical proximity, for example. In a context such as ours where there is also a need to coordinate activities, there is a reinforcing mechanism: By locating close to other supporters of

[9]Early in his paper, Tomz (2007) recalls, "In the second debate, for example, Bush stated that he did not see how Kerry 'could lead this country in a time of war, in a time of uncertainty, if [Kerry] changed his mind because of politics.' A country at war, he argued, 'requires a president who is steadfast and strong and determined.'"

[10]We could add more types of players with other levels of sensitivity to leadership, but this would clutter the results while adding little insight.

leader L, players can play like L, and the players close to them are more likely to coordinate. The CF players have no such preferences, so it is fine with them to be in more neutral territory further away from L players. The mechanism whereby this location outcome is implemented need not be particularly complex. If the space close to leaders L is allocated by means of an auction or a contest, the LF players would very naturally win those spaces. In addition, Van Huyck, Battalio and Beil (1993) show how a preplay auction mechanism can serve to align actions in a coordination game.

The other important distinction is between the LF players, who obtain a boost only by imitating the L player, and the CF players, who receive a boost by imitating their peers. We believe that this is a realistic feature of human interaction: We are a hierarchical species, but the group also matters to us. As we mentioned earlier, these concerns are probably present to different degrees in all people, but we simplify the analysis by assuming that only either concern with imitating the leader or concern with imitating one's peers is relevant at a given point in time for a specific person.

4 Stationary states with fixed types

We first analyze the stationary states of a dynamic process in which, at time $t = 0$, the LF players play the action of their closest leaders and CF players take a random action.[11] From this period onward, every player best-responds to the actions of the players relevant to her in the previous period. The types of the players stay fixed throughout. The best-response dynamics therefore determine whether a leader is successful, i.e., preserves loyal followers in steady state. We now show that while the success of leaders espousing the risk-dominant strategy (B-leaders) is guaranteed in steady state, leaders espousing the Pareto-dominant strategy (A leaders) might lose the loyalty of some or all their followers, and we discuss the consequences for steady-state play. Proposition 1, the main proposition in this section, is:

[11] We think this is the relevant initial condition since it converts the agents we label leaders in our model into real leaders by giving them at least an initially "loyal" followership choosing the action proposed by the leader. Otherwise, the leaders in our model would be indistinguishable from stubborn agents who always play the same strategy no matter what.

Proposition 1 *Suppose that* $d+e+2\alpha_L < b+f-\frac{2(d-f+b-e)}{k}$. *Then, if there is at least one B leader, everyone converges to playing B except the A leaders.*

Suppose that $b+f > d+e+2\alpha_L \geq b+f-\frac{2(d-f+b-e)}{k}$. *Then, everyone converges to playing B except the A leaders and the players between two consecutive A leaders, who can converge to all playing B or all playing A, depending on initial conditions.*

Suppose, on the other hand, that $d+e+2\alpha_L \geq b+f$. *Then, the LF regions next to A- and B leaders play the same actions as their leaders. The CF players between a B leader and an A leader converge to playing B. Moreover, CF regions between two consecutive A leaders can converge to all playing B or all playing A in a stationary state, depending on initial conditions.*

This proposition is proved with a sequence of lemmas, all of which refer to outcomes in stationary states.[12]

Lemma 1 *All LF players with a B leader always follow their leader choosing strategy B.*

Lemma 2 *If* $d+e+2\alpha_L \geq b+f$, *all LF players with an A leader follow their leader choosing strategy A.*

Lemma 3 *All CF players located in an area where at least one of the leaders is a B leader choose strategy B.*

Lemma 4 *Any B cluster of CF can invade the LF region of an A-leader until it runs into that leader if*

$$d+e+2\alpha_L < b+f \tag{2}$$

and jump to the LF followers on the other side of the leader if

$$d+e+2\alpha_L < b+f-\frac{2\,(d-f+b-e)}{k}, \tag{3}$$

in which case all the LF players of the A leader will switch to strategy B.

[12] All the lemmas in this section are formally proven in Appendix A.

Lemma 5 *There is no stationary configuration that is not a cluster of all A or all B among CF players between two A leaders.*

We now provide an intuitive discussion of how Proposition 1 follows from the lemmas and why they are important.

First, we start with Lemmas 1 and 2, which relate to the limit behavior of LF players. To understand whether LF players stay loyal to their leader, we have to look at the most distant and therefore most exposed LF, who is the LF sitting at the boundary between the LF and the CF players. In the initial round, all the LF players following the same leader play the same strategy, but the CF players can play a different strategy; therefore, the most exposed LF will have at least half of her neighbors playing the same strategy that she plays and at most half of her neighbors playing the other strategy. Loyalty to the leader is guaranteed if the most exposed LF still best-responds with the strategy proposed by her leader even in the worst-case scenario, when all her neighboring CF players play the other strategy. By the definition of risk dominance, this is always true for the most exposed LF with a risk-dominant leader; however, the most exposed LF of a Pareto-dominant and therefore risk-dominated leader requires a sufficiently high payoff from following the leader to maintain her loyalty; in other words, the A leader must be sufficiently charismatic.

Second, we study how CF players behave under a B-leader, as per Lemma 3. The conditions in Lemmas 2 and 1 for full loyalty to the leader depend only on the most exposed LF player in the circle and therefore extend to other interaction structures where we could have one or more of these most exposed players, e.g., an interaction structure in which all LF players play half of their coordination games with other LF players following their same leader and play the other half of the games with random members of society. The conditions ensure that LF-led regions are immune to invasions of the strategy not proposed by the respective leader. However, will the strategy played by the LF players spread to the CF regions? Risk-dominant play will spread from LF to CF regions since the boundary CF player has at least half of her neighbors playing the risk-dominant strategy (namely, all her LF neighbors), which by definition makes risk dominance the best response.

The third case relates to CF players between A- and B-leaders, as in Lemma 4. From the previous lemmas, all players between two B leaders end up always playing strategy B. What happens in areas between a B- and an A leader? We have already established that all LF players under the B leader's influence and all CF players choose strategy B. When can risk-dominant play invade the adjacent A-led region? Clearly, this occurs when the most exposed A-led LF does not stay loyal to her leader because the leader is not sufficiently charismatic. This will unravel to all A-led LF players being located on the side of the leader from which the B invasion occurs. In this case, all LF players of A leaders located between two B leaders will be invaded by risk-dominant play.

What happens if the A leader is located between a B leader and another A leader? Will she be able to serve as a barrier to protect her LF players on her other side (who are located between two A leaders) from the invasion of risk-dominant play triggered by the B-led region? Now, the most exposed LF is the player located directly next to the invaded A leader but on the other side of the invasion. Since the A leader is stubbornly playing strategy A, the most exposed LF player has two more neighbors playing the Pareto-dominant strategies than neighbors playing the risk-dominant strategy. Whether the invasion occurs now depends not only on the leader's charisma α_L but also on the size of the neighborhood k. If $k = 2$, there cannot be any invasion because the leader serves as a complete barrier against the risk-dominant invasion: Both neighbors of the LF player located next to the A leader on the other side (that is, both the A leader and the neighboring LF player) play strategy A, and therefore this player's best response is to stick to strategy A. When the size of the neighborhood grows, this player will encounter LF players on the other side of the leader who have been invaded by risk-dominant play. The larger the neighborhood, the easier is it for the invasion to jump the barrier of the leader, which can only be prevented by the leader's having sufficiently high charisma. Lemma 4 states the exact conditions under which risk-dominant play from a CF risk-dominant cluster spreads to an A-led LF region.

Under condition (3), an A leader whose closest leaders are of different classes cannot serve as a barrier against a one-sided B invasion. Moreover,

risk-dominant play will spread to the other side of the A leader and convert not only all of her followers but also the CF players next to the A-led area, who will then invade the LF players of the adjacent A-led region. The invasion will jump to the other side of the A leader and will continue to spread until all CF and LF players play strategy B. In steady state, everyone plays B except the A leaders.

If condition (3) is violated but condition (2) holds, an A leader whose closest leaders are of different classes might serve as a barrier against the invasion of risk-dominant play initiated by the B-led area, but risk-dominant play may still originate from the CF players located between two A-led areas. Whether this happens depends on the random initial conditions that affect the first-period choice of CF players located between two A leaders.

The final case is that of CF players between A-leaders, as in Lemma 5. If the CF players between two A leaders settle on a risk-dominant cluster, risk-dominant play will spread to the LF players between two A leaders if and only if condition (2) holds. If, on the other hand, the CF players between two A leaders settle on a Pareto-dominant cluster and condition (3) is violated, everyone between the two A leaders will choose the Pareto-dominant strategy. Hence, at intermediate levels of leader charisma (if condition (3) is violated but condition (2) holds), the players located between two A leaders will all play the same strategy; whether it is the Pareto-dominant or risk-dominant strategy is determined by initial conditions. Hence, LF players of A leaders may stay loyal to their leader, but the only role of the leader is to serve as an invasion barrier. The loyalty itself is driven by the random initial conditions that make CF players converge on A.

At high levels of leader charisma, the LF players stay fully loyal to their leader and are no longer affected by CF players between two A leaders, who will converge randomly to an A or B cluster.

To understand Proposition 1, let us consider the possible cases.

If the neighborhood parameter k is sufficiently large, there are three possible outcomes in a steady state of this game.

1. At a sufficiently low impact of leadership α_L, only the risk-dominant strategy B is capable of surviving.

2. At intermediate α_L, there is a possibility of the Pareto-dominant strategy A surviving, but this can occur only in regions between two adjacent A leaders and if the initial conditions happen to be conducive, in the sense that sufficient CF players played strategy A as their initial random action. Even between two A leaders, initial conditions may favor convergence to all-B play. In all other regions, everyone plays the risk-dominant strategy B.[13]

3. Finally, if the value of leadership is high enough, then with certainty the LF players play the same strategy in the coordination game as their closest leader all the time. In addition, the CF players in between A leaders can converge to playing the Pareto-dominant strategy A under appropriate initial conditions. All other CF players play the risk-dominant strategy B in the limit.[14]

A key aspect of this result is that once a sufficiently large cluster of agents playing one strategy or the other forms, the action happens at the boundaries of the cluster. This is why risk dominance is so important. Someone at the boundary has half of her neighbors playing one strategy and half of them playing the other. In the absence of extra elements, such as leadership, risk dominance would take over the population. This explains why, in Proposition 1, clustering and relatively strong A leaders are crucial for the survival of the Pareto-dominant (but risk-dominated) strategy in the limit. Propagation of the risk-dominant strategy does not rely on leadership as much since norm following and even the pure dynamic reaction over time are sufficient to keep this strategy in play.

The proposition does not mention the possibility of limiting states that are not stationary. There can be some of those between A leaders. For example, when $k = 2$ and $d + e + 2\alpha_L \geq b + f$, one can have a limit state in which the CF players between two A leaders fluctuate between $ABABABAB....AB$ and $BABABABA....BA$. The literature on chaotic dynamics usually calls such

[13]These results also hold at low α_L and sufficiently low k.

[14]Note that α_C does not play a role in the results of this section. This is because the coordination game already provides the premium to play the action more prevalent in the neighborhood of a CF player.

states "blinkers". They are a curiosity, but even if they exist, they do not qualitatively challenge our argument since, as we now show, such blinker states have a relatively large amount of A play.

Lemma 6 *The absorbing sets in which there is any A play must have at least 50% A play.*

Proof. In any state of an absorbing set, there cannot be any cluster with more than $k/2$ players using strategy B. In between those clusters, the sets that can survive with A play must have at least $k/2$ players using strategy A or they will be completely eliminated. Thus, there must be at least equal numbers of A and B players in the limiting states. ■

5 Stochastic stability

In Proposition 1, we saw that for $d + e + 2\alpha_L \geq b + f - \frac{2(d-f+b-e)}{k}$, there are multiple possible stationary states between two A leaders. A natural question in this context is whether one of those steady states is more likely to emerge in the long run. Borrowing techniques from Young (1993) and Kandori, Mailath and Rob (1993), one can answer that question if the evolution of strategic choice has some randomness. Rather than best-responding to the actions of other agents in previous periods, the agents best-respond to their environment with probability $1 - \varepsilon$ and choose the alternative action with probability ε. In this case, the result is straightforward to show.

Proposition 2 *Suppose that $b + f > d + e + 2\alpha_L \geq b + f - \frac{2(d-f+b-e)}{k}$. Then, the limit distribution of play between two consecutive A leaders as $\varepsilon \to 0$ gives probability one to all players choosing B.*

Suppose, on the other hand, that $d + e + 2\alpha_L \geq b + f$. Then, the limit distribution of play between two consecutive A leaders as $\varepsilon \to 0$ gives probability one to all crowd followers choosing B and all leader followers playing A.

Proof. See appendix. ■

The steady state with all-A crowd followers is still interesting because, if one starts from a steady state with all A, the transition time to an all-B situation with very small ε should be very long and the kind of social phenomena that we are interested in are likely to not last an exceedingly long amount of time.

6 Steady states for the evolution of types

In this section, we study the evolution of types after convergence to a stationary steady state over strategies in the coordination games has been reached. Even in one of the most extreme, and often pathological, versions of leader-followership, cult following, the literature (Bainbridge and Stark 1979, Rousselet et al. 2017) has long established that individuals come in and out of the cult. In the process, they consider the costs and benefits of doing so. Obviously, there is an attachment benefit (our α_L) lost upon an individual's moving out of the cult. However, life outside a cult could be more in sync with that in the rest of society (our α_C), with one's actions potentially better matching those of one's peers. This reasoning with respect to the decision of whether to follow a leader of course extends to more standard leadership situations. The literature has explored this issue, and it is quite complex (Välikangas, and Okumura 1997, Messick 2004, Liborius 2014); however, our simplified model captures important elements of the observations made in the empirical literature.

To be precise, we assume that a CF can become an LF if the payoff of a CF in the stationary steady state, at the player's current position and given the current population state, is lower than that of an LF and vice versa. In other words, players choose the type expected to maximize their utility in the current period.

In our study of the evolution of types, the premium given to crowd following becomes important. To clarify why, we compare the conditions under which a CF does better or worse than an LF playing the same strategy in the coordination games.

Lemma 7 *An LF following her leader outperforms a CF playing the same strategy as the leader iff*

$$\alpha_L > \frac{x_k}{k}\alpha_C, , \tag{4}$$

where x_k is the number of neighbors playing the same strategy as the player under consideration.

Proof. Since the LF and CF play the same strategy, they obtain the same payoffs from playing the coordination games, while the LF additionally obtains α_L since she follows her leader and CF obtains an extra payoff from conforming to the crowd $\frac{x_k}{k}\alpha_C$. The strategy with the higher extra payoff outperforms the other strategy. ■

Lemma 7 states that the extra payoff from leader following must be higher than the extra payoff from crowd following weighted by the proportion of people playing the strategy in the neighborhood. If $\alpha_C = 0$, then by Lemma 7, an LF always outperforms a CF playing the same strategy, so the introduction of the α_C parameter allows the types to be more balanced.

We want to understand whether the evolution of types favors or harms Pareto-dominant play. Here, we summarize our results; the exact propositions and their proofs can be found in Appendix C. We distinguish the different stationary steady states with fixed types and discuss the new steady states that can be reached when types can evolve. In the exposition, leaders' strategies are set in blackboard bold font ($\mathbb{A}$ or $\mathbb{B}$), LF players' in bold font ($\mathbf{A}$ or $\mathbf{B}$), and CF players' in italic font (A or B).

First, observe that the evolution of types does not change risk-dominant play between two risk-dominant leaders but does convert all the CF players into LF players or vice versa.

$$\mathbb{B}\underbrace{\mathbf{B}...\mathbf{B}}_{l_L}\ \underbrace{B.....B}_{l_C}\ \underbrace{\mathbf{B}...\mathbf{B}}_{l_L}\mathbb{B} \rightarrow \begin{cases} \mathbb{B}\underbrace{\mathbf{B}...\mathbf{B}\mathbf{B}.....\mathbf{B}\mathbf{B}...\mathbf{B}}_{l_L+l_C+l_L}\mathbb{B} \text{ for } \alpha_L > \alpha_C \\ \mathbb{B}\underbrace{B...BB.....BB...B}_{l_L+l_C+l_L}\mathbb{B} \text{ for } \alpha_L < \alpha_C \end{cases}$$

Nor does the evolution of types change the risk-dominant play between a risk-dominant and a Pareto-dominant leader with low charisma, i.e., when

$d + e + 2\alpha_L < b + f$.

$$\mathbb{A}\underbrace{\mathbf{B}...\mathbf{B}}_{l_L}\ \underbrace{B.....B}_{l_C}\ \underbrace{\mathbf{B}...\mathbf{B}}_{l_L}\mathbb{B} \rightarrow \begin{cases} \mathbb{A}\underbrace{B...BB..B}_{l_L+\frac{1}{2}l_C}\underbrace{\mathbf{B}..\mathbf{BB}...\mathbf{B}}_{l_L+\frac{1}{2}l_C}\mathbb{B} \text{ for } \alpha_L > \alpha_C \\ \mathbb{A}\underbrace{B...BB.....BB...B}_{l_L+l_C+l_L}\mathbb{B} \text{ for } \alpha_L < \alpha_C \end{cases}$$

The players whose closest leader is the Pareto-dominant leader will convert to CF since risk-dominant play is their best response.

However, when leaders are highly charismatic, i.e., when $d+e+2\alpha_L > b+f$ such that with fixed types their LF players stay loyal to them, when types can evolve, Pareto-dominant play will increase if the benefit from crowd following is not too high and will be harmed otherwise.

- For $d + e + 2\alpha_L > b + f + \alpha_C$

$$\mathbb{A}\underbrace{\mathbf{A}...\mathbf{A}}_{l_L}\ \underbrace{B.....B}_{l_C}\ \underbrace{\mathbf{B}...\mathbf{B}}_{l_L}\mathbb{B} \rightarrow \begin{cases} \mathbb{A}\underbrace{\mathbf{A}...\mathbf{AA}..\mathbf{AB}..\mathbf{BB}...\mathbf{B}}_{l_L+\frac{1}{2}l_C\qquad l_L+\frac{1}{2}l_C}\mathbb{B} \text{ for } \alpha_L > \alpha_C \\ \mathbb{A}\underbrace{A...A\mathbf{A}..\mathbf{AB}..\mathbf{B}B...B}_{l_L+\frac{1}{2}l_C\qquad l_L+\frac{1}{2}l_C}\mathbb{B} \text{ for } \alpha_L < \alpha_C \end{cases}$$

  Observe that when types can evolve, the most vulnerable type playing the Pareto-dominant strategy is an LF playing A just at the boundary of an all-B cluster played by CF players. When $d + e + 2\alpha_L > b + f + \alpha_C$, this most vulnerable LF-A player does not want to switch to CF-B but starts to invade the CF-B cluster. In this case, all players will choose the same strategy as their closest leader. They will all be LF iff $\alpha_L > \alpha_C$. Otherwise, the players closest to their leaders will be CF while those furthest away from the leader will be LF. The latter are close to the border between payoff-dominant and risk-dominant play and hence prefer to follow the leader and not the crowd since one section of their neighbors plays the other strategy.

- For $b + f < d + e + 2\alpha_L < b + f + \alpha_C$, the CF-B cluster starts to invade

the LF-A cluster; hence,

$$\mathbb{A}\underbrace{\mathbf{A}...\mathbf{A}}_{l_L}\ \underbrace{B.....B}_{l_C}\ \underbrace{\mathbf{B}...\mathbf{B}}_{l_L}\mathbb{B} \to \begin{cases} \mathbb{A}\underbrace{B...BB..B}_{l_L+\frac{1}{2}l_C}\underbrace{\mathbf{B}..\mathbf{BB}...\mathbf{B}}_{l_L+\frac{1}{2}l_C}\mathbb{B} \text{ for } \alpha_L > \alpha_C \\ \mathbb{A}\underbrace{B...BB.....BB...B}_{l_L+l_C+l_L}\mathbb{B} \text{ for } \alpha_L < \alpha_C \end{cases}$$

Similarly, when types are allowed to evolve, an excessively high α_C may harm Pareto-dominant play between two A leaders. In this case, in the stationary equilibria with fixed types, the CF players either converge to all-A or all-B play. In the former case, we obtain the following:

- For α_C high, Pareto-dominant play is eliminated:

$$\mathbb{A}\underbrace{\mathbf{A}..\mathbf{A}}_{l_L}\ \underbrace{A....A}_{l_C}\ \underbrace{\mathbf{A}..\mathbf{A}}_{l_L}\mathbb{A} \to \mathbb{A}\underbrace{B....BB.....BB...B}_{l_L+l_C+l_L}\mathbb{A}$$

$$\mathbb{A}\underbrace{\mathbf{A}..\mathbf{A}}_{l_L}\ \underbrace{B....B}_{l_C}\ \underbrace{\mathbf{A}...\mathbf{A}}_{l_L}\mathbb{A} \to \mathbb{A}\underbrace{B...BB.....BB...B}_{l_L+l_C+l_L}\mathbb{A}$$

- For low α_C, Pareto-dominant play always survives:

$$\mathbb{A}\underbrace{\mathbf{A}..\mathbf{A}}_{l_L}\ \underbrace{A....A}_{l_C}\ \underbrace{\mathbf{A}..\mathbf{A}}_{l_L}\mathbb{A} \to \begin{cases} \mathbb{A}\underbrace{\mathbf{A}...\mathbf{AA}.....\mathbf{AA}...\mathbf{A}}_{l_L+l_C+l_L}\mathbb{A} \text{ for } \alpha_L > \alpha_C \\ \mathbb{A}\underbrace{\mathbf{A}..\mathbf{A}}_{<l_L}\ \underbrace{A..AA.A}_{>l_C}\ \underbrace{\mathbf{A}..\mathbf{A}}_{<l_L} \text{ for } \left(\frac{\alpha_C}{2}+\frac{\alpha_C}{k}\right) < \alpha_L < \alpha_C \\ \mathbb{A}\underbrace{A...AA.....AA...A}_{l_L+l_C+l_L}\mathbb{A} \text{ for } \alpha_L < \left(\frac{\alpha_C}{2}+\frac{\alpha_C}{k}\right) \end{cases}$$

  or is even created:

$$\mathbb{A}\underbrace{\mathbf{A}..\mathbf{A}}_{l_L}\ \underbrace{B....B}_{l_C}\ \underbrace{\mathbf{A}...\mathbf{A}}_{l_L}\mathbb{A} \to \begin{cases} \mathbb{A}\underbrace{\mathbf{A}...\mathbf{AA}.....\mathbf{AA}...\mathbf{A}}_{l_L+l_C+l_L}\mathbb{A} \text{ for } \alpha_L > \alpha_C \\ \mathbb{A}\underbrace{A...AA.....AA...A}_{l_L+l_C+l_L}\mathbb{A} \text{ for } \alpha_L < \alpha_C \end{cases}$$

In short, the evolution of types can affect Pareto-dominant play only if leaders are sufficiently charismatic. In this case, if α_C is low, the evolution of

types can increase Pareto-dominant play. However, if α_C is high, the evolution of types can further harm Pareto-dominant play. A higher α_C reinforces the advantage of risk-dominant play.

7 Relative importance of A and B leaders

We have claimed that leadership is less important for the long-run survival of the risk-dominant but Pareto-dominated strategy B than for the Pareto-dominant but risk-dominated strategy A. To explore the extent to which this matters, we study what happens when the charisma of leaders disappears. In particular, we analyze the case where $\alpha_{L_B} = 0$ and $\alpha_{L_A} > 0$ and then observe what happens for $\alpha_{L_B} \geq 0$ and $\alpha_{L_A} = 0$.

Lemma 8 *The stationary steady states with fixed types as described in Proposition 1 are unaffected when $\alpha_{L_B} = 0$ and $\alpha_{L_A} > 0$. However, when $\alpha_{L_A} = 0$ and $\alpha_{L_B} \geq 0$, payoff-dominant play can occur only between two consecutive A leaders when $d + e \geq b + f - \frac{2(d-f+b-e)}{k}$ and initial conditions are favorable for the CF followers.*

Proof. Observe that even when $\alpha_{L_B} = 0$ when types are fixed, all LF players, including the furthest away from a B leader, will follow this leader choosing strategy B simply because B is risk dominant (see Lemma 1). On the other hand, when types are fixed, unless the A leader has a minimum of charisma, namely, $\alpha_{L_A} > \underline{\alpha_{L_A}} = \frac{b+f-(d+e)}{2}$ (see Lemma 2), the LF players furthest away from this A leader might not follow her and may deviate to the risk-dominant strategy, which will unravel to the LF closest to the A leader. The remainder of the lemma follows from setting $\alpha_{L_A} = 0$ in Proposition 1. ■

Given this result, to understand the implications of $\alpha_{L_B} = 0$ for the evolution of types, we need only to evaluate how this assumption affects Propositions 8 and 9 when $a_{L_A} > 0$. It is easy to see that, with $\alpha_{L_B} = 0$, the only difference with respect to the previous results is that, in regions close to B leaders, there will be CF players because B leaders can no longer attract LF players. However, strategy A still cannot invade regions with a B leader: These regions will be populated by CF players playing B. This is so because a CF playing B

does better than a CF playing A when half of her neighbors are playing B and half are playing A.While the existence of B leaders guarantees the formation of risk-dominant clusters, their charisma, i.e., how attractive they are, does not matter for the choice of actions.

On the other hand, since $\alpha_{L_A} = 0$ already affects the steady states when type are fixed, it also has important implications for the survival of all-A regions when types can evolve. Introducing this assumption into Proposition 8, we learn that an all-A cluster between two A leaders can survive the evolution of types with everybody becoming a CF between the two A leaders only if the payoff from rule following is sufficiently high. In particular, condition (9) needs to hold, which leads to

$$\alpha_C > \underline{\alpha_C} = \frac{k\left(b + f - (e + d)\right)}{2} - (d - f + b - e)\,.$$

The above results allow us to establish the following remarks, which provide further insights.

Remark 1 *The risk-dominant action B is always guaranteed to survive as long as there is a B leader somewhere, while the risk-dominated action disappears if α_L is low, and its expansion is always limited by the existence of a B leader. In other words, the risk-dominated action can never infiltrate regions where the closest leader is a B leader. The B leader is a shield against infiltration no matter how charismatic A leaders are.*

Remark 2 *If there were only A leaders, then, at sufficiently high α_L, the stationary steady state could converge to the whole population playing A.*

8 Policies to maximize payoff-dominant play

In this section, we examine different strategies that an authority might use to maximize A play. We first study the optimal distribution of a fixed amount of charisma among A leaders. Then, we allow the authority to strategically either remove or place a leader. These instruments may seem rather abstract, but there is a growing interest in the development literature in policies targeting

community leaders to effect behavioral change in their communities (see, e.g., Valente and Pumpuang 2007 and Vyborny 2021), and our framework can inform the implementation of such policies.

8.1 Optimal distribution of charisma

Assume that the amount of charisma α_{L_i} is specific to each leader and is also a choice variable of the authority, who has at her disposal the total amount of charisma $\Lambda = \sum_i \alpha_{L_i}$. We analyze how Λ should be distributed if this principal wants to maximize the amount of A play in the stationary state.

Lemma 9 *Charisma should be distributed to maximize the number of A leaders for whom* $\alpha_L^* = (b + f - (e + d))/2$ *and should be given to A leaders with the highest* l_L*. Moreover, consecutive A leaders should be favored to receive* α_L^**.*

Proof. By Lemma 2, α_L^* is the minimum amount of charisma that ensures that all leader-followers stay loyal to their A leader in the stationary steady state. Additional amounts of charisma will not lead to additional limit A play. In addition, lower amounts are not enough to keep the leader's followers with her, so they are not useful for the objective of the principal. The higher l_L is, the more A play due to LF players. By Lemma 5, for sufficiently high α_L, CF regions between two consecutive A leaders in a stationary steady state converge to either playing all A or playing all B, depending on initial conditions. Hence, if the A leaders are consecutive, there is a probability that the crowd-followers between them also converge to playing A. ■

Since l_C is also random in our model, a question arises: namely, how to allocate charisma if there is a choice between different pairs of consecutive A leaders. This is complicated by the fact that there is a tradeoff. One can choose A leaders with a high l_C between them. If play converges to an all-A state, this involves a larger number of A players. On the other hand, with a high l_C, there is a higher likelihood of an initial cluster of B play that then leads to an all-B state. Indeed,

Proposition 3 *From a random initial condition between two A leaders, if the distance between them becomes sufficiently large, in a stationary steady state, all CF players converge to playing B.*

Proof. If a k-cluster of B players forms, all players end up playing B. The chance of a k-player cluster forming at random at $t = 0$ increases as the distance between two A leaders grows. ■

In section 8.3, we explore in depth the tradeoff between the probability of convergence to all A and the number of players using A in the limit.

8.2 Removal of leaders

Suppose that a social planner considers removing B leaders to increase the amount of A play.[15] To avoid the simplistic case where all B leaders can be removed, suppose that she can remove only one leader.

Suppose that one B leader can be removed after play in the game with fixed types has reached the steady state or after the game with evolving types has reached the steady state.[16] Since the leader has been active before, now, in the first round after the removal, the LF types remain as LF, and the CF players remain as CF but reoptimize the strategy in the coordination game.

Proposition 4 *The removal of a B leader after the game with fixed types or evolving types has reached a stationary steady state makes a difference to the final outcome only when* $d + e + 2\alpha_L > b + f + \alpha_C$*, implying that (11) holds and at least one of the leaders closest to the removed B leader is an A leader.*

1. *If the removed B leader was located between two A leaders, all players formerly under the influence of this B leader play A.*

2. *If the removed B leader was located between an A- and a B-leader, the number of players using A will grow in the new area of influence of the A leader.*

[15] One alternative policy that produces qualitatively similar results is to reconfigure the network such that an A leader is proximate to a B cluster of LF players.

[16] In Appendix E, we study the case where the leader is removed at the beginning of the game.

Proof. See Appendix D. ■

The following corollaries are immediate consequences of Proposition 4:

Corollary 1 *The best candidate for removal of a B leader located between two A leaders is the one at the greatest distance from these two A leaders.*

Corollary 2 *The best candidate for removal of a B leader between an A leader and a B leader is the one resulting in the largest new area of influence of an A leader who was the unique closest A leader of the removed B leader.*

Corollary 3 *The overall gain in A play is greatest with the removal of a B leader between two A leaders if the area of influence of this B leader is larger than the largest new area of influence of an A leader who was the unique A leader closest to the removed B leader. Otherwise, the unique A leader should be removed.*

Observe that, independently of the timing of the removal of the B leader, the removal of a B leader between A leaders might enhance all-A play. If A leaders are sufficiently charismatic (α_L is large), then this is guaranteed if the removal happens either after the steady state in strategies is reached with fixed types or after the steady state of the evolution of types is reached. In both cases, the greatest impact emerges if the B leader who is removed is located between the two A leaders who are furthest apart. If the removal of the B leader happens at the beginning of the game (see Appendix E), whether A play is enhanced depends on initial conditions. By Proposition 3, this is more likely the closer the two consecutive A leaders surrounding the eliminated B leader are located. Of course, at the same time, if these two leaders are close to one another, the number of affected players is smaller.

If A leaders are sufficiently charismatic (α_L is large), A play also grows if a B leader located between an A leader and a B leader is removed either after the steady state in strategies is reached with fixed types or after the steady state of the evolution of types is reached. The area of influence of the neighboring A leader will grow, and everybody in this area will play the Pareto-dominant action. Hence the size of the growth of Pareto-dominant play corresponds to the size of the increase in the area of influence of this neighboring A leader.

8.3 Strategic placement of leaders

Suppose again that a social planner wanted to increase the number of people playing the Pareto-efficient outcome A and could strategically place a fixed number of A leaders in the circle. What would be the optimal location of those leaders? Since, from Propositions 1 and 8, A play by nonleaders can occur in steady state only when A leaders are sufficiently charismatic, this question is relevant only when $d + e + 2\alpha_L \geq b + f - \frac{2(d-f+b-e)}{k}$. We start our analysis with the case in which leaders are highly charismatic. We allow types to evolve as in section 6.

Proposition 5 *Suppose that $d + e + 2\alpha_L > b + f + \alpha_C$, i.e., that condition (11) holds and there are at least two A leaders to be placed. Then, one way to maximize A play is to place all B leaders in a cluster next to each other at the minimal possible distance and place on each side of the cluster an A leader at the minimal possible distance from the B leaders limiting the cluster.*

Proof. If condition (11) holds, by Proposition 9, everybody plays the same strategy as their closest leader. Hence, A play is maximized by minimizing the area of influence of B leaders. ■

If condition (11) is violated, A-play after the evolution of types can be achieved in steady state only between two consecutive A leaders and requires favorable initial conditions. Now, there is a first decision as to the optimal distance between the A leaders trading off the probability of converging to all-A play and the area of influence of the consecutive A leaders. Assume that the planner has solved this tradeoff,[17] and call this optimal distance h^*. Then, the only remaining question is the relative position of the A leaders among themselves and other B leaders.

Proposition 6 *Suppose that $b + f + \alpha_C > d + e + 2\alpha_L \geq b + f - \frac{2(d-f+b-e)}{k}$. The location of A leaders that maximizes the possibility of A play is a cluster of them next to one another at distance h^*.*

[17] We deal with this optimal tradeoff in detail in section 8.3.1.

Proof. Taking an A leader surrounded by B leaders and placing her next to a cluster of A leaders clearly increases the likelihood of steady state A play. The reason is that an isolated A leader never induces A play but two consecutive A leaders might do so. Similarly, merging two clusters of A leaders increases the likelihood of A play at the new boundary between the two clusters. ■

Clustering leaders of the same class always maximizes the probability that the Pareto-efficient outcome is reached. However, the reasons for clustering and the optimal distance between leaders depends on their charisma. If leaders are so charismatic that, after the evolution of types, everybody in the leaders' area of influence chooses their preferred action, the area of influence of B leaders should be reduced to the minimum. For less charismatic leaders, A play requires clustering of A leaders at the distance that optimally resolves the tradeoff between the area of influence of the A leaders and the probability of converging to all-A play.

8.3.1 Optimal placement of leaders: An explicit characterization

To give a more concrete illustration of how this tradeoff between the probability of converging to all-A play and the area of influence of the consecutive A leaders works, we study the optimal placement in detail. We then study its comparative statics. The free parameters in our model are the number of neighbors k, the number of crowd followers l_C, and the numbers in the payoff table for the game. The influence of the numbers in the payoff table can be summarized in m, the minimum number of extra players choosing A that would lead a player to prefer playing A when $k/2+m$ neighbors choose A and the rest choose B (m must be even in our model). Note that m can be defined as the smallest even natural number for which (5) holds.

$$d + e > b + f - \frac{m\,(d - f + b - e)}{k}. \tag{5}$$

Then, let $p\,(l_C, k, m)$ be the probability that a cluster with at least $k/2+2-m/2$ playing B forms at time zero among CF players. Observe that if a cluster of CF players with at least $k/2+2-m/2$ playing B forms, then we are guaranteed to converge to all B from the initial condition. We argue heuristically (in

Appendix F) that maximizing the objective function $(1 - p(l_C, k, m))\, l_C$ over l_C will generally yield an upper bound on the optimal l_C, and it is a feasible number to characterize.[18]

Proposition 7 *Assume that the objective function is* $(1 - p(l_C, k, m))\, l_c$*; then, there is generically one* $l_C^*(k, m)$ *that maximizes the objective. The* $l_C^*(k, m)$ *increases with* k *and decreases with* m*.*

Proof. See Appendix F. ■

The explanation for why $l_C^*(k, m)$ decreases with m is straightforward. With a larger m, the advantage of the risk-dominant strategy is larger. A smaller cluster of B players can invade because B is a best response with a lower probability of B play. Hence, there are many more possible initial conditions from which all-B play in the limit can be attained. For k, the reason is that the clusters that are needed to start a successful invasion are larger when k grows and the likelihood of their forming randomly at time zero is smaller.

9 Interaction in a lattice

Up to now, we have assumed that players interact in a circle. We now discuss (informally; for a formal discussion, see section G in the appendix) how our insights with fixed player types extend to more general patterns of interaction, a lattice with 2 dimensions, either infinite or folded around a three-dimensional torus so there are no boundaries. In this case, we assume that each player interacts with all players who are fewer than n steps away in each of the 2 dimensions. Hence, $(2n+1)^2 - 1 = k$ is the total number of neighbors for any player.

Note that, with a lattice, it is no longer true that the risk-dominant equilibrium always invades the population in a simple coordination game where

[18] Part of the reason why this is so is that some initial conditions will go to blinker states in the limit. At small k, it is not too difficult to deal with these cases, but for a general treatment of k, it is very complicated to completely characterize all initial conditions that lead to blinker states.

players best-respond to the population choices from the previous period. The reason why risk dominance is so powerful in a linear environment (the circle) is that the players at the boundary between clusters have half of their neighbors playing each strategy. (Recall that a risk-dominant strategy is defined precisely as the strategy that is a best response when half of the population uses it.) In a lattice, the neighborhood structure is different. In case some A clusters and some B clusters form, at the linear boundary between clusters of these two types, each person would interact with $n(2n+1)$ players of the different type and $n(2n+3)$ of the same type. For example, with $n=1$, we have a two-dimensional lattice, and players interact with the neighbors who are one step away in each dimension. Hence, the total number of neighbors $k=(2n+1)^2-1=8$. At the linear boundary between an A and a B cluster, each person would interact with 5 people of her own type and 3 people of the different type. Hence, for players at the boundary of an A cluster, only $\frac{3}{8}<\frac{1}{2}$ of their neighbors are using the risk-dominant strategy, so risk dominance is no longer a sufficient condition for the risk-dominant strategy to invade. This indicates that the payoff-dominant strategy has a better chance of survival in a lattice than in a circle.

Remark 3 *The survival of A is easier in the lattice than in the circle because exposure to outsiders is more limited. With more general interaction structures, Morris (2000) provides conditions under which risk-dominated strategies may survive even in the absence of leaders. A key condition for a group to be uninvadable is the exposure to external influence of the most externally connected person in the group (relative to her degree of connections inside the group).*

10 Global leaders

We have assumed throughout the paper that L players have local influence. This is realistic because some leaders do in fact have only local influence. However, there do exist people with a more global followership. We now argue that the existence of global leaders does not necessarily alter the qualitative

conclusions of our paper. Assume there are two global L players, one choosing action A and another choosing action B, whose action influences all LF players. The influence of the global L player choosing A can be written as α_L^{GA}, and the influence of the one choosing B is α_L^{GA}. It can be easily seen that the model would now be equivalent to one where a local LF close to an A leader would have utility $u_x + \alpha_L I_A + \alpha_L^{GA} I_A + \alpha_L^{GB} I_B$. However, since $I_B = 1 - I_A$, this is equivalent to $u_x + \left(\alpha_L + \alpha_L^{GA} - \alpha_L^{GB}\right) I_A + \alpha_L^{GB}$. Similarly, for an LF close to a B leader, her utility is $u_x + \left(\alpha_L + \alpha_L^{GB} - \alpha_L^{GA}\right) I_A + \alpha_L^{GA}$. Thus, the new situation is analogous to one where A- and B leaders now have different degrees of charisma. We have already studied the situation where leaders have different charisma (in section 7), and it is clear that what matters is the sign of $\alpha_L^{GA} - \alpha_L^{GB}$. If the global A leader has higher charisma, it will make survival of the A strategy more likely. Low charisma on the part of the global A leader, on the other hand, might even destroy the chance of local survival of the A strategy.

11 Conclusion

We have postulated a game in which leadership and norm following interact, in an environment where individuals play a coordination game with local interaction. We find that the survival of Pareto-efficient outcomes over time depends heavily on clustering and on the existence and strength of leaders willing to support the actions leading to those outcomes.

Several important extensions to this model could be considered. We assume that people either follow leaders or follow their peers. Mixed motivations could be important. The extent of peer influence is limited to a small environment, consistent with evidence about the cognitive limitations on the scope of human relationship networks (i.e., Dunbar numbers; see e.g., Dunbar 1992 and Dunbar and Shultz 2007). However, we have focused on particularly simple network structures, where the evolution is relatively tractable. More complex structures might produce interesting results. In particular, the effects of leadership that can reach differently sized segments of the population seem a worthwhile avenue for future research. In the same vein, a model that al-

lows leaders to influence the size of their followership and compete with other leaders for followers seems a fruitful avenue for future research.

References

Acemoglu, Daron, and Matthew O. Jackson. "History, expectations, and leadership in the evolution of social norms." The Review of Economic Studies 82.2 (2015): 423-456.

Acemoglu, Daron, and Matthew O. Jackson. "Social norms and the enforcement of laws." Journal of the European Economic Association 15.2 (2017): 245-295.

Afolabi, A. , & Ilesanmi, O. S. (2021). Addressing COVID-19 vaccine hesitancy: Lessons from the role of community participation in previous vaccination programs. Health Promotion Perspectives, 11(4), 434.

Alós-Ferrer, Carlos, and Simon Weidenholzer. "Contagion and efficiency." Journal of Economic Theory 143.1 (2008): 251-274.

Alós-Ferrer, Carlos, and Simon Weidenholzer. "Imitation, local interactions, and efficiency." Economics Letters 93.2 (2006): 163-168.

Ang, James B., Per G. Fredriksson, and Swati Sharma. "Individualism and the adoption of clean energy technology." Resource and Energy Economics 61 (2020): 101180.

Bainbridge, W. S., & Stark, R. (1979). Cult formation: Three compatible models. Sociological Analysis, 40(4), 283-295.

Ballester, C., Calvó-Armengol, A., & Zenou, Y. (2006). Who's who in networks. Wanted: The key player. Econometrica, 74(5), 1403-1417.

Barney, J. B. (1995). Looking inside for competitive advantage. Academy of Management Perspectives, 9(4), 49-61.

Belk, Russell and Gülnur Tumbat. "The Cult of Macintosh". Consumption Markets & Culture, 8.3 (2005): 205-217

Besley, T., & Coate, S. (1998). Sources of inefficiency in a representative democracy: A dynamic analysis, American Economic Review, 139–156.

Bicchieri, Cristina. The grammar of society: The nature and dynamics of social norms. Cambridge University Press, (2005).

Bicchieri, Cristina. Norms in the wild: How to diagnose, measure, and change social norms. Oxford University Press, (2016).

Bilancini, Ennio, and Leonardo Boncinelli. "Social coordination with locally observable types." Economic Theory 65.4 (2018): 975-1009.

Binmore, Ken, and Larry Samuelson. "An economist's perspective on the evolution of norms." Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft 150.1 (1994): 45-63.

Binmore, Ken. "Social norms or social preferences?." Mind & Society 9.2 (2010): 139-157.Burke, Mary A., and H. Peyton Young. "Social norms." Handbook of social economics. Vol. 1. North-Holland, (2011). 311-338.

Canes-Wrone, B., Herron, M. C., & Shotts, K. W. (2001). Leadership and pandering: A theory of executive policymaking. American Journal of Political Science, 532-550.

Chen, Hsiao-Chi, Yunshyong Chow, and Li-Chau Wu. "Imitation, local interaction, and coordination." International Journal of Game Theory 42.4 (2013): 1041-1057.

Clifford, P., & Sudbury, A. (1973). A model for spatial conflict. Biometrika, 60(3), 581-588.

Connaughton, Stacey L., and John A. Daly (2004). "Identification with leader: A comparison of perceptions of identification among geographically dispersed and co-located teams." Corporate Communications: An International Journal 9.2: 89-103.

Cui, Zhiwei. "More neighbors, more efficiency." Journal of Economic Dynamics and Control 40 (2014): 103-115.

Cummins, Denise. "Dominance, status, and social hierarchies." The handbook of evolutionary psychology (2005): 676-697.

DeGroot, M. H. (1974). Reaching a consensus. Journal of the American Statistical association, 69(345), 118-121.

Dhaliwal, B. K., Seth, R., Thankachen, B., Qaiyum, Y., Closser, S., Best, T., & Shet, A. (2023, June). Leading from the frontlines: community-oriented approaches for strengthening vaccine delivery and acceptance. In BMC proceedings (Vol. 17, No. Suppl 7, p. 5). London: BioMed Central.

Dunbar, Robin IM. "Neocortex size as a constraint on group size in pri-

mates." Journal of human evolution 22.6 (1992): 469-493.

Dunbar, Robin IM, and Susanne Shultz. "Evolution in the social brain." science 317.5843 (2007): 1344-1347.

Dwivedi, A., Dwivedi, P., Joshi, K., Sharma, V., Sengar, A., Agrawal, R., ... & Barthwal, M. (2022). Local leader's impact on adoption of renewable energy generation technology by rural communities in the Himalayan region. Journal of Cleaner Production, 352, 131479.

Dziubiński, M., & Goyal, S. (2017). How do you defend a network?. Theoretical Economics, 12(1), 331-376.

Dziubiński, M., Goyal, S., Vigier, A., Bramoullé, Y., Galeotti, A., & Rogers, B. (2016). Conflict and networks. The Oxford Handbook of the Economics of Networks.

Ellison, Glenn. "Learning, local interaction, and coordination." Econometrica: Journal of the Econometric Society 61.5 (1993): 1047-1071.

Eshel, Ilan, Larry Samuelson, and Avner Shaked. "Altruists, egoists, and hooligans in a local interaction model." American Economic Review 88.1 (1998): 157-179.

Fang, Tommy Pan, Andy Wu, and David R. Clough. "Platform diffusion at temporary gatherings: Social coordination and ecosystem emergence." Strategic Management Journal 42.2 (2021): 233-272.

Farrell, Joseph, and Garth Saloner. "Coordination through committees and markets." The RAND Journal of Economics (1988): 235-252.

Galeotti, A., Golub, B., & Goyal, S. (2020). Targeting interventions in networks. Econometrica, 88(6), 2445-2471.

Galeotti, A., & Goyal, S. (2009). Influencing the influencers: a theory of strategic diffusion. The RAND Journal of Economics, 40(3), 509-532.

Galeotti, A., & Goyal, S. (2010). The law of the few. American Economic Review, 100(4), 1468-1492.

Gilbert, Paul. "Evolution and social anxiety: The role of attraction, social competition, and social hierarchies." Psychiatric Clinics 24.4 (2001): 723-751.

Glynn, Eugene, Brian Fitzgerald, and Chris Exton. "Commercial adoption of open source software: an empirical study." International Symposium on Empirical Software Engineering, IEEE (2005).

Goyal, Sanjeev, and Fernando Vega-Redondo. "Network formation and social coordination." Games and Economic Behavior 50.2 (2005): 178-207.

Goyal, S., & Vigier, A. (2014). Attack, defence, and contagion in networks. The Review of Economic Studies, 81(4), 1518-1542.

Den Hartigh, E., Ortt, J. R., Van de Kaa, G., & Stolwijk, C. C. (2016). Platform control during battles for market dominance: The case of Apple versus IBM in the early personal computer industry. Technovation, 48, 4-12.

Hallgren, E., Moore, R., Purvis, R. S., Hall, S., Willis, D. E., Reece, S., ... & McElfish, P. A. (2021). Facilitators to vaccination among hesitant adopters. Human Vaccines & Immunotherapeutics, 17(12), 5168-5175.

Hauk E and Mueller H. (2015). Cultural leaders and the clash of civilizations. Journal of Conflict Resolution. 59(3):367–400

Holley, R. A., & Liggett, T. M. (1975). Ergodic theorems for weakly interacting infinite systems and the voter model. The annals of probability, 643-663.

Holmberg, Tove, Marcus Logander, and Fredrik Lindqvist. ""Living on the Edge"-A Case Study of Important Factors for the Survival of Apple Computers, Inc." (2005).

Hunter, D. S., & Zaman, T. (2022). Optimizing opinions with stubborn agents. Operations Research, 70(4), 2119-2137.

Iriberri, Nagore, and José-Ramón Uriarte. "Minority language and the stability of bilingual equilibria." Rationality and Society 24.4 (2012): 442-462.

Jackson, Matthew O., and Alison Watts. "On the formation of interaction networks in social coordination games." Games and Economic Behavior 41.2 (2002): 265-291.

Khan, Abhimanyu. "Coordination under global random interaction and local imitation." International Journal of Game Theory 43.4 (2014): 721-745.

King, Andrew J., Dominic DP Johnson, and Mark Van Vugt. "The origins and evolution of leadership." Current biology 19.19 (2009): R911-R916.

Lempert, Robert J., Alan H. Sanstad, and Michael E. Schlesinger. "Multiple equilibria in a stochastic implementation of DICE with abrupt climate change." Energy economics 28.5-6 (2006): 677-689.

Levine, David, Salvatore Modica and Aldo Rustichini. Cooperating through

leaders. Mimeo (2022).

Levy, Steven. Insanely great: The life and times of Macintosh, the computer that changed everything. New York: Penguin (2000)

Liborius, P. (2014). Who is worthy of being followed? The impact of leaders' character and the moderating role of followers' personality. The Journal of psychology, 148(3), 347-385.

Maner, Jon K., and Charleen R. Case. "Dominance and prestige: Dual strategies for navigating social hierarchies." Advances in experimental social psychology. Vol. 54. Academic Press, (2016): 129-180.

Meirovich, Gavriel, and Ashita Goswami (2021). "Psychosocial and tangible distance between a leader and a follower: The impact on dyadic relations." Journal of Leadership Studies 14.4: 6-20.

Morris, Stephen. "Contagion." The Review of Economic Studies 67.1 (2000): 57-78.

Müller, Malte, Christian Kimmich, and Jens Rommel. "Farmers' adoption of irrigation technologies: experimental evidence from a coordination game with positive network externalities in India." German Economic Review 19.2 (2018): 119-139.

Neufeld, Derrick J., Linying Dong, and Chris Higgins. "Charismatic leadership and user acceptance of information technology." European Journal of Information Systems 16 (2007): 494-510.

Ostrom, Elinor. "Collective action and the evolution of social norms." Journal of economic perspectives 14.3 (2000): 137-158.

Page, W. H., and Lopatka, J. E. (1999). Network externalities. Encyclopedia of law and economics, 760, 952-980.

Prummer A and Siedlarek JP. (2017). Community leaders and the preservation of cultural traits. Journal of Economic Theory 168:143–76

Rousselet, M., Duretete, O., Hardouin, J. B., & Grall-Bronnec, M. (2017). Cult membership: What factors contribute to joining or leaving?. Psychiatry Research, 257, 27-33.

Smith, K. B., Larimer, C. W., Littvay, L., & Hibbing, J. R. (2007). Evolutionary theory and political leadership: Why certain people do not trust decision makers. The Journal of Politics, 69(2), 285-299.

Sunstein, Cass R. "On academic fads and fashions." Mich. L. Rev. 99 (2000): 1251.

Tomz, M. (2007). Domestic audience costs in international relations: An experimental approach. International Organization, 61(4), 821-840.

Ushchev, Philip, and Yves Zenou. "Social norms in networks." Journal of Economic Theory 185 (2020): 104969.

Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior change. Health education & behavior, 34(6), 881-896.

Välikangas, L., & Okumura, A. (1997). Why do people follow leaders? A study of a US and a Japanese change program. The Leadership Quarterly, 8(3), 313-337.

Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil (1993). "Asset markets as an equilibrium selection mechanism: Coordination failure, game form auctions, and tacit communication." Games and Economic Behavior 5.3: 485-504.

Van Vugt, Mark, Robert Hogan, and Robert B. Kaiser. "leadership, followership, and evolution: some lessons from the past." American Psychologist 63.3 (2008): 182.

Venkatraman, N., and Chi-Hyon Lee. "Preferential linkage and network evolution: A conceptual model and empirical test in the US video game sector." Academy of Management Journal 47.6 (2004): 876-892.

Verdier T and Zenou Y. (2018*)*. Cultural leader and the dynamics of assimilation. Journal of Economic Theory 175:374–414

Vincenzo, J. L., Spear, M. J., Moore, R., Purvis, R. S., Patton, S. K., Callaghan-Koru, J., ... & Curran, G. M. (2023). Reaching late adopters: factors influencing COVID-19 vaccination of Marshallese and Hispanic adults. BMC Public Health, 23(1), 631.

Vyborny, K. (2021). Persuasion and public health: Evidence from an experiment with religious leaders during COVID-19 in Pakistan. Available at SSRN 3842048.

Wu, Ruijuan, Cheng Lu Wang, and Andy Hao. "What makes a fan a fan?: The connection between Steve Jobs and Apple fandom." Handbook of research on the impact of fandom in society and consumerism. IGI Global,

(2020): 378-396.

Yengoh, G. T., Armah, F. A., & Svensson, M. G. (2010). Technology adoption in small-scale agriculture.

Yildiz, E., Ozdaglar, A., Acemoglu, D., Saberi, A., & Scaglione, A. (2013). Binary opinion dynamics with stubborn agents. ACM Transactions on Economics and Computation (TEAC), 1(4), 1-30.

Young, H. Peyton. "The evolution of conventions". Econometrica 61 (1993): 57-84.

Young, H. Peyton. "Social norms and economic welfare." European Economic Review 42.3-5 (1998): 821-830.

Zenou, Y. (2016). Key players. Oxford Handbook on the Economics of Networks, 244-274.

Zhou, J., & Chen, Y. J. (2015). Key leaders in social networks. Journal of Economic Theory, 157, 212-235.

Zimmermann, M. G., & Eguíluz, V. M. (2005). Cooperation, social networks, and the emergence of leadership in a prisoner's dilemma with adaptive local interactions. Physical Review E, 72(5), 056118.

A Proofs of lemmas leading to Proposition 1.

Lemma 1: LF **with a** B**-leader always follows her Leader in choosing strategy** B.

Proof. It suffices to look at the LF most distant from her B leader, who has a payoff of at least $\frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)+\alpha_L$ from choosing B and at most $\frac{1}{k}\left(d\frac{k}{2}+e\frac{k}{2}\right)$ from choosing A; thus, from $b+f>d+e$, strategy B is the best response. ■

Lemma 2: If $d+e+2\alpha_L \geq b+f$ **all** LF **players with an** A**-leader follow their leader in choosing strategy** A.

Proof. It again suffices to look at the LF most distant from the A leader. She has a payoff of at least $\frac{1}{k}\left(d\frac{k}{2}+e\frac{k}{2}\right)+\alpha_L$ from choosing A and of at most $\frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)$ from choosing B, so from $d+e+2\alpha_L \geq b+f$, she continues to play A. ■

Lemma 3: All CF players located in an area where at least one of the leaders is a B-leader choose strategy B.

Proof. Take a CF located next to a B-led region where all LF players play B by Lemma 1. Her payoff from playing B is at least $\frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)+\alpha_C\frac{1}{2}$. Her payoff from playing A is at most $\frac{1}{k}\left(d\frac{k}{2}+e\frac{k}{2}\right)+\alpha_C\frac{1}{2}$. Hence, she will choose B because $b+f>d+e$. By induction, all the CF players next to a B-led region flip to B. ■

Lemma 4: Any B cluster of CF players can invade the LF region of an A leader till the invasion reaches the leader if $d+e+2\alpha_L < b+f$ and jump to the LF players on the other side of the leader if $d+e+2\alpha_L < b+f-\frac{2(d-f+b-e)}{k}$ in which case all the LF players of the A-leader will switch to strategy B.

Proof. We know that the LF most distant from her A leader facing a B cluster invasion (from the left) has a payoff $\frac{1}{k}\left(d\frac{k}{2}+e\frac{k}{2}\right)+\alpha_L$ from choosing A and of $\frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)$ from choosing B, so from $d+e+2\alpha_L < b+f$, she flips to playing B. By induction, this frontier keeps advancing until it reaches the A leader. Now, the LF to the right of the A leader has a payoff $\frac{d}{k}\left(\frac{k}{2}+1\right)+\frac{e}{k}\left(\frac{k}{2}-1\right)+\alpha_L$. Her payoff from playing B is $\frac{b}{k}\left(\frac{k}{2}-1\right)+\frac{f}{k}\left(\frac{k}{2}+1\right)$, so she flips if

$$d+e+2\alpha_L < b+f-\frac{2\left(d-f+b-e\right)}{k}.$$

■

Lemma 5: There is no stable configuration that is not a cluster of all A or all B among CF players between two A-leaders.

Proof. Take a CF player who is in a sector with x A neighbors and $k-x$ B neighbors. The payoff from A is $\frac{xd+(k-x)e}{k}+\alpha_C\frac{x}{k}$. The payoff from B is $\frac{xb+(k-x)f}{k}+\alpha_C\frac{k-x}{k}$. A is better than B if

$$\frac{x}{k} > \frac{(f-e)+\alpha_C}{(d+f)-(e+b)+2\alpha_C}. \tag{6}$$

Suppose that 6 holds for an A sitting next to a B to the left of B. Then, we will show that B wants to flip to A. Observe that the difference in the neighborhood between A and B is that there is one person to the extreme left

of the A interval—call her C—who does not belong to the B interval and one person to the extreme right of the B interval who does not belong to the A interval—call her D—and A has B as a neighbor and B has A as a neighbor. Assume first that condition (6) holds.

Case 1. C is A and D is A. Then, B has one more A neighbor than A, so B wants to switch to A.

Case 2. C is A and D is B. Then, B has the same number of A neighbors as A, so B wants to switch to A.

Case 3. C is B and D is A. Then, B has two more A neighbors than A, so B wants to switch to A.

Case 4. C is B and D is B. Then, B has one more A neighbor than A, so B wants to switch to A.

By induction, this unravels to all players being CF between two A leaders.

Suppose that (6) does not hold for an A sitting next to a B to the left of B. Then, an analogous argument shows that A wants to flip to B. By induction, this unravels to all players being CF between two A leaders. ■

B Proof of Proposition 2

Proof. Take first $d + e + 2\alpha_L \geq b + f$. From Proposition 1, we know that the only stationary steady states have all CF players choosing A or all CF players choosing B. To exit the state where all CF players are choosing A toward a state where all CF players choose B, it is enough that a cluster of mutants with $k/2 + 1$ players that plays B be created. On the other hand, to exit the state where all CF players are choosing B toward a state where all CF players choose A, we need at least half of the players between the two A leaders to switch to A.

To see this, note that, in this case, there can never be a cluster of more than $k/2 - 1$ B-CF players together. This means that, next to any group of $k/2-1$ B-CF players, there must be A-CF players. To prevent their switching back to B, these must number at least $k/2$—which means that, from the initial position, there must be more A-CF players than B-CF players. This is in the most favorable case where having one more neighbor playing A than B makes

A the best reply, which will not always hold.

Since the number of CF players between two leaders is larger than k, the number of mutations required to go from all A to all B is smaller than the number of mutations required to go from all B to all A. Then, from Young (1993) and Kandori, Mailath and Rob (1993), the result follows.

The other case follows a similar argument.

In case the steady state is a blinker, as before, to shift from a blinker to all B, only $k/2-1$ mutants are needed. On the other hand, we know from Lemma 6 that at least half the population plays A. Hence, the number of mutants needed to go from all B to a blinker will be larger than $k/2-1$. ■

C Evolution of types

We first study leaders who are not sufficiently charismatic to keep all their followers loyal. In this case, the evolution of types never increases Pareto-dominant play but might harm it by converting some all$-A$ regions into all-B regions.

Proposition 8 *Let $d+e+2\alpha_L < b+f$.*

All-B regions before the evolution of types remain all-B regions after the evolution of types. Within these all$-B$ regions, everybody closest to an A leader now is a CF, while all players closest to a B leader are LF iff $\alpha_L > \alpha_C$ and are CF otherwise.

For all-A regions before the evolution of types, which can arise for $d+e+2\alpha_L \geq b+f-\frac{2(d-f+b-e)}{k}$, different cases apply:

1. *They remain all-A regions after the evolution of types:*

   (a) *Everybody becomes LF playing A if*

$$d+e+2\alpha_L > b+f+\alpha_C\left(1-\frac{2}{k}\right)-\frac{2\left(d-f+b-e\right)}{k} \qquad (7)$$

   and $\alpha_L > \alpha_C$.

(b) *All former CF players playing A continue to play A and partially invade the LF players playing A but do not reach all the way to the A leader if (7) holds and $\alpha_C > \alpha_L > (\frac{1}{2} + \frac{1}{k})\alpha_C$. The first LF not to convert is the LF with the smallest y, where y is the number of her neighbors playing B such that, by Lemma 7, condition (4) holds for $x_k = k - y$.*

(c) *All players become CF players playing A if*

$$\alpha_L < \left(\frac{1}{2} + \frac{1}{k}\right)\alpha_C \tag{8}$$

and

$$b + f < e + d + \alpha_C \frac{4}{k} + \frac{2\,(d - f + b - e)}{k}. \tag{9}$$

2. *They become part of an all-B region after the evolution of types with only CF when both (7) and (9) do not hold.*

Proof.

Suppose that $d+e+2\alpha_L < b+f-\frac{2(d-f+b-e)}{k}$, implying that everyone except the A leaders chooses strategy B before the evolution of types. Then, the LF next to an A leader who played B will switch to a CF playing B because she does not receive any benefit from following the leader but does obtain benefits from following the crowd followers who all play the same strategy. For the LF and CF closest to a B leader, they choose to be LF by Lemma (7) iff $\alpha_L > \alpha_C$ since $x_k = k$ because all neighbors play B.

Suppose that $b + f > d + e + 2\alpha_L \geq b + f - \frac{2(d-f+b-e)}{k}$. In this case, all the regions playing B continue to play B, but the CF players playing B will invade the LF players who play B in regions next to an A leader because these LF players do not follow their leader and choose B, so they are better off following the crowd. All the players in a B region who are closest to a B leader will become LF iff $\alpha_L > \alpha_C$ since they are surrounded by only all$-B$ neighbors and become CF iff $\alpha_L < \alpha_C$.

Now, we study what happens to the regions between two A leaders who converged to playing A before the evolution of types set in. We first study the choice of the most exposed LF player. For illustrative purposes and without

loss of generality, consider a sequence of leaders B, A, A and assume that all-A play has been reached between the two A leaders. The most exposed LF is the LF to the right of the first A leader, who faces B play to the left of the A leader. Hence, all of her $\frac{k}{2}$ neighbors to the right play A, while to her left, the leader plays A and the remaining $\frac{k}{2}-1$ neighbors play B. Therefore, this most vulnerable LF prefers to remain an LF playing A instead of switching to a CF playing B if

$$\frac{d}{k}\left(\frac{k}{2}+1\right)+\frac{e}{k}\left(\frac{k}{2}-1\right)+\alpha_L>\frac{b}{k}\left(\frac{k}{2}-1\right)+\frac{f}{k}\left(\frac{k}{2}+1\right)+\frac{\alpha_C}{k}\left(\frac{k}{2}-1\right),$$

which simplifies to (7).

This most vulnerable LF prefers switching to being a CF playing A instead of remaining an LF playing A if

$$\frac{d}{k}\left(\frac{k}{2}+1\right)+\frac{e}{k}\left(\frac{k}{2}-1\right)+\alpha_L<\frac{d}{k}\left(\frac{k}{2}+1\right)+\frac{e}{k}\left(\frac{k}{2}-1\right)+\frac{\alpha_C}{k}\left(\frac{k}{2}+1\right),$$

which simplifies to (8).

If we combine (7) and (8), the condition to remain an LF is

$$2\alpha_L>\max\{\alpha_C\left(1+\frac{2}{k}\right),b+f-(d+e)+\alpha_C\left(1-\frac{2}{k}\right)-\frac{2\left(d-f+b-e\right)}{k}\}. \tag{10}$$

Note that, for this most vulnerable person, being a CF playing B is worse than being a CF playing A if

$$\frac{b}{k}\left(\frac{k}{2}-1\right)+\frac{f}{k}\left(\frac{k}{2}+1\right)+\frac{\alpha_C}{k}\left(\frac{k}{2}-1\right)<\frac{d}{k}\left(\frac{k}{2}+1\right)+\frac{e}{k}\left(\frac{k}{2}-1\right)+\frac{\alpha_C}{k}\left(\frac{k}{2}+1\right),$$

which simplifies to (9).

1. Assume that (7) holds and $\alpha_L>\alpha_C$, which implies that (8) is violated. Then, the most vulnerable LF will remain an LF playing A. Moreover, everybody between two A leaders becomes an LF playing A since LF-A dominates CF-A even when all neighbors play A.

2. Assume that (7) holds and (8) is violated. Moreover, $\alpha_L<\alpha_C$, so that, in

combination with (8) being violated, the parameter restriction becomes $\alpha_C > \alpha_L > (\frac{1}{2} + \frac{1}{k})\alpha_C$. In this case, the CF players playing A invade the LF regions but do not take it over completely. The CF invasion stops at the greatest distance $k - y$ from the A (where y is the number of neighbors in the B region of the A leader), satisfying $\alpha_L > \frac{k-y}{k}\alpha_C$, which guarantees that condition (4) of Lemma 7 is satisfied.

3. Assume that (8) holds and (9) holds. The CF players playing A dominate both CF players playing B and LF players playing A, so everyone becomes a CF playing A.

4. Assume that (7) and (9) are both violated. Then, CF players playing B dominate both CF players playing A and LF players playing A. The most exposed player will switch to a CF playing B, converting her neighbor into the most exposed player in the same position. By induction, the former all-A region becomes an all-B region, with all players between the A leaders becoming CF players playing B.

■

We next study leaders who are sufficiently charismatic to keep all their followers loyal before the evolution of types. The next proposition shows that the extra benefit from crowd following is crucial in determining whether the evolution of types favors or harms the spread of Pareto-optimal play.

Proposition 9 *Suppose that* $d + e + 2\alpha_L \geq b + f$.

1. *Suppose that*

$$d + e + 2\alpha_L > b + f + \alpha_C. \tag{11}$$

   (a) *If* $\alpha_L > \alpha_C$*, all players will become LF players following the strategy of their closest leader.*

   (b) *If* $\alpha_L < \alpha_C$*, all players will play the same strategy as the closest leader in the coordination game, but some will be CF and some will be LF, in particular:*

i. All players located between two leaders of the same class will be CF.

ii. For players located between two leaders of different classes, those furthest away from their closest leader will be LF, imitating their leader, while those sufficiently close to the leader will be CF, choosing the same strategy as their closest leader. By Lemma 7, the number of neighbors x_k who play the same strategy in the coordination game as the player in question is defined by the lowest x_k for which condition (4) is violated.

2. If (11) does not hold, so that

$$b + f + \alpha_C > d + e + 2\alpha_L \geq b + f \ ,$$

all players between two B leaders will play B and are LF iff $\alpha_L > \alpha_C$ and are CF otherwise. All players between two leaders with different strategies will play B in the coordination game where everybody closest to an A leader becomes a CF, while all players closest to a B-leader are LF iff $\alpha_L > \alpha_C$ and are CF otherwise. All players between two B leaders will also play B and are LF iff $\alpha_L > \alpha_C$ and are CF otherwise. We must distinguish the following cases for players between two A leaders:

(a) If before the evolution of types the CF players converged to playing B, those CF players invade the area of LF-A, and in the long run, all players become CF players playing B.

(b) For all-A regions between two A leaders before the evolution of types, the results of Proposition 8 apply.

Proof.

Look at the B region boundary with an A-LF who must decide whether to switch to being a CF playing B (she cannot switch to being a CF playing A because, with half of the neighborhood playing B, a CF always plays B).

This person remains an LF if

$$\begin{aligned} \frac{1}{k}\left(d\frac{k}{2}+e\frac{k}{2}\right)+\alpha_L &> \frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)+\alpha_C\frac{1}{2} \\ d+e+2\alpha_L &> b+f+\alpha_C, \end{aligned}$$

which coincides with (11). Under (11), all the CF players playing B next to an A-LF playing A decide to switch to LF players playing A as long as their closest leader plays A since they face exactly the tradeoff described to derive (11). When the closest leader switches to B, then the CF players playing B must decide whether to become LF players playing B. The first one who is surrounded on one side by all-B players and on the other by all-A players compares

$$\begin{aligned} \frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)+\alpha_L &> \frac{1}{k}\left(b\frac{k}{2}+f\frac{k}{2}\right)+\alpha_C\frac{1}{2} \\ b+f+2\alpha_L &> b+f+\alpha_C, \end{aligned}$$

which definitely holds because $b+f>d+e$ and (11*)* holds. The next player compares

$$\frac{1}{k}\left(b\frac{k+1}{2}+f\frac{k-1}{2}\right)+\alpha_L > \frac{1}{k}\left(b\frac{k+1}{2}+f\frac{k-1}{2}\right)+\alpha_C\frac{k/2+1}{k},$$

so the frontier keeps advancing until $\frac{x_k}{k} < \frac{a_L}{\alpha_C} = d^* > \frac{1}{2}$ so that condition (4) of Lemma 7 is violated.

If $\frac{a_L}{\alpha_C} = d^* > 1$, this will never happen, and then all CF players playing B closest to a B leader will become LF players playing B.

If (11) holds and $\alpha_L > \alpha_C$ or equivalently $d^* > 1$, the A-LF players will advance as long they are closest to an A leader, and everybody closest to a B leader becomes an LF playing B. All areas between two A leaders are converted to A-LF areas that invade either the former B-CF area (since (11) holds) or the former A-CF area (since $\alpha_L > \alpha_C$) located in between these two A leaders. All players located between two B leaders are surrounded by only B neighbors and will become B-LF players since $\alpha_L > \alpha_C$.

If (11) holds and $\alpha_L < \alpha_C$, or equivalently $d^* < 1$, then at the frontier between an A-LF area and a B-CF area, the A-LF will invade the neighboring B-CF players as long they are closest to an A leader, and players closest to a B leader play B. Those whose closest B leader is nearest the frontier of the all-A area, so that $\frac{x_k}{k} > d^*$, are LF players playing B. All the others with a closest B leader are CF players playing B. Everyone located between two leaders of the same type becomes a CF playing the strategy of her nearest leader since $\alpha_L < \alpha_C$ and the leader is always surrounded by players playing her same strategy in the coordination game.

If (11) does not hold, then the A-LF players switch to B-CF until they reach the A leader. From this point on, the analysis that we performed for the previous proposition holds.

D Proof of Proposition 4

Observe that eliminating a B leader located between two B leaders will never make a difference regardless of whether this leader is removed in the steady state with fixed or with evolving types. It is also easy to see that removing a B leader when $d+e+2\alpha_L < b+f$ can never make a difference. By Proposition 8, no new all-A regions can evolve, and some might be preserved if we already have an all-A region between two A leaders before the evolution of types, so removing a B leader will not create a new all-A region. Let $b+f < d+e+2\alpha_L$, and consider that the leader is removed after the steady state in fixed types has been reached and before the evolution of types has started. If condition (11) does not hold, no new A clusters can be created since the B-CF of the removed B leader will invade any A-LF, and there is no difference in final outcome. Thus, assume that condition (11) holds.

When $b + f < d + e + 2\alpha_L$, all LF players with an A leader follow their leader in choosing strategy A before the types are allowed to change. Thus, if we eliminate a B leader between two A leaders, all the LF players of that eliminated B leader now have as their closest leader an A leader and play A. This proves (1) for leader removal after the steady state in fixed types has been reached. Similarly, if the eliminated B leader is located between an A- and a

B leader, the LF players of the eliminated leader closest to the A leader will play strategy A, while those closest to the B leader will play strategy B in the first round after the removal. All the CF players will continue playing B since B is risk-dominant and at least half of their neighbors play B. We have shown in the proof of Proposition 9 that if (11) holds when the possibility to change types holds, A-LF will invade the B-CF regions, and everyone will play the same strategy in the coordination game as their closest leader. Therefore, if a B leader between two A leaders is removed, the entire influence area of the removed B leader converts to playing A, while if the removed B leader had only one closest A leader, A play grows in the new influence area of this A leader. This proves (2) for leader removal after the steady state in fixed types has been reached.

Now, consider the case where the removal of one B leader happens after the steady state in the evolution of types has been reached.

Assume that $d + e + 2\alpha_L > b + f + \alpha_C$. When a B leader between two A leaders is removed, in the first round, the B-LF players around the removed B leader become A-LF. By Proposition 9, if $\alpha_L > \alpha_C$, this is the final outcome. If $\alpha_L < \alpha_C$, then in the first round, the B-CF players around the removed B leader remain B-CF because at most half of their neighbors play A. From the next round onward, the B-CF players start being invaded by the A-LF players, while on the other side, the A-LF area is invaded by the A-CF players who border the A-LF. In the end, though, all those players will be A-CF, and hence, all players formerly under the influence of the removed B leader will be added A players. This proves (1) for leader removal after the steady state in the evolution of types has been reached. When a B leader between an A- and a B leader is removed, in the first round, only the B-LF players around the removed B leader who fall into the new area of influence of the A leader become A-LF. If $\alpha_L > \alpha_C$ all players in the new influence area of the A leader become A-LF and this is the final outcome. If $\alpha_L < \alpha_C$ then by Proposition 9, the former B-LF players who become A-LF in the first round are those located furthest away from the removed B leader and hence bordering the A-LF area before the removal of the B leader. The B-CF players now under the influence of the A leader continue playing B since B is risk dominant and at

least half of their neighbors play B. Since (11) holds, A-LF players can invade these $C-LF$ players, while on the other side, the A-LF area is invaded by the A-CF players who border the A-LF area. In the end, though, everyone under the influence of the A leader will play A, with those closest to the A leader being CF and those furthest away being LF, so everyone in the new influence area of the A leader will be a new A player. This proves (2) for leader removal after the steady state in the evolution of types has been reached.

■

E Removal of leaders at the beginning of the game

We now study leader removal at the beginning of the game after all leaders are already located in the circle. Since the removed leader has never been active, we assume that the LF players surrounding that B leader become CF players because now they are no longer located next to a leader.

Proposition 10 *Suppose that there are at least* 2 B *leaders and one* B *leader is taken out before the game starts. The only removal of a* B *leader that can make a difference in increasing* A *play is that of a* B *leader whose nearest leaders on both sides are* A *leaders when* $d+e+2\alpha_L \geq b+f-\frac{2(d-f+b-e)}{k}$.

Proof. This is a corollary of Proposition 1. ■

The situation is now identical to that in the beginning of the game in general but with one fewer B leader. We know from Proposition 1 that, if $d+e+2\alpha_L < b+f-\frac{2(d-f+b-e)}{k}$, everybody except A leaders plays B in the limit so the removal of B-leaders does not make a difference. When $d+e+2\alpha_L \geq b+f-\frac{2(d-f+b-e)}{k}$, there can be clusters of A players in between two A leaders. Hence, change can only occur if the extirpated B leader is in between two A leaders.

When removing one B leader surrounded by the two A leaders before the game has started, the social planner faces the same tradeoff as in section 8.1. The probability of reaching an A cluster will be very small if the distance

between two A leaders is large. However, this is how one can create a large A cluster.

F Proof of Proposition 7

We first develop the heuristic argument for focusing on our objective function.

Remark 4 *Let $q(l_c, k, m)$ be the probability of all the other initial states that converge to all B outside those counted in $p(l_c, k, m)$. Then, a more precise objective function (albeit one less feasible to characterize) is $(1 - p(l_c, k, m) - q(l_c, k, m)) l_c$. The characterization of the solution would use*

$$\begin{aligned} & (1 - p(l_c, k, m) - q(l_c, k, m)) l_c - (1 - p(l_c - 2, k, m) - q(l_c - 2, k, m)) (l_c - 2) \\ = \; & 2 - (p(l_c, k, m) - p(l_c - 2, k, m)) l_c - p(l_c - 2, k, m) \\ - \; & (q(l_c, k, m) - q(l_c - 2, k, m)) l_c - q(l_c - 2, k, m). \end{aligned}$$

We use in the characterization

$$2 - (p(l_c, k, m) - p(l_c - 2, k, m)) l_c - p(l_c - 2, k, m),$$

so the optimal l_c^* with the objective function including $q(l_c, k, m)$ will be lower if

$$-(q(l_c, k, m) - q(l_c - 2, k, m)) l_c - q(l_c - 2, k, m) < 0. \tag{12}$$

Clearly, $-q(l_c, k, m)$ is negative, so the only reason (12) would not be true is if $-(q(l_c, k, m) - q(l_c - 2, k, m)) l_c > 0$. This can actually happen when l_c is relatively large because, when l_c is large (relative to k), most of the probability of arriving at all B arises from initial conditions with at least $k/2+2-m/2$-sized B clusters. However, large values of l_c are also clearly not optimal. Thus, in fact, the relevant cases appear those for which $-(q(l_c, k, m) - q(l_c - 2, k, m)) l_c < 0$.

Proposition 7: *Assume that the objective function is $(1 - p(l_c, k, m)) l_c$; then, there is generically one $l_c^*(k, m)$ that maximizes the objective. The $l_c^*(k, m)$ increases with k and decreases with m.*

Proof. Let $C(l_C)$ be the total number of initial configurations with at least $k/2+2-m/2$-sized B clusters.

Let n_B the number of B players in a cluster. Note that there is just one configuration with a cluster where all the CF players play B no matter what the size of l_C is. Similarly, the number of initial configurations where clusters with $-x$ positions fewer than the maximum possible is always the same no matter the value of l_C. Observe that, when we move to a cluster of size l_C-1, there are always just two configurations where the cluster is located at the two extremes of the CF players. For smaller-sized clusters, we need to distinguish two situations: (i) that where the cluster is located at the border of the LF players, in which case it has one CF neighbor playing B, and (ii) that where the cluster is interior among the CF players and therefore has two CF neighbors playing B.

In case (i), for each of the 2 border positions, $2^{(l_c-(n_B+1))}$ clusters exist, and hence, the total number is $2*2^{(l_c-(n_B+1))}=2^{(l_c-n_B)}$.

In case (ii), there are (l_C-n_B-1) interior positions, for each of which $2^{(l_c-(n_B+2))}$ clusters exist, and hence, there is a total of $(l_C-n_B-1)2^{(l_c-(n_B+2))}$ clusters. Given this, when we increase the size of l_C (which is even), we need to add to $C\,(l_C-2)$ the cases of the smallest $(n_B=k/2+2-m/2)$ and second-smallest $(n_B=k/2+3-m/2)$ cluster sizes. Therefore, $C(l_C)$ can be defined recursively as follows:

$$\begin{aligned} C(l_c) &= C(l_c-2)+2^{(l_c-(k/2+2-m/2))} \\ &+ (l_C-(k/2+2-m/2)-1)2^{(l_c-(k/2+2-m/2+2))} \\ &\quad +2^{(l_c-(k/2+3-m/2))}+(l_C-(k/2+3-m/2)-1)2^{(l_c-((k/2+3-m/2)+2))} \end{aligned} \tag{13}$$

Of the minimum-sized cluster, there are $2^{(l_C-(k/2+2-m/2))}$ configurations where that cluster is at the edge of the interval and $(l_C-(k/2+2-m/2)-1)2^{(l_C-(k/2+2-m/2+2))}$ in the interior. Analogously, with the second-smallest cluster, there are $2^{(l_C-(k/2+3-m/2))}$ configurations at the edge and $(l_C-(k/2+3-m/2)-1)2^{(l_C-((k/2+3-m/2)+2))}$ in the interior.

The formula 14 simplifies to:

$$\begin{aligned} C(l_C) &= C(l_C-2)+2^{(l_C-(k/2+2-m/2))}+(l_C-(k/2-m/2)-3)2^{(l_C-(k/2-m/2+4))} \\ &\quad +2^{(l_C-(k/2+3-m/2))}+(l_C-(k/2-m/2)-4)2^{(l_C-((k/2-m/2)+5))}. \end{aligned}$$

We are now in a position to define $p\left(l_C,k,m\right)$

$$\begin{aligned} p\left(l_C,k,m\right) &= \frac{C(l_C)}{2^{l_C}}=\frac{C(l_C-2)}{2^{l_C}}+\frac{2^{(l_C-(k/2+2-m/2))}}{2^{l_C}} \\ &+ (l_C-(k/2-m/2)-3)\frac{2^{(l_C-(k/2+2-m/2+2))}}{2^{l_C}} \\ &\quad +\frac{2^{(l_C-(k/2+3-m/2))}}{2^{l_C}}+(l_C-(k/2-m/2)-4)\frac{2^{(l_C-((k/2-m/2)+5))}}{2^{l_C}} \\ &= \frac{C(l_C-2)}{2^{l_C}}+2^{-(k/2+2-m/2)}+2^{-(k/2+3-m/2)} \\ &\quad +(l_C-(k/2-m/2)-3)2^{(-(k/2+2-m/2+2))} \\ &+ (l_C-(k/2-m/2)-4)2^{(-((k/2-m/2)+5)).} \end{aligned}$$

Note that the last two terms are increasing in l_C and the second and third are independent. Since the recursive terms all have the same structure, the full formula is increasing in l_C.[19]

The change in the expected number of l_C is:

$$\begin{aligned} &\left(1-p\left(l_C,k,m\right)\right)l_C-\left(1-p\left(l_C-2,k,m\right)\right)\left(l_C-2\right) \\ &=2-\left(p\left(l_C,k,m\right)-p\left(l_C-2,k,m\right)\right)l_C-p\left(l_C-2,k,m\right). \end{aligned}$$

$p\left(l_C,k,m\right)$ is increasing in l_C, so $-\left(p\left(l_C,k,m\right)-p\left(l_C-2,k,m\right)\right)l_C$ is negative. Thus, the change might be first positive for $l_C=l_C^{\min}=k+2$ when $\left(p\left(l_C^{\min},k,m\right)-p\left(l_C^{\min}-2,k,m\right)\right)$ is not too large (if not, the max is already at $l_C^{\min}$) and then negative, so there is a unique maximum generically.

For the comparative statics, we need to show that $2-\left(p\left(l_C,k\right)-p\left(l_C-2,k\right)\right)l_C-p\left(l_C-2,k\right)$ moves up with k (and down with

[19]Except possibly for the first term. However, the variation in l_c is important only for $p\left(l_C,k,m\right)-p\left(l_C-2,k,m\right)$, and in this computation, the first terms of the recursion cancel out.

m), such that the new zero will be to the right of the old one so the maximum will be higher with k and lower with m.[20]

Let $C(l_C)$ be the number of initial conditions given m, k, for which a cluster with at least $k/2 + 2 - m/2$ players play B:

$$\begin{aligned}
C(l_C) &= C(l_C - 2) + 2^{(l_C-(k/2+2-m/2))} \\
&\quad +(l_C - (k/2 + 2 - m/2) - 1)2^{(l_C-(k/2+2-m/2+2))} \\
&\quad +2^{(l_C-(k/2+3-m/2))} + (l_C - (k/2 + 3 - m/2) - 1)2^{(l_C-((k/2+3-m/2)+2))} \\
C(l_C) &= C(l_C - 2) + 2^{(l_C-(k/2+2-m/2))} \\
&\quad +(l_C - (k/2 - m/2) - 3)2^{(l_C-(k/2-m/2+4))} \\
&\quad +2^{(l_C-(k/2+3-m/2))} + (l_C - (k/2 - m/2) - 4)2^{(l_C-((k/2-m/2)+5)).}
\end{aligned}$$

Note that

$$p\left(l_C, k, m\right) = \frac{C(l_C)}{2^{l_C.}}$$

We first check that $p\left(l_C, k, m\right)$ increases with k and decreases with m:

$$\begin{aligned}
\frac{C(l_C)}{2^{l_C}} &= \frac{C(l_C - 2)}{2^{l_C-2}} + \frac{2^{(l_C-(k/2+2-m/2))}}{2^{l_C}} + (l_C - (k/2 - m/2) - 3)\frac{2^{(l_C-(k/2-m/2+4))}}{2^{l_C}} \\
&\quad +\frac{2^{(l_C-(k/2+3-m/2))}}{2^{l_C}} + (l_C - (k/2 - m/2) - 4)\frac{2^{(l_C-((k/2-m/2)+5))}}{2^{l_C}} \\
&= \frac{C(l_C - 2)}{2^{l_C-2}} + 2^{(-(k/2+2-m/2))} + (l_C - (k/2 - m/2) - 3)2^{(-(k/2-m/2+4))} \\
&\quad +2^{(-(k/2+3-m/2))} + (l_C - (k/2 - m/2) - 4)2^{(-((k/2-m/2)+5)).}
\end{aligned}$$

Thus, since $\frac{C(l_C)}{2^{l_C}}$ is simply a summation of terms that all increase with k and decrease with m, the result follows.

We now check that the differences also increase with k and decrease with

[20] Observe that this needs to be done for l_c constant, so the first term in the recursion does not shift and we thus do not need to characterize it.

m :

$$\begin{aligned}\frac{C(l_C)}{2^{l_C}}-\frac{C(l_C-2)}{2^{l_C-2}} &= \left(\frac{C(l_C-2)}{2^{l_C}}-\frac{C(l_C-4)}{2^{l_C-2}}\right)+\frac{2^{(l_C-(k/2+2-m/2))}}{2^{l_C}}\\ &+(l_C-(k/2-m/2)-3)\frac{2^{(l_C-(k/2-m/2+4))}}{2^{l_C}}\\ &+\frac{2^{(l_C-(k/2+3-m/2))}}{2^{l_C}}+(l_C-(k/2-m/2)-4)\frac{2^{(l_C-((k/2-m/2)+5))}}{2^{l_C}}\\ &-\left(\frac{2^{(l_C-(k/2+2-m/2)-2)}}{2^{l_C-2}}+(l_C-(k/2-m/2)-3-2)\frac{2^{(l_C-(k/2-m/2+4)-2)}}{2^{l_C-2}}\right)\\ &-\left(\frac{2^{(l_C-(k/2+3-m/2)-2)}}{2^{l_C-2}}+(l_C-(k/2-m/2)-4-2)\frac{2^{(l_C-((k/2-m/2)+5)-2)}}{2^{l_C-2}}\right)\end{aligned}$$

$$\begin{aligned}\frac{C(l_C)}{2^{l_C}}-\frac{C(l_C-2)}{2^{l_C-2}} &= \left(\frac{C(l_C-2)}{2^{l_C}}-\frac{C(l_C-4)}{2^{l_C-2}}\right)+2^{(-(k/2+2-m/2))}\\ &+(l_C-(k/2-m/2)-3)2^{(-(k/2-m/2+4))}\\ &+2^{(-(k/2+3-m/2))}+(l_C-(k/2-m/2)-4)2^{(-((k/2-m/2)+5))}\\ &-\left(2^{(-(k/2+2-m/2))}+(l_C-(k/2-m/2)-3-2)2^{(-(k/2-m/2+4))}\right)\\ &-\left(2^{(-(k/2+3-m/2))}+(l_C-(k/2-m/2)-4-2)2^{(-((k/2-m/2)+5))}\right)\end{aligned}$$

$$\begin{aligned}\frac{C(l_C)}{2^{l_C}}-\frac{C(l_C-2)}{2^{l_C-2}} &= \left(\frac{C(l_C-2)}{2^{l_C}}-\frac{C(l_C-4)}{2^{l_C-2}}\right)+2*2^{(-(k/2-m/2+4))}+2*2^{(-((k/2-m/2)+5))}\\ &= \left(\frac{C(l_C-2)}{2^{l_C}}-\frac{C(l_C-4)}{2^{l_C-2}}\right)+2^{(-(k/2-m/2+3))}+2^{(-((k/2-m/2)+4)).}\end{aligned}$$

Thus, since $\frac{C(l_C)}{2^{l_C}}-\frac{C(l_C-2)}{2^{l_C-2}}$ is just a summation of terms that all increase with k and decrease with m, the result follows. ■

G Interaction in a lattice

The L players are given a neighborhood with a fixed number of LF players in both dimensions—call it $l_L > n$. All players who are not LF or L are CF. Define by l_C the number of CF players between the closest two groups

of LF players in any dimension. We assume that $l_C > 2n$. The combined assumptions of l_C and l_L imply that the smallest distance between two leaders is at least $2l_L + 2n$ in any dimension. Respecting this condition, we assume that leaders are placed at random in the lattice and that their type A or B is also random. For the ease of exposition, we set $\alpha_C = 0$.

As we mention in the body of the paper, the payoff-dominant strategy has a better chance of survival in a lattice than in a circle. Clusters of payoff-dominant strategies may survive even in the absence of charismatic A leaders if risk dominance is not too strong. Hence, we need to strengthen risk dominance if we want the risk-dominant strategy to have a similar advantage as in the circle.

Proposition 11 *Assume that*

$$(b-e)\frac{(n+1)^2-1}{(2n+1)^2-1} > (d-f)\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right) \tag{14}$$

and set $\alpha_C = 0$.

Then, all LF players with a B leader always follow their leader in choosing strategy B. All CF players adjacent to a B-led region also choose strategy B.

Moreover, if

$$(d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L+\frac{1}{k}\left(b+d-(f+e)\right) < (b-e)\frac{n(2n+1)}{(2n+1)^2-1},$$

these B players will fully invade a neighboring A-led region. If

$$\begin{aligned} & (d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L+\frac{1}{k}\left(b+d-(f+e)\right) \\ > \quad & (b-e)\frac{n(2n+1)}{(2n+1)^2-1} > (d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L, \end{aligned}$$

the B invasion will stop once it has converted the boundary of the LF area where the A leader is located into playing B if there is no other B-led region in the neighborhood invading from another direction.

All the LF players of an A leader stay loyal to the A leader if $(d-f)\frac{(n+1)^2-1}{(2n+1)^2-1}+$

$\alpha_L > (b-e)\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)$.

Proof. We proceed by establishing each part of the proposition separately.

Claim 1 *LF players with a B leader always follow their leader in choosing strategy B.*

Proof. It suffices to show that the most distant and most exposed LF players with a B leader will not want to switch to playing A. These most distant and exposed LF players live at the corners of the square of LF players with their B leader in the center of the square, and all players outside this square are CF players. Hence the most distant and exposed LF players have $(n+1)^2-1$ neighbors who are LF players playing B and are surrounded by a fraction $\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)$ of CF players who, in the worst case, all play A. Thus, these most exposed LF players obtain a payoff of at least $b\frac{(n+1)^2-1}{(2n+1)^2-1}+f\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)+\alpha_L$ from choosing B and at most $d\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)+e\frac{(n+1)^2-1}{(2n+1)^2-1}$ from choosing A. Hence, for B to be a best response, we need $b\frac{(n+1)^2-1}{(2n+1)^2-1}+f\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)+\alpha_L > d\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)+e\frac{(n+1)^2-1}{(2n+1)^2-1}$ or, equivalently, $(b-e)\frac{(n+1)^2-1}{(2n+1)^2-1}+\alpha_L > (d-f)\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)$, which is implied by Assumption 14. ■

Claim 2 *All CF players located in an area where at least one of the leaders is a B leader end up choosing strategy B.*

Proof. We know from Lemma 1 that a square of LF players playing B will form around a B leader. Let us consider the CF players located at the frontline of this square. By Assumption 14, the best response for all CF players with at least a fraction of LF neighbors $\frac{(n+1)^2-1}{(2n+1)^2-1}$ playing B is to play B. Note that this applies to all the CF players at the frontline of the LF-Bsquare who are located at two steps or more from a corner of the Bsquare since the fraction of their neighbors in the LF-B square is exactly $\frac{(n+1)^2-1}{(2n+1)^2-1}$. Hence, all these CF players will play B. Now, consider the players who are one step from the corner of the B square: They used to have at least a fraction of $\frac{n(n+1)}{(2n+1)^2-1}$ neighbors playing B, namely, all their LF neighbors in the B square, but now

they also have $2n-1$ CF players playing B due to those located at two steps or more from a corner of the B square now choosing B with certainty. Thus, the fraction playing B for those CF players located at one step from the corner of the LF-B square is now at least $\frac{n(n+1)+2n-1}{(2n+1)^2-1} = \frac{n^2+3n-1}{(2n+1)^2-1} \geq \frac{(n+1)^2-1}{(2n+1)^2-1} = \frac{n^2+2n}{(2n+1)^2-1}$, and hence, by Assumption 14, they will change to playing B. Finally, consider the corner players, who originally had at least a fraction of $\frac{n^2}{(2n+1)^2-1}$ neighbors playing B, namely, all their LF neighbors in the B square. Now, however, at least $2n$ of their CF neighbors are best responding by playing B. Thus, the fraction playing B is now at least $\frac{n^2+2n}{(2n+1)^2-1} = \frac{(n+1)^2-1}{(2n+1)^2-1}$. Therefore, by Assumption 14, they will switch to playing B. Hence, all the CF players at the frontline of the LF-B square are choosing B. Now, by the same argument, we have a new frontline of CF players located next to the B square that will convert to playing B. The result that all CF players next to a B-led region end up choosing B follows by induction. ■

Claim 3 *Any B cluster of CF players can invade the LF region from one dimension of an A leader till it reaches the leader if* $(b-e)\frac{n(2n+1)}{(2n+1)^2-1} > (d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L$ *and can jump to the LF followers on the other side of the leader if* $(d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L+\frac{1}{k}(b+d-(f+e)) < (b-e)\frac{n(2n+1)}{(2n+1)^2-1}$ *in which case all the LF players of the A leader will switch to strategy B.*

Proof. We know that the LF players most distant from their A leaders who are located at the boundary of the B cluster (from one dimension) have $n(2n+1)$ B neighbors and therefore a payoff of $d\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+e\frac{n(2n+1)}{(2n+1)^2-1}+\alpha_L$ from choosing A and a payoff of $b\frac{n(2n+1)}{(2n+1)^2-1}+f\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)$ from choosing B, so if $b\frac{n(2n+1)}{(2n+1)^2-1}+f\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right) > d\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+e\frac{n(2n+1)}{(2n+1)^2-1}+\alpha_L$ or, Equivalently,

$$(d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L < (b-e)\frac{n(2n+1)}{(2n+1)^2-1},$$

they switch to playing B. By induction, this frontier keeps advancing until it reaches the border where the A leader is located and all the LF players in this boundary also switch to playing B. Now, the LF players to the other side of the

A leader have a payoff $d\left(1-\frac{n(2n+1)}{(2n+1)^2-1}+\frac{1}{k}\right)+e\left(\frac{n(2n+1)}{(2n+1)^2-1}-\frac{1}{k}\right)+\alpha_L$ from playing A. Their payoff from playing B is $b\left(\frac{n(2n+1)}{(2n+1)^2-1}-\frac{1}{k}\right)+f\left(1-\frac{n(2n+1)}{(2n+1)^2-1}+\frac{1}{k}\right)$, so they switch if

$$\begin{aligned} & d\left(1-\frac{n(2n+1)}{(2n+1)^2-1}+\frac{1}{k}\right)+e\left(\frac{n(2n+1)}{(2n+1)^2-1}-\frac{1}{k}\right)+\alpha_L \\ < \; & b\left(\frac{n(2n+1)}{(2n+1)^2-1}-\frac{1}{k}\right)+f\left(1-\frac{n(2n+1)}{(2n+1)^2-1}+\frac{1}{k}\right) \end{aligned}$$

or, equivalently,

$$(d-f)\left(1-\frac{n(2n+1)}{(2n+1)^2-1}\right)+\alpha_L+\frac{1}{k}\left(b+d-(f+e)\right)<(b-e)\frac{n(2n+1)}{(2n+1)^2-1}.$$

■

Claim 4 *If* $(d-f)\frac{(n+1)^2-1}{(2n+1)^2-1}+\alpha_L>(b-e)\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)$ *, all LF players with an A leader follow their leader in choosing strategy A.*

Proof. It again suffices to look at the LF players most distant from the A leader (those located at the corner of the LF players' A-led square) and show that they do not want to switch. They have a payoff of at least $d\frac{(n+1)^2-1}{(2n+1)^2-1}+e\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)+\alpha_L$ from choosing A and of at most $b\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)+f\frac{(n+1)^2-1}{(2n+1)^2-1}$ from choosing B, so they keep playing A if $(d-f)\frac{(n+1)^2-1}{(2n+1)^2-1}+\alpha_L>(b-e)\left(1-\frac{(n+1)^2-1}{(2n+1)^2-1}\right)$. ■

This concludes the proof of Proposition 11. ■

Assumption 14 guarantees that the most distant and most exposed LF with a B leader will always follow her B leader. This player is located at a corner of the square of LF players who in the first period all choose the same strategy as their leader. This LF player therefore has $(n+1)^2-1$ LF neighbors out of her $k=(2n+1)^2-1$ neighbors. All the remaining neighbors are CF players who might be playing strategy A. Hence, if this player wants to stick to B when all her CF neighbors are playing A, then all LF players with a B leader will follow their leader. By Assumption 14, this is indeed the case even if the B leader lacks charisma ($\alpha_L=0$). Hence, a square of LF players playing B forms around a B leader.

Similarly, Assumption 14 also guarantees that all CF players adjacent to a B-led region will also choose strategy B. This guarantees that all CF players located directly at the boundary but at least two steps away from the corner of the LF-B square around the B-leader will choose strategy B, and this will spread first to the entire boundary and then to the entire CF region. If there is an adjacent A-led region, this region will be invaded unless the A leader is sufficiently charismatic (sufficiently high α_L), which mirrors Lemma 4 for the circle.

In contrast to what holds for the circle, it is not always true that, between A leaders, there is convergence to all CF players playing A or all playing B, even when Equation 14 holds. To give an example, consider a situation in which $n = 1$. If a cluster with a square of four CF players using B forms in that region and the remaining CF players use A, this can be stable. The B players have 3 B neighbors, and if Assumption 14 holds, namely, $\frac{3}{8}(b-e) > \frac{5}{8}(d-f)$, that is enough for them to keep playing B. The most exposed CF players playing A have six neighbors playing A, and hence, it is possible that this is also enough for them to keep playing A, namely, if $\frac{2}{8}(b-e) < \frac{6}{8}(d-f)$. In this case, the situation is stable. However, if the B players form a cluster of 9 players, they would invade, since the most exposed A player now would have 3 neighbors playing B and hence, by Assumption 14, would convert to playing B.

In the same vein, let us consider a case where Assumption 14 does not hold. Then, in the previous example, a cluster of 9 B players would not invade the area between two A leaders. In addition, in that case, consider a situation with $\alpha_L = 0$ and half the plane occupied by players using A. Then, it is no longer true that B will invade the A region since the A players at the boundary each have 5 neighbors playing A and only 3 playing B. If Assumption 14 is not satisfied, the A strategy may survive even in the absence of charismatic A leaders.[21]

[21] In this specific example, A would invade, but this is not true in general for $n > 1$.