Purges and Predictability

Sam Kapon
UC Berkeley, Haas BPP

Matteo Camboni
UWisconsin–Madison*

September 23, 2026

Abstract

We study a novel dynamic game in which a principal can purge a group of agents, sequentially targeting its members for elimination. Agents can organize a protest that, if sufficiently large, can end the purge permanently. When targeting is predictable, next period's targets coordinate with current targets, limiting the purge to few agents and at most one period. When targeting is unpredictable, currently untargeted agents face diffuse risk of near-term elimination, weakening their incentive to protest. Unpredictability typically benefits the principal by allowing multi-period purges that eliminate a larger share of agents. We illustrate these results with historical elite purges.

1 Introduction

Authoritarian leaders purge their enemies, their friends, and the masses. These purges often unfold gradually, with unpredictable criteria for selecting targets (Gregory, 2009). Participation in repression and past loyalty offer no guarantee of safety: authoritarian leaders routinely turn against old allies and powerful insiders in their quest to centralize power. Potential victims understand that they too may be targeted and, especially early in the process, often collectively possess the capacity to resist. Yet in many historical purges, such as those perpetrated by Stalin or Saddam Hussein, elites remained largely compliant.

*Kapon: skapon@berkeley.edu; Camboni: camboni@wisc.edu. We are grateful to Nageeb Ali and Ernesto Dal Bó for many useful comments. We also thank Matilde Bombardini, Matias Iaryczower, Alessandro Lizzeri, Massimo Morelli, Harry Pei, Marzena Rostek, Konstantin Sonin, Francesco Squintani, and Francesco Trebbi. We also thank audiences at SITE Political Economy 2026, Wallis Institute 2026, IBEO (Alghero) 2026, McMaster, and UC Berkeley.

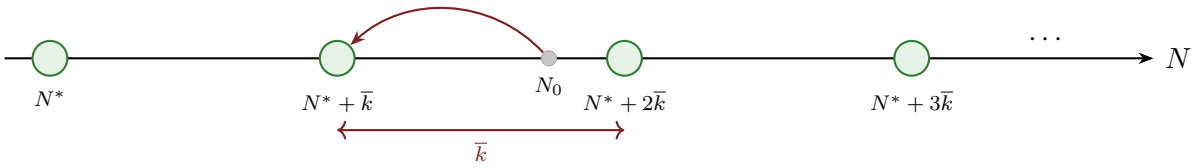
We provide a theory of how the predictability of future targeting shapes this compliance and, more broadly, the dynamics of sequential purges. When future targeting is predictable, for example because the leader's priorities can be inferred from observable characteristics like faction, rank, or past behavior, elites who expect to be targeted are incentivized to join current targets in resisting the leader. This generates blocking coalitions that limit purges to few agents and at most one period. When instead targeting is unpredictable, the risk of near-term elimination is spread across all untargeted elites, weakening their individual incentives to resist. Unpredictability can therefore sustain multi-period purges that eliminate a larger share of agents.

We formalize this mechanism in a dynamic infinite-horizon game between a principal and a finite group of forward-looking agents whom she may repeatedly target for elimination (or equivalently, purge). While in power, the principal's flow payoff is single-peaked in the number of survivors, which captures the trade-off between consolidating authority and retaining enough agents to govern. Agents receive a flow payoff while alive and zero thereafter. In each period, the principal targets a subset of survivors for elimination. After observing the target set, all survivors (targeted and untargeted) simultaneously choose whether to protest. A protest that reaches a known participation threshold $\overline{k}$ succeeds with positive probability. A successful protest permanently stops the purge and leaves the principal with zero continuation payoff. If such a protest fails, both targeted and protesting agents are eliminated, and the game continues. We focus on environments in which defeat is so costly for the principal that she prefers to forgo elimination rather than trigger a protest with a positive probability of success.

To abstract from coordination frictions and isolate the role of target uncertainty, we restrict attention to subgame-perfect equilibria (SPE) that are robust to profitable joint one-shot deviations by coalitions of agents, i.e., coalition-proof SPE. We compare two polar cases for target predictability: one in which untargeted agents can identify who would be targeted next if the purge continued, and one in which all untargeted agents perceive the same risk of future targeting. In the model, we capture these environments through *predictable-order* and *unpredictable-order* coalition-proof SPE, as if the principal followed a prespecified priority order or selected targets uniformly at random, respectively. As the microfoundation in Section 6 clarifies, however, the economic distinction behind these labels can be interpreted as concerning agents' information, not the principal's literal targeting procedure.

The contrast between the two environments is stark. Under predictable-order targeting, agents who would be eliminated next can identify themselves and prefer to join current targets in a protest, even at the risk of immediate elimination. Instead, under unpredictable-order targeting, these agents perceive the same near-term elimination risk as every other untargeted agent, weakening their incentives to protest. This fragments opposition and makes blocking coalitions harder to assemble. As a result, while under predictable-order targeting purges are limited to fewer than $\overline{k}$ agents and at most one period, under unpredictable-order targeting they can extend to multiple periods and eliminate a much larger share of the population.

To understand our results, we first describe the case of predictable-order targeting. Equilibrium behavior can be characterized by what we call *stable* population sizes, namely population sizes at which no further eliminations occur. The figure below shows the sequence of stable population sizes starting from the principal's ideal size N^*, where further elimination would be suboptimal, and depicts a typical one-period elimination path from initial population size N_0.

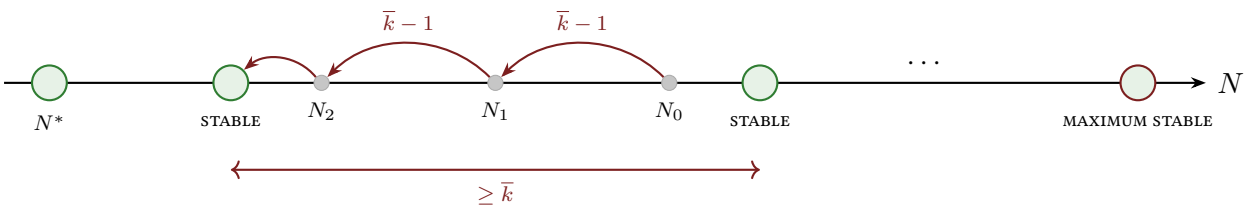


Starting from N^*, stable population sizes recur every $\overline{k}$ agents, ad infinitum. To see why, consider an initial population size $N^*+\overline{k}$. Suppose that in period 1, the principal targets $k \in \{1, \ldots, \overline{k}-1\}$ agents. If nobody protests, the population falls to $N^* + \overline{k} - k$. Then in period 2, the principal would eliminate the next $\overline{k}-k$ agents without triggering a protest, as untargeted agents correctly anticipate no further targeting once the stable population size N^* is reached. But then these $\overline{k}-k$ agents, who could identify themselves in period 1 under predictable-order targeting, would have preferred to join the period 1 targets in a potentially successful protest. To avoid protest, the principal stops in period 1, rendering $N^* + \overline{k}$ stable. For any integer $m \geq 1$, a similar argument applies to each population size $N^* + (m+1)\bar{k}$, given that $N^* + m\bar{k}$ is a stable population size. The location of stable population sizes under predictable-order targeting thus implies that the principal eliminates at most $\overline{k} - 1$ agents in the first period and then stops elimination at the largest stable population size below the initial population size.

Whether a similar structure can be recovered under unpredictable-order targeting is, ex-ante,

unclear. The principal's ideal population size N^* is stable, but recursively constructing further stable sizes requires characterizing the path of eliminations, which is more challenging than in the predictable-order case. Suppose we already established that if the principal eliminates any agents today, she will be able to continue eliminating until reaching N^*, and that targeting one agent today would trigger a potentially successful protest. Would targeting $m > 1$ also do so? Unlike under predictable-order targeting, the answer depends on a comparison between two opposing effects. Targeting fewer agents delays elimination of future targets, *discouraging* protest, just like in the predictable-order case. But unlike in the predictable-order case, targeting fewer agents increases each untargeted agent's chance of ultimately being eliminated, *encouraging* protest. We exploit the quasi-convexity of (a lower bound on) untargeted agents' continuation values to compare these forces and characterize equilibria. As in the predictable-order case, the principal stops at the largest stable size below the initial population size (hereafter, the *final stable size*). Unlike the predictable-order case, however, this sequence is finite, with consecutive sizes separated by *at least* $\overline{k}$ agents and strictly larger gaps at sufficiently large population sizes.

On the path to the final stable size, the principal follows a *greedy* strategy, eliminating $\overline{k} - 1$ agents in all but at most two periods. Slower-than-greedy elimination reduces future targets' incentives to protest and can be optimal if the principal can commit to an elimination path (Online Appendix B). However, without commitment, two forces limit the scope for slower-than-greedy elimination. First, eliminating more agents decreases each untargeted agent's chance of ultimately being eliminated. Near the final stable size, this effect dominates, so the principal eliminates greedily. Second, in large populations, untargeted agents face a low near-term risk of elimination, so slow elimination is unnecessary to discourage protest. We show that genuine delay occurs at most once in equilibrium, at an intermediate population size.[1] We illustrate below a multi-period elimination path, from N_0, to N_1, to N_2, to a stable population size.



[1] In unpredictable-order equilibria, we show that the principal's equilibrium payoff is unique after every history. Outside this class, history-dependent punishments can support a broader set of equilibrium values; see Section 7.

Longer purges (more *periods*) and deeper purges (more *agents*) become feasible under unpredictable order targeting. We characterize when consecutive stable population sizes are spaced strictly more than $\overline{k}$ apart (Proposition 4), and show that the set of stable population sizes scales approximately linearly in $\overline{k}$ (Proposition 5). The differences we identify between predictable and unpredictable-order targeting imply that unpredictable-order targeting typically yields a higher payoff for the principal, unambiguously so for sufficiently large initial populations (Proposition 4).

The broader lesson is that unpredictability can sustain a dynamic form of divide-and-conquer even when agents can coordinate profitable joint deviations. The principal cannot commit not to eliminate untargeted agents after current targets are eliminated, and target predictability determines whether the near-term future targets can identify themselves. This provides a lens through which to understand why authoritarian purges appear arbitrary even when rulers have underlying priorities, and why insiders fail to organize resistance.

We illustrate the mechanism through several historical episodes. During Stalin's purge of the Red Army officer corps, senior officers understood that loyalty and participation in past repression offered little protection, yet future targets remained difficult to predict and effective resistance failed to emerge (Zakharov and Sonin, 2024, Montefiore, 2003). Saddam Hussein's 1979 single-day purge of the Ba'ath Party leadership provides a compressed illustration of the same logic: names were disclosed sequentially while officials not yet named remained compliant (Matthews, 2018). The fall of Robespierre and the events of 9 Thermidor illustrate the cost of mismanaging expectations. Once a sufficiently large set of revolutionary leaders believed they might be targeted soon, they coordinated against Robespierre (Linton, 2013).

The remainder of the paper is organized as follows. Section 2 discusses the related literature. Section 3 presents the model and solution concept. Section 4 characterizes equilibrium under predictable and unpredictable-order targeting and compares their implications for the share of agents purged and the principal's payoff. Section 5 relates the mechanism to historical purges. Section 6 provides the heterogeneous-priority foundation for the two benchmark environments. Section 7 discusses extensions, and Section 8 concludes.

2 Literature

Our paper relates to several strands of research in political economy. One strand studies repression and elite purges in authoritarian regimes. Tyson (2018) analyzes the agency problem between an autocrat and her repressive apparatus. Montagnes and Wolton (2019) study mass purges as a device to create accountability among subordinates. Zakharov and Sonin (2024) provide evidence from the Great Terror and distinguish between informed and preventive purges. By contrast, we abstract from agency frictions within the regime and from the ruler's information about potential coup plotters, focusing instead on how uncertainty about future purge targets shapes protest incentives among elites and the endogenous path of eliminations pursued by the ruler.

Closely related is Che et al. (2026), who study divide-and-conquer when potential victims are uncertain about the ultimate scope of an aggressor's campaign. As in our model, uncertainty weakens coordination among potential victims, but the object of uncertainty differs: they hold the aggressor's ranking of victims fixed and study uncertainty about how many will ultimately be targeted, whereas we hold the desired number of targets fixed and commonly known and study uncertainty about the targeting order. Also related is the analysis in Zeilberger (2024), which considers a version of De Mesquita et al. (2005) in which elites face uncertainty about their future inclusion in the ruling coalition. There, uncertainty plays a fundamentally different role: it makes it harder for the leader to maintain elite support because elites are uncertain about whether they will remain in the coalition. As a result, uncertainty harms, rather than benefits, the principal.

Our paper is also closely related to the literature on coalition formation and bargaining with an endogenous status quo, including Bernheim et al. (2006), Acemoglu et al. (2008), Diermeier and Fong (2009), Acemoglu et al. (2012), Diermeier and Fong (2011), Diermeier et al. (2017), Ali et al. (2019), Iaryczower and Oliveros (2023), and Ali et al. (2023). In these models, forward-looking agents internalize how current redistribution or elimination decisions affect their future bargaining positions, which disciplines equilibrium proposals. Acemoglu et al. (2008) study a dynamic environment in which a ruling coalition is formed through supermajority-approved elimination proposals. They show that this process converges in a single transition to a unique self-enforcing coalition that cannot be successfully challenged thereafter. Diermeier and Fong (2011) analyze an infinite-horizon agenda-setting model with an evolving status quo and show that the possibility

of future revision induces non-proposers to protect one another, limiting the proposer's ability to form minimal winning coalitions and yielding more egalitarian outcomes. Ali et al. (2023) consider a finite-horizon agenda-setting model with an evolving status quo and give conditions under which the agenda-setter achieves her preferred policy even with sophisticated voters.[2]

Relative to this literature, our main contribution is to show how uncertainty about the targeting order shapes the elimination path by weakening agents' forward-looking incentives to coordinate on blocking coalitions. When the order is known, elimination collapses to a single period, echoing Acemoglu et al. (2008). When the order is unpredictable, longer elimination chains can be sustained. We thereby provide a novel understanding of how randomness can be strategically valuable for the principal and connect the mechanism to historical episodes of repression.

Our paper is also connected to the literature studying divide and conquer, such as Abreu and Matsushima (1992), Spiegler (2000), Segal and Whinston (2000), Segal (2003),Winter (2004), Genicot and Ray (2006), Dal Bó (2007), Eliaz and Spiegler (2015), Chen and Zápal (2022), Kapon et al. (2024), Chan (2025), Halac et al. (2020, 2021, 2026). In particular, our comparison between predictable and unpredictable order equilibria highlights that a coarse form of divide-and-conquer, in which the targeted group is singled out for immediate elimination while the untargeted agents hold identical beliefs about when they will be targeted, can dominate the finer strategy of differentiating all agents. In the Online Appendix, we study the principal's commitment problem, allowing her to choose both the elimination path and a public disclosure rule. There, we find that an optimal public disclosure rule ensures that most untargeted agents are kept uniformly uncertain about whether and when the principal will target them.

Our focus on the effects of randomization also connects the paper to work on mixed-strategy equilibria in bargaining, such as Kalandrakis (2004) and Diermeier and Fong (2009). In Diermeier and Fong (2009), mixing over future coalition partners allows the agenda-setter to reach her most preferred policy within three periods, whereas pure-strategy equilibria may severely constrain her. Beyond differences in substantive focus, this mixing differs fundamentally from our randomization, both in its basic mechanics and in the extreme proposal power it confers.[3]

[2] At the heart of the difference between our predictable-order elimination environment and Ali et al. (2023) is that their principal retains proposal power even after a failed proposal, whereas our principal cannot eliminate agents after a successful protest. Appendix A discusses this connection further.

[3] See Appendix Section A for a more detailed discussion of the connection.

Our paper also relates to work on dynamic collective action. Shadmehr and Bernhardt (2019) study how revolutionary vanguards shape participation in collective action, while Gieczewski and Kocak (2024) show that expectations of future protest can reduce citizens' willingness to protest today. Our mechanism highlights a complementary force: predictable-order targeting strengthens protest by identifying the agents with the greatest incentives to join current targets, while unpredictability weakens protest by diffusing the perceived risk of future targeting.

Finally, in our dynamic setting, our coalition-proof refinement is a natural analogue of the perfect coalitional equilibrium (PCE) introduced by Ali and Liu (2025) for repeated games.[4]

3 Model

We study a dynamic game in which a principal seeks to eliminate a subset of agents, whose only protection is the ability to organize a protest that can permanently halt further eliminations.

Time, players, and state. Time is discrete, with an infinite horizon and periods indexed by $t \in \{0, 1, \dots\}$. Initially, there is a *principal* and a finite set of homogeneous agents $\mathcal{N}_0 = \{1, ..., N_0\}$. For each period t, let $\mathcal{N}_t \subseteq \mathcal{N}_0$ denote the set of surviving agents, i.e., those not previously eliminated, and let $N_t \equiv |\mathcal{N}_t|$ denote its size. Let $\Omega_t \in \{0, 1\}$, with $\Omega_0 = 0$, indicate whether a successful protest has occurred before period t: $\Omega_t = 1$ if so and $\Omega_t = 0$ otherwise. All players are risk-neutral and share discount factor $\delta \in (0, 1)$.

Agents receive flow payoff 1 while alive and 0 after elimination. With n survivors, the principal's period-t flow payoff is $u(n)$ when $\Omega_t = 0$ (i.e., when still in power) and 0 when $\Omega_t = 1$, where $u : \{0, \dots, N_0\} \to \mathbb{R}_+$ is strictly single-peaked at N^*:

$$u(n+1) > u(n) \quad \text{for } n < N^*, \qquad u(n+1) < u(n) \quad \text{for } n \geq N^*.$$

This payoff structure captures a trade-off between consolidating power and retaining a sufficiently large base from which to govern.[5]

Actions and state transitions. At the start of each period t with $\Omega_t = 0$, the principal chooses a set $\mathcal{K}_t \subseteq \mathcal{N}_t$ of agents to *target* for elimination, with $k_t \equiv |\mathcal{K}_t|$. After observing $\mathcal{K}_t$, each agent

[4] See the Online Appendix for details.

[5] For instance, suppose a government functioning with n elites produces a pie of size $A + \alpha \times \min\{0, n - N^*\}$ per period, with $\alpha > 1$, $A > \alpha N_0$, and $N^* \geq 0$ denoting the minimum efficient government size. Also, suppose utilities are linear in consumption and that each surviving elite gets 1 per period while the leader gets the remainder. Then the leader's payoff is $u(n) = A + \alpha \times \min\{0, n - N^*\} - n$, which is strictly single-peaked at N^*.

$i \in \mathcal{N}_t$ decides whether to protest, $a_i \in \{0,1\}$, with action $a_i = 1$ denoting *protest* and $a_i = 0$ denoting *no protest*. Let $\mathcal{P}_t \subseteq \mathcal{N}_t$ be the set of protesters in period t, and let $p_t \equiv |\mathcal{P}_t|$ be its size.

A protest in period t is *potentially successful* if its size is at least $\overline{k}$, i.e., $p_t \geq \overline{k}$, where $\overline{k} \in \{2,3,\dots\}$. Denote by $\hat{\mathcal{P}}_t$ the set of its participants. Such a protest succeeds if and only if an exogenous binary state $\omega_t \in \{0,1\}$, realized after actions are taken, equals 1; otherwise it fails. The states are independent across periods, with $\Pr(\omega_t = 1) = \rho \in (0,1)$.[6]

If the protest succeeds, the game proceeds to period $t+1$ with

$$\mathcal{N}_{t+1} = \mathcal{N}_t \qquad \text{and} \qquad \Omega_{t+1} = 1.$$

If no protest succeeds, all targeted agents and all participants in a potentially successful protest are permanently eliminated, so [7]

$$\mathcal{N}_{t+1} = \mathcal{N}_t \setminus (\mathcal{K}_t \cup \hat{\mathcal{P}}_t) \qquad \text{and} \qquad \Omega_{t+1} = 0.$$

In each period t such that $\Omega_t = 1$, the principal cannot target any agent and agents do not protest. Accordingly, we set $\mathcal{K}_t = \emptyset$, $\mathcal{P}_t = \emptyset$, $\Omega_{t+1} = \Omega_t = 1$, and $\mathcal{N}_{t+1} = \mathcal{N}_t$. If $\mathcal{K}_t = \emptyset$, we assume that protest does not occur, so $\mathcal{P}_t = \emptyset$.[8]

Histories. Given the initial set of agents $\mathcal{N}_0$, a *history* up to period t is $h^t \equiv (\mathcal{K}_s, \mathcal{P}_s, \omega_s)_{s=0}^{t-1}$. Denote by H the set of all histories. Each history $h^t \in H$ induces a state $(\mathcal{N}_t(h^t), \Omega_t(h^t))$, where

$$\Omega_t(h^t) = \max_{0 \leq s < t} \{\mathbf{1}\{p_s \geq \overline{k}\}\, \omega_s\} \qquad \textit{and} \qquad \mathcal{N}_t = \mathcal{N}_0 \setminus \bigcup_{0 \leq s < t : \Omega_{s+1} = 0} (\mathcal{K}_s \cup \hat{\mathcal{P}}_s).$$

We suppress the dependence of the state on h^t when it causes no confusion.

A *history with proposal* (i.e., after the principal's move at t) is $\ell^t \equiv \left(h^t, \mathcal{K}_t\right)$. Denote by L the set of all histories with proposal.

Strategies. A (behavioral) strategy for the principal is a mapping $\sigma^p : H \to \Delta\left(2^{\mathcal{N}_0}\right)$ such that, after any history $h^t \in H$, the support of $\sigma^p(h^t)$ is contained in $2^{\mathcal{N}_t(h^t)}$, and if $\Omega_t(h^t) = 1$, $\sigma^p(h^t)$

[6] In our setting, participation is costly only because $\rho < 1$ creates a risk of elimination following a failed protest. A direct protest cost would play a similar role.

[7] Consider an alternative in which all participants in a failed protest are eliminated, regardless of its size. This specification is equivalent to ours if protesters may withdraw when participation falls below $\overline{k}$, or if protest decisions are sequential. However, without withdrawal and with simultaneous protest decisions, some proposal subgames may lack a pure-strategy coalition-proof equilibrium as defined in this paper; our specification avoids this technical issue. Our results would also be unchanged if untargeted protesters are not eliminated but instead pay a fixed protest cost.

[8] Allowing agents to protest when no elimination is proposed would not affect the analysis.

assigns probability 1 to $\emptyset$. A (behavioral) strategy for agent i is a mapping $\sigma^i : L \to \Delta(\{0,1\})$, with the convention that $\sigma^i(\ell^t) = 0$ whenever $i \notin \mathcal{N}_t$ or $\Omega_t = 1$ or $\mathcal{K}_t = \emptyset$.[9]

Continuation payoffs. Fix a strategy profile $\sigma = (\sigma^p, (\sigma^i)_{i \in \mathcal{N}_0})$. Let $\mathbb{E}_\sigma$ denote expectation with respect to the distribution of the process $(\mathcal{N}_t, \Omega_t)_{t \geq 0}$ induced by σ and $(\omega_t)_{t \geq 0}$. The principal's continuation payoff after history h^t is

$$U^p(\sigma \mid h^t) = \mathbb{E}_\sigma \Big[\sum_{s=t}^{\infty} \delta^{s-t} \mathbf{1}\{\Omega_s = 0\} u(N_s) \;\Big|\; h^t \Big].$$

For agent i, the continuation payoff after history with proposal ℓ^t is

$$U^i(\sigma \mid \ell^t) = \mathbb{E}_\sigma \Big[\sum_{s=t}^{\infty} \delta^{s-t} \mathbf{1}\{i \in \mathcal{N}_s\} \;\Big|\; \ell^t \Big].$$

We impose two assumptions.

Assumption 1. *Agents strictly prefer participating in a protest that succeeds with probability ρ, to waiting one period and then being eliminated for sure, i.e.,* $1 < \frac{\rho}{1-\delta}$.

Assumption 1 ensures that the payoff of being eliminated next period, $1+\delta$, is below $1 + \frac{\rho\delta}{1-\delta}$, the expected payoff from joining a potentially successful protest immediately: the protest succeeds with probability ρ, yielding survival thereafter, and fails with probability $1-\rho$, causing immediate elimination. If Assumption 1 fails, the model is uninteresting, as only targeted agents would have an incentive to protest.

Assumption 2. *The principal strictly prefers to remain forever at the initial population size than to choose any action that induces a protest of size $p_t \geq \overline{k}$, i.e.,* $u(N_0) > (1-\rho)u(N^*)$.

Assumption 2 ensures that the principal prefers not eliminating any agents to triggering a potentially successful protest. It captures settings in which the consequences of a successful protest are severe, for example because the ruler risks being overthrown or killed.

Sequential optimality and solution concept. Fix a strategy profile $\sigma = (\sigma^p, (\sigma^i)_{i \in \mathcal{N}_0})$. The principal's strategy σ^p is *sequentially optimal* if, for every history h^t with induced state $(\mathcal{N}_t, \Omega_t)$, and for every alternative strategy $\tilde{\sigma}^p$ that agrees with σ^p up to $t-1$,

$$U^p(\sigma^p, \sigma^{-p} \mid h^t) \;\geq\; U^p(\tilde{\sigma}^p, \sigma^{-p} \mid h^t).$$

[9] With abuse of notation, we denote by a_i the mixed action that assigns probability 1 to $a_i \in \{0,1\}$

Similarly, agent i's strategy σ^i is *sequentially optimal* if, for every history with proposal ℓ^t such that $i \in \mathcal{N}_t$, and every alternative strategy $\tilde{\sigma}^i$ that agrees with σ^i up to $t-1$,

$$U^i(\sigma^i, \sigma^{-i} \mid \ell^t) \;\geq\; U^i(\tilde{\sigma}^i, \sigma^{-i} \mid \ell^t).$$

A strategy profile σ is a *subgame perfect equilibrium (SPE)* if every player's strategy is sequentially optimal after every history at which that player moves.

Coalition-proof refinement. We refine SPE by imposing robustness to joint one-shot deviations by coalitions of agents.

Definition 1. *A strategy profile σ is a* coalition-proof SPE *if it is an SPE and satisfies the following:*

1. *there exists no history with proposal ℓ^t and nonempty coalition $C \subseteq \mathcal{N}_t$ that has a joint pure one-shot deviation at ℓ^t that strictly increases the payoff of every $i \in C$;*
2. *for every agent i and history with proposal ℓ^t such that $i \in \mathcal{N}_t$, if agent i is indifferent between protesting and not protesting, then $\sigma^i(\ell^t)$ assigns probability 1 to not protest.*

The second condition only resolves nongeneric indifferences and is incorporated into the definition to avoid additional notation.[10]

We use coalition-proofness to abstract from coordination failures at the protest stage, which are important but distinct from the mechanism we seek to isolate. To see the issue, suppose that $N_t - N^* \geq \overline{k}$. There is always an SPE in which the principal eliminates all $N_t - N^*$ agents immediately. This is because no agent can reach the participation threshold unilaterally if they expect no one else to protest. Coalition-proofness eliminates such equilibria. [11]

To study how uncertainty about future purge targets shapes protest incentives and the endogenous path of eliminations, we focus on two classes of coalition-proof SPE:[12]

- A *predictable-order coalition-proof SPE* is a coalition-proof SPE for which there exists a strict total order $\succ$ on $\mathcal{N}_0$ such that, after every history with $\Omega_t = 0$, conditional on targeting k_t agents, the principal targets the k_t highest-ranked surviving agents according to $\succ$.

[10] Mixed strategies for agents are ruled out except for a nongeneric set of parameters even without the tie-breaking rule, by the possibility of joint deviations.

[11] Our results would not change if we used the concept of Markov Voting Equilibrium in Acemoglu et al. (2015), though the concept also imposes Markovian strategies, which we derive as a result. Our results would also be unchanged if, after the principal proposes the set of agents to eliminate, agents make their protest decisions sequentially.

[12] See Section 7 for a brief discussion of the set of all coalition-proof SPE that can arise.

- An *unpredictable-order coalition-proof SPE* is a coalition-proof SPE such that, after every history with $\Omega_t = 0$, conditional on targeting k_t agents, the principal chooses uniformly at random among the k_t-size subsets of $\mathcal{N}_t$.

To avoid confusion, we emphasize that these are both classes of *coalition-proof SPE*: we do not restrict the deviations available to the principal.

The two classes capture how agents perceive their near-term elimination risk: an unpredictable order diffuses it across all untargeted agents, while a predictable order concentrates it on a subset of them. Section 5 relates these classes to several historical elite purges.

Foundation for the equilibrium classes. Section 6 provides a microfoundation for the two equilibrium classes in a richer environment where agents differ in the cost they impose on the principal. The principal targets more costly agents first in both cases, but which outcome emerges depends on agents' information. When these costs are common knowledge, the predictable-order outcome emerges; when only the principal observes them and all untargeted agents perceive the same elimination risk, the unpredictable-order outcome emerges. Thus, the two equilibrium classes isolate the effect of transparency about future targets: what matters is not whether the principal literally randomizes or follows a fixed order, but how well agents can predict who is next. We would expect predictable-order outcomes when the leader's priorities can be inferred from easily observed characteristics like faction, rank, or past behavior, and the unpredictable-order outcomes when her ranking depends on opaque criteria that only the leader knows.

4 Coalition-Proof Subgame Perfect Equilibria

4.1 Predictable-order targeting

Under predictable-order elimination, agents who expect to be targeted next join current protesters. As a result, the purge collapses to a single period, and eliminates at most $\overline{k} - 1$ agents.

Fix a priority order $\succ$ and let $E_t := \max\{0, N_t - N^*\}$ be called the *excess population* in period t. For each nonnegative integer E, let $r(E) := E \bmod \overline{k} \in \{0, \ldots, \overline{k} - 1\}$ denote the remainder when E is divided by $\overline{k}$, and set $r_t := r(E_t)$.

Proposition 1 (Predictable-order targeting). *Under Assumptions 1 and 2, in every predictable-order coalition-proof SPE with priority order $\succ$, the principal targets the r_0 highest-ranked agents in period 0 and then stops permanently. No protest succeeds, and $N_t = N_0 - r_0$ for any $t \geq 1$.*

Thus, the purge lasts at most one period, and is limited to eliminating no more than $\overline{k} - 1$ agents, independently of the initial population size. This result echoes the one-step transition in Acemoglu et al. (2008), with agents blocking the principal because they anticipate becoming future targets. As we show next, this conclusion fails under unpredictable-order elimination.

Sketch of the proof. Predictability allows next-period targets (if any) to identify themselves, giving them an incentive to join current targets in protest and constraining the principal. We use this idea in a proof by induction on the excess population size, denoted E, showing that the principal targets the $r(E)$ highest-ranked survivors and then stops permanently. If $E = 0$, the principal stops, as any further elimination would lower her payoff. If $1 \leq E < \overline{k}$, she eliminates all $r(E) = E$ excess agents and reaches her ideal size N^* in one period: the targets are too few to mount a potentially successful protest alone, while untargeted agents anticipate no further elimination and therefore prefer not to join.

Next suppose that $E = q\overline{k} + r$ for some $q \geq 1$, where $r := r(E)$, and that the claim holds for every smaller excess population. Targeting $m > r$ agents induces a potentially successful protest, either by the targets alone when $m \geq \overline{k}$, or with the participation of untargeted agents when $m \in \{r+1, \ldots, \overline{k} - 1\}$. Indeed, in the latter case, the induction hypothesis implies that the $r(E-m) = \overline{k} + r - m$ highest-ranked untargeted agents would be eliminated next period if the current targets were eliminated. If fewer than $\overline{k}$ agents protested, Assumption 1 implies that a coalition of current and prospective targets could profitably deviate by joining it and bringing participation to $\overline{k}$. Coalition-proofness therefore implies that the proposal induces a potentially successful protest. Assumption 2 thus makes every target count $m > r$ suboptimal. If $r = 0$, the principal therefore stops. If instead $r > 0$, targeting exactly r agents induces no such protest because $E - r = q\overline{k} < E$ has remainder zero, so the induction hypothesis implies no further elimination. Untargeted agents therefore do not join, while the $r < \overline{k}$ targets alone cannot mount a potentially successful protest. Targeting fewer agents only delays reaching $E = q\overline{k}$ agents and is strictly worse because u is strictly decreasing above N^*. Hence, the principal targets exactly

the $r(E)$ highest-ranked surviving agents and then stops permanently.

Example 1. *Consider $N^* = 2$ and $\bar{k} = 3$. For $N_0 \in \{3, 4\}$, $E_0 = N_0 - 2 \in \{1, 2\}$ and $r(E_0) = E_0$, so the principal eliminates one agent if $N_0 = 3$ and two if $N_0 = 4$, leaving $N^* = 2$ survivors. At $N_0 = 5$, $E_0 = 3$ and $r(E_0) = 0$, so the principal eliminates no agents. To see why, note that eliminating $m \in \{1, 2\}$ agents would leave $3 - m$ excess agents who know they would be eliminated next period and so prefer to protest today rather than letting the purge proceed. Thus, by Assumption 1 and coalition-proofness, a potentially successful protest of size $m + (3 - m) = 3$ forms if the principal targets $m \in \{1, 2\}$ agents, making such targeting suboptimal by Assumption 2. Similarly, the principal would not target 3 or more agents, since targets alone would mount a potentially successful protest. At $N_0 \in \{6, 7\}$, $E_0 = N_0 - 2 \in \{4, 5\}$ and $r(E_0) = E_0 - 3 \in \{1, 2\}$, so the principal eliminates one agent if $N_0 = 6$ and two if $N_0 = 7$, reaching $N_1 = 5$ in one step and then stopping. The same argument implies no elimination at $N_0 = 8$, and the pattern repeats.*

As we show next, under unpredictable-order targeting the purge can last for multiple periods, and its depth can grow with the initial population.

4.2 Unpredictable-order targeting

When future targets are unpredictable, the risk of near-term targeting is spread uniformly across all untargeted agents, weakening their incentives to join current targets in protest. This can sustain multi-period purges and eliminate many more agents than predictable-order targeting.

Because current targets always benefit from a potentially successful protest, the key question is whether untargeted agents join them. Their decision depends on the probability and timing of future elimination if the purge continues. Both are endogenous, depending on the principal's continuation strategy and on whether further targeting induces a potentially successful protest.

First, we show that (i) every unpredictable-order coalition-proof SPE yields the same path of target counts, (ii) equilibrium target counts and agents' continuation values depend only on the current population size, and (iii) the principal's continuation payoff strictly decreases in population size above N^* (Lemma 1 in the appendix). Combined with Assumption 2, this implies that the principal targets as many agents as possible without inducing a potentially successful protest.

Let $\kappa(n)$ denote the equilibrium target count with n survivors, and let $V(n)$ denote an agent's ex ante continuation payoff before the principal chooses the target count. As under predictable-

order targeting, the principal targets $\kappa(n) = 0$ when $n \leq N^*$ and $\kappa(n) = n - N^*$ when $0 < n - N^* < \overline{k}$, without inducing any protest.

Suppose instead that $n \geq N^* + \overline{k}$. Targeting $m \geq \overline{k}$ agents induces a potentially successful protest by the targets alone, so consider targeting $m < \overline{k}$ agents. If the principal's proposed elimination is implemented without a potentially successful protest, an untargeted agent obtains $1 + \delta V(n-m)$, whereas joining a potentially successful protest yields $1 + \delta\rho/(1-\delta)$. Coalition-proofness and the tie-breaking rule therefore imply that the proposal avoids a potentially successful protest if and only if

$$1 + \delta V(n-m) \geq 1 + \delta \frac{\rho}{1-\delta}$$

If this inequality fails, a coalition of $\overline{k}$ targeted and untargeted agents would strictly benefit from protesting. Thus, for all $n \geq N^* + \overline{k}$ the principal optimally targets $\kappa(n) = 0$ agents if $V(n-m) < \frac{\rho}{1-\delta}$ for all $m \in \{1, \ldots, \overline{k} - 1\}$ and otherwise targets

$$\kappa(n) = \max\left\{ m \in \{1, \ldots, \overline{k} - 1\} : V(n-m) \geq \frac{\rho}{1-\delta} \right\}.$$

We now describe our main result, which we break into two propositions to aid readability.

Proposition 2 (Blocking population sizes)**.** *Under Assumptions 1 and 2, all unpredictable-order coalition-proof SPE induce the same path of target counts and population sizes $(k_t, N_t)_{t \geq 0}$. Moreover, there exists a unique finite set $\mathcal{B} = \{B_0, B_1, \ldots, B_n\} \subseteq \mathbb{N}$, independent of u and N_0, such that $N^* = B_0 < B_1 < \cdots < B_n$, and the following holds. For any initial population size $N_0 \in \mathbb{N}$, the equilibrium path reaches population size*

$$B^*_{N_0} := \begin{cases} N_0, & \text{if } N_0 < N^*, \\ \max\{B_j \in \mathcal{B} : B_j \leq N_0\}, & \text{if } N_0 \geq N^*. \end{cases}$$

in finite time, after which no further eliminations occur. Along the equilibrium path, no untargeted agent protests, and no protest succeeds. In every equilibrium, the principal's strategy is Markovian in $(\mathcal{N}_t, \mathcal{K}_t)$.

A population size $N_t = B$ is terminal either because the principal has reached or passed her ideal point, $B \leq N^*$, or because any further elimination would lower the continuation value of untargeted agents enough to induce them to join the current targets in a potentially successful

protest. In the latter case, the principal refrains from targeting once the population reaches B to avoid this risk. Crucially, if targeting k agents at $N_t = B$ induces a protest of size $\overline{k}$, then targeting $k + m$ agents at $N_t = B + m$ does so as well: both proposals leave $B - k$ agents after the current elimination and hence generate the same continuation value for untargeted agents. Thus, from any $N_t \geq B$, the principal can never eliminate more than $N_t - B$ agents without triggering a potentially successful protest.

The set $\mathcal{B}$ consists of the ideal population size N^* and precisely these endogenous blocking sizes above it, at which any further targeting would raise future elimination risk enough to induce untargeted agents to join the current targets in protest. From any initial population $N_0 \geq N^*$, the principal eliminates agents until reaching the largest $B \in \mathcal{B}$ such that $B \leq N_0$, then stops permanently. Therefore, diffusion weakens resistance but does not eliminate it.

Our next result characterizes elimination speed using the following definition.

Definition 2. *We say that the principal uses the* greedy *elimination strategy in period t if she targets $\overline{k} - 1$ agents.*

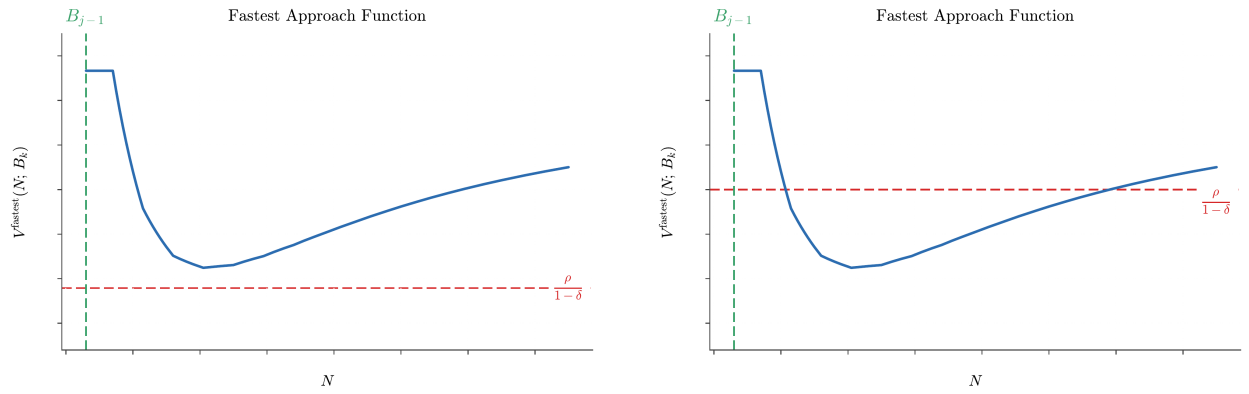
The greedy strategy is thus the fastest targeting pace that prevents the targets alone from mounting a potentially successful protest.

Proposition 3 (Minimal delay)**.** *Along the unique unpredictable-order coalition-proof SPE path of target counts and population sizes from N_0 to $B^*_{N_0}$, the principal uses the greedy elimination strategy in every period except at most two.*

Intuitively, a slower pace of elimination could help the principal by raising untargeted agents' continuation values and discouraging them from joining a protest. Two forces limit this motive. First, when the current population is close to the final blocking coalition size $B^*_{N_0}$, eliminating more agents today raises each untargeted agent's probability of belonging to the final coalition, and that probability reaches one exactly when $B^*_{N_0}$ untargeted agents remain. In this region, the continuation value of untargeted agents increases with the number of eliminations, so the principal has no incentive to delay and will target $k_t = \min\{N_t - B^*_{N_0}, \overline{k} - 1\}$ agents. Second, when the population is far above $B^*_{N_0}$, untargeted agents know that they are unlikely to be targeted soon, because the principal never targets more than $\overline{k} - 1$ agents per period, since doing so would trigger a potentially successful protest. Although their continuation value still declines

as elimination proceeds, that decline is too small to induce them to protest, even if they expect elimination to proceed at the maximum speed (i.e., greedily).

Taken together, we show these two forces imply that, under the greedy strategy, the continuation value of untargeted agents when the principal faces no risk of successful protest—which we call the *fastest approach function*—is quasi-convex in N_t along the path from N_0 to $B^*_{N_0}$, reaching its minimum at $n_{min} > B^*_{N_0}$.[13] This quasi-convexity implies that there is at most one contiguous region around n_{min} in which untargeted agents find it optimal to join a protest. However, either this region is small enough that there is at most one period of delay, or it contains another element of $\mathcal{B}$, contradicting the fact that $B^*_{N_0}$ is the final blocking coalition size.



(a) No intersection

(b) Intersection

Figure 1: Fastest approach function (V^{aux}) v.s. value of protesting

Proof sketch and constructing B_j. We sketch here the recursive construction of thresholds B_j, which underlies the proof of both Propositions 2 and 3; these are in fact proved simultaneously. We initialize the construction with $B_0 = N^*$, and suppose that we have found values $(B_0, \ldots, B_{j-1})$ satisfying Propositions 2 and 3. To construct B_j, we compute the fastest approach function as defined in the previous paragraph, and if it never falls below the payoff from protesting, then we set $B_j = \infty$.[14] Otherwise, B_j is the first integer from which every feasible positive elimination would induce a population size at which the fastest approach function falls below

[13] More precisely, we define the fastest approach function from N_0 to $B^*_{N_0}$ as the continuation value of untargeted agents when the principal uses the greedy strategy in every period, except perhaps in the final period, in which the principal uses $(N_0 - B^*_{N_0}) \bmod (\overline{k} - 1)$ to exactly reach $B^*_{N_0}$.

[14] More generally, we also set $B_j = \infty$ when the fastest approach function falls below the protest payoff only over a region that the principal can jump over in a single period.

the payoff from protesting. The fastest approach function in these two main cases is displayed in Figure 1, where we denote the fastest approach function by V^{aux}. It is straightforward to verify that for any $N \in (B_{j-1}, B_j)$, the greedy strategy is optimal for the principal, and hence the fastest approach function indeed describes the continuation value of an untargeted agent. [15]

Example 2. *Consider $\delta = 9/10$, $\rho = 3/4$, $\overline{k} = 3$, and $N^* = 3$. Then $B_0 = 3$; we now describe how to construct B_1. At $N_0 \in \{4, 5\}$, the principal eliminates $N_0 - 3$ agents without protest, so $N_t = 3$ for all $t \geq 1$. At $N_0 = 6$, suppose the principal targets 2 agents. An untargeted agent who is pivotal for forming a protest of size $\overline{k}$ obtains $1 + \frac{\delta}{1-\delta} \times \rho = 7.75$ from protesting and $1 + \delta\left[\frac{1}{1-\delta} \times \frac{3}{4} + \frac{1}{4}\right] \approx 8$ from not protesting. Thus, the principal eliminates 2 agents without protest, leaving $N_1 = 4$; she then eliminates 1 agent, leaving $N_2 = 3$, and stops.*

Next, consider $N_0 = 7$. The principal avoids targeting 3 or more agents, as the targets alone would form a potentially successful protest. Suppose instead that she targets 2 agents. A pivotal untargeted agent again receives 7.75 from protesting and $1 + \delta\left[\frac{1}{1-\delta} \times \frac{3}{5} + \frac{2}{5}\right] \approx 6.8$ from not protesting. Thus, she strictly prefers to protest if pivotal, so targeting 2 agents induces a potentially successful protest and is suboptimal for the principal. The question is thus whether the principal could avert protest by targeting only 1 agent. However, a pivotal untargeted agent who does not protest when 1 agent is targeted receives $1 + \delta\left[\frac{1}{1-\delta}\frac{3}{6} + (1+\delta)\frac{1}{6} + \frac{2}{6}\right] \approx 6.1$, less than when 2 agents are targeted. Because her payoff from joining a potentially successful protest is unchanged, targeting only 1 agent also induces a potentially successful protests. So the principal proposes no elimination, and $B_1 = 7$.

To understand why targeting fewer agents does not avert protest, observe that whether the principal targets 1 or 2 agents, only 3 survive the purge. Reducing the targets from 2 to 1 decreases an untargeted agent's chance of survival. When the population is close to B_0, as it is here, this survival effect dominates, so targeting fewer agents cannot reduce her incentive to protest. For sufficiently large populations, however, targeting more agents decreases an untargeted agent's continuation value. As the proof shows, this region of the value function does not determine the cutoffs in $\mathcal{B}$.

[15] This explanation is valid except for the possible one period of delay mentioned in the previous paragraph. In that case, the fastest approach function falls beneath the payoff to protesting but only for an interval of length smaller than $\overline{k}$. We formalize and deal with this case in the proof. In that case, the fastest approach function serves as a lower bound on the continuation value of an untargeted agent, and the solution in this case produces one of the non-greedy periods of elimination in Proposition 3.

4.3 Comparative statics and the value of unpredictability

Our first comparative static result shows how unpredictability *typically* increases the total number of agents eliminated, which we call *purge depth*, and raises the principal's payoff.

For fixed primitives, Proposition 1 implies that an equivalent notion to the blocking population sizes from unpredictable-order elimination can be defined under predictable-order as

$$\mathcal{B}^F := \left(B_q^F\right)_{q=0}^{\infty} = \left(N^* + q\overline{k}\right)_{q=0}^{\infty}.$$

The depth of a purge starting from any $N_0 \geq N^*$ is the distance between N_0 and the largest blocking size weakly below it.[16]

Because blocking population sizes generally differ across regimes, purge depth has no pointwise ranking: for instance, starting from $N_0 = B_j \in \mathcal{B}$, unpredictable-order targeting yields no elimination, whereas predictable-order targeting typically yields strictly positive purge depth, except when $B_j \in \mathcal{B}^F$ as well. Nevertheless, the spacing between blocking population sizes provides a useful comparison of purge depth across the two regimes.

Under predictable-order elimination, agents are able to form blocking coalitions with their neighbors in the elimination order. Hence, blocking population sizes recur every $\overline{k}$ agents, $B_{q+1}^F - B_q^F = \overline{k}$. This limits the purge depth to at most $\overline{k} - 1$ agents. Under unpredictable-order elimination, instead, the risk of being in the next targeted group is diffuse rather than concentrated among a few untargeted agents, weakening their incentive to join the protest. As a result, consecutive unpredictable-order blocking sizes are at least as far apart as their predictable-order counterparts, allowing for deeper purges and, at least for large N_0, benefiting the principal.

Proposition 4. *Suppose Assumptions 1 and 2 hold. Let $(B_0, ..., B_n)$, with $n < \infty$ and $B_0 = N^*$, denote the sequence of blocking population sizes under unpredictable-order coalition proof SPE in Proposition 2, and, by convention, set $B_{n+1} = \infty$. Then,*

- *For every $j \in \{0, \ldots, n\}$ and every $q \in \mathbb{N}$, $B_{j+1} - B_j \geq \overline{k} = B_{q+1}^F - B_q^F$, with $B_{j+1} - B_j > \overline{k}$ if and only if $B_j \geq \frac{\delta}{1-\rho} - 1$.*
- *There exists $\bar{N}$ independent of N_0 and u such that, if $N_0 > \bar{N}$, the principal's payoff is strictly higher under unpredictable-order coalition-proof SPE than under predictable-order coalition-proof SPE.*

[16] Of course, purge depth under predictable-order elimination coincides with r^* as defined in Section 4.1.

The first part implies that if $B_j \geq \frac{\delta}{1-\rho} - 1$, then $B_j + \overline{k} < B_{j+1}$, so that if $N_0 \in \{B_j + \overline{k}, ..., B_{j+1} - 1\}$, unpredictable-order purge depth is at least $\overline{k}$, whereas predictable-order purge depth is at most $\overline{k}-1$. Except for the possibility that the principal's first elimination is non-greedy or N_0 lies within $\overline{k}-1$ of $B^*_{N_0}$—in which case the payoff ranking is ambiguous—the principal will therefore have a higher payoff under unpredictable-order than predictable-order elimination. The second part of the proposition shows that these two qualifiers can be dropped for N_0 large enough. Thus, unpredictability is typically valuable for the principal, and unambiguously so for large N_0.

The next comparative static examines how the principal's strength, measured by the protest threshold, $\overline{k}$, affects blocking sizes and hence purge depth under unpredictable-order elimination. Let $\mathcal{B}(\overline{k}) = (B_0, B_1, \ldots, B_n)$ denote the tuple of blocking sizes from Proposition 2 for protest threshold $\overline{k}$. Recall that this tuple is independent of u.

Proposition 5. *Assume $N^* = 0$, and Assumption 1 holds. For every $\kappa \geq 1$ and a.e. on $(\rho, \delta) \in (0,1)^2$, there exists $\overline{N}(\kappa, \delta, \rho)$ s.t. if $\overline{k} > \overline{N}(\kappa, \delta, \rho)$, then $\mathcal{B}(\lfloor \kappa \overline{k} \rfloor)$ has the same length as $\mathcal{B}(\overline{k})$, and*

$$\frac{\mathcal{B}(\lfloor \kappa \overline{k} \rfloor)}{\overline{k}} = \kappa \frac{\mathcal{B}(\overline{k})}{\overline{k}} + O\left(\frac{1}{\overline{k}}\right).$$

In words, when $N^* = 0$, scaling up the principal's strength by κ scales up every blocking threshold by approximately κ.[17]

One interesting implication of Proposition 5 is that, even when N_0 is well above $\overline{k}$, *decreasing* $\overline{k}$ (strengthening the agents, since fewer are needed to succeed) can lead to an *increase* in the number of agents the principal is able to eliminate.[18] This echoes results in Diermeier et al. (2017) that show how increasing supermajority requirements may increase expropriation.[19]

5 Uncertainty in authoritarian regimes: examples

Authoritarian regimes—in which purges, enemy elimination, and expropriation take center stage—can be highly unpredictable. During the Great Terror, orchestrated by Stalin in the 1930s, the identities of individual targets were often difficult to predict. Purges of the broad population were implemented *by quota*, without clear criteria for selecting individuals beyond broad categories

[17] A version of the result can be stated for the case $N^* > 0$, but is slightly more cumbersome.

[18] If $N_0 = \overline{k}$, decreasing $\overline{k}$ increases purge depth from 0 to 1, and the statement holds.

[19] The same effect can arise under predictable-order elimination. If $N_0 = 5$ and $\overline{k} = 5$, then $r = 0$ and no elimination occurs; reducing $\overline{k}$ to 4 yields $r = 1$, allowing one elimination.

(Gregory, 2009). Mobutu Sese Seko, president of the Democratic Republic of Congo (Zaire, under Mobutu), was known to regularly shuffle ministers in and out of office, evidently to stop them from amassing independent power, leaving them uncertain about their tenure (Haskin, 2005).[20]

These cases illustrate a general uncertainty that pervades authoritarian regimes, especially as they eliminate opposition. Masha Gessen, writing in the *New York Times* in 2026,[21] crystallizes the uncertainty faced by those living under terror regimes:

> *When we think of the terror regimes of the past, it is tempting to superimpose a logical narrative on them, as though totalitarian leaders had an extermination to-do list and worked their way through it methodically. This, I think, is how most people understand Martin Niemöller's classic poem "First They Came." In reality, though, the people living under those regimes never knew which group of people would be designated an enemy of the state next.*

Our paper places this uncertainty at the center of the analysis, showing how unpredictability regarding the identity of future purge targets undermines resistance. At the same time, historical episodes suggest that mismanaging the expectations of potential purge targets can have severe consequences for authoritarian leaders themselves. During the Reign of Terror of the French Revolution, Maximilien Robespierre, along with other members of the Committee of Public Safety, orchestrated the elimination of (among others) Girondins, Dantonists and Hébertists, groups spanning the ideological spectrum. Robespierre himself was eventually arrested and executed by a coalition that included members of the Committee of Public Safety, who feared their own downfall. Linton (2013) writes about the Thermidorian reaction

> *Most of the members of the Committee of Public Safety along with the Committee of General Security had actually signed the arrest warrant for the Dantonists … Now they feared that such tactics might be used against themselves. Several of them were eyeing Robespierre warily, afraid that he was about to precipitate another purge … and that their own names might be on his list.*

and quoting Pierre Louis Prieur, a member of the Committee of Public Safety,[22]

[20] Haskin (2005) attributes the plunder and stealing of politicians under Mobutu's regime to this uncertainty.
[21] https://www.nytimes.com/2026/01/24/opinion/state-terror-has-arrived.html
[22] Robespierre is commonly believed to have made an error by failing to narrow the scope of the purge effectively

Storm clouds banked up over our heads, and at each moment we came closer to an inevitable crisis. However the work of the Committee did not suffer. In the midst of our anguish, uncertain whether the hour would come that would see us sent before the Revolutionary Tribunal, to go from there to the scaffold, without perhaps having the time to say goodbye to our families and our friends . . . We ended by becoming so accustomed to these inextricable situations, that we pursued our daily task, so that business should not be left pending, as though we had a whole lifetime before us, when it was perfectly likely that we would not see the sun rise tomorrow.

We discuss three episodes that highlight the role of uncertainty on collective action in purges. The 1937–1938 purge of Red Army officers and Saddam Hussein's 1979 Ba'ath Party purge illustrate settings in which elites could not predict who would be targeted next. Through the lens of our model, we interpret this uncertainty as weakening their ability to coordinate. Zimbabwe's 2017 military coup provides a contrast: targeting became predictable at the level of an identifiable group, helping potential targets coordinate.

We make an important caveat before proceeding. These episodes involve at least *two* dimensions of uncertainty: (a) the intended *number* of targets, and (b) *who* will be targeted (our focus). We chose the first two cases because, while the intended number of targets was almost certainly unknown, their identity appears highly uncertain as well.[23]

1937-38 Purges of Officers in the Red Army. Between 1937 and 1938, Stalin purged thousands of officers from the Red Army, including many high ranking officers. While scholars have attributed substantial discretion over the choice of targets to Stalin (and the People's Commis-

publicly announced on 8 Thermidor, leaving too many with the belief that they may be on his list of immediate targets. It is also established that a number of his enemies were either clearly identified by Robespierre or could identify themselves. Linton (2013) writes, "...At the same time he [Robespierre] accused several of his fellow Committee members, more or less openly, of being part of the plot against him. He compounded this extraordinarily dangerous public declaration by refusing to give the names of the men on the list of those whose arrest he sought, leaving many of his terrified listeners to speculate whether they themselves, or their friends, appeared on it..." Thus, it seems plausible that even if Robespierre had named the immediate purge targets, which, according to Saint-Just's undelivered speech, were confined to Jaques-Nicolas Billaud-Varenne and Jean-Marie Collot d'Herbois, other leaders (e.g., Joseph Fouché, Jean-Lambert Tallien who spoke out and coordinated on or immediately after the speech of 8 Thermidor) would have feared they would be next, and that this concentrated fear was indeed a fundamental factor in Robespierre's downfall.

[23] Our model also offers limited insight into purges involving the simultaneous, one-time, executions of opponents caught completely by surprise, such as Hitler's Night of the Long Knives in 1934. On the other hand, if such surprise attacks are repeated and constrained in scale by purge technology, then similar results can be obtained.

sariat for Internal Affairs, NKVD) directly, the archives which may host internal documents revealing the precise nature and scope of Stalin's discretion are unavailable to the public. From the recent analysis in Zakharov and Sonin (2024), however, we learn about the characteristics of those among the 1,844 general-grade army officers who were targeted for arrest and execution (at least 780 of whom were executed) and who, as noted in Zakharov and Sonin (2024) likely presented a substantial risk of collective action and resistance to the purge. As the authors show, being purged is not substantially correlated with one's superior or subordinate being purged, suggesting that it would have been difficult to predict future purge targets based on past purge targets. The authors propose a network based theory to differentiate *informed* purges—the leader knows at least one coup plotter's identity—from *preventive* purge—the leader does not know the identity of any coup plotters, and purges to prevent the formation of a coup plot—and argue that the evidence is more consistent with preventive as opposed to informed purges.

The theory we develop in this paper produces a rationale for why purges whose future targets are difficult to predict, as in a preventive purge in the language of Zakharov and Sonin (2024), more successfully undermine collective action and resistance than those for which future targets are easy to predict, as in an informed purge. It also suggests that the weak empirical predictability of purge targets documented by Zakharov and Sonin (2024) among high-ranking army officers could have been the result of a *deliberate* attempt to maintain uncertainty about future targets, thereby reducing the likelihood of collective resistance. While direct evidence on intent is scarce, some qualitative observations are consistent with this interpretation.

First, there is evidence that high-ranking regime insiders often understood that participation in repression did not guarantee, or even make likely, future safety. As Corps Commander Ivan Belov, a member of the Military Collegium and one of the judges presiding over the 1937 trial of Marshal Tukhachevsky, remarked, "*Tomorrow I'll be put in the same place.*" (Montefiore, 2003, p. 230). Second, this fear was not misplaced, as Belov was arrested in January 1938 and executed in July 1938. Several other members of the tribunal, including marshals Vasily Blyukher and Alexander Yegorov and corps commanders Yakov Alksnis, Pavel Dybenko, and Nikolai Kashirin, were also arrested and executed in subsequent purge waves (Montefiore, 2003, Conquest, 2008).[24]

[24]Of the high-ranked officials involved in the trial only a few, such as Marshal Semyon Budyonny and Army Commander Boris Shaposhnikov, survived (Montefiore, 2003, Conquest, 2008).

Third, despite this fear and despite holding substantial coercive power, these high-ranking insiders did not mount any effective resistance or coup attempt prior to their arrests and execution.[25]

Consistent with our model, this suggests that Stalin successfully maintained insiders' perceived risk of becoming future targets sufficiently small relative to the risks associated with opposing the regime, making continued compliance individually optimal for those not yet targeted.

Saddam Hussein's Ba'athist Purge 1979. In this section, we detail Saddam Hussein's 1979 Ba'ath party purge. In doing this, our goal is to simply establish the existence of a nearly simultaneous purge in which the leader chooses to maintain uncertainty over future purge targets, even within the highly condensed sequence of events that unfolded over a single day. This illustrates one of the core premises of our model: that uncertainty over future purge targets undermines collective action, and a leader would prefer to maintain uncertainty rather than to make the ranking of potential purge targets known ex-ante.

In 1979, Saddam Hussein, having newly ascended to the Iraqi presidency, executed a televised purge of dozens of members of the Ba'ath party leadership in Baghdad's al-Khuld conference center. As described in Matthews (2018),

> *Muhie Mashhadi, visibly broken from torture by the secret police, was pushed onstage by his security handlers and forced to read a "confession" that implicated himself and other members of the RCC in a Syrian plot to take over the Iraqi Ba'ath party. Mashhadi then proceeded to slowly read a prepared list of co-conspirators, many of whom were assembled in the hall. As names were called, each of those named were confronted by men in suits and pulled from the conference hall by Hussein's security forces.*

While it is impossible to know whether those eventually called out by Mashhadi could have mounted any resistance if they were all informed simultaneously of their supposed guilt, nearly a third of the present members of the Ba'ath party were arrested (over 60 members), 21 of whom were then executed. By removing the "guilty" one-by-one, Hussein ensured that no large group who *knew* they would be named could coordinate resistance. While it is certainly true that some

[25]Following Stalin's death, Soviet authorities formally overturned many purge-era convictions and posthumously rehabilitated numerous senior military officers who had been accused of participating in a military conspiracy, including Tukhachevsky, Blyukher, Yegorov, Belov and several other senior officers involved in the Tukhachevsky trial. These rehabilitation decisions established that the charges against them had been unfounded.

would have predicted a higher probability of their own "guilt" than others, the structure of the purge was such that no one was certain until their name was called, at which point they were immediately removed from the room by security forces.

Zimbabwe's 2017 political purge and military coup. Zimbabwe provides a useful contrast: as targeting became more predictable, potential future victims could identify one another and coordinate. The following account of the episode draws throughout on Tendi (2020). In December 2014, Robert Mugabe, then the President of Zimbabwe, purged Vice President Joice Mujuru and several of her supporters from the ruling ZANU PF party. He subsequently marginalized liberation-era officials while elevating younger politicians commonly labelled Generation 40 (G40). By July 2017, he had publicly announced plans to "retire" members of the military leadership; senior officers also feared arrest and prosecution. Because the risk of being purged could be inferred from observable characteristics—military rank and association with the liberation generation—the threatened group was increasingly identifiable (Tendi, 2020, pp. 51–56). The process became more concrete on November 6, 2017, when Mugabe dismissed Vice President Emmerson Mnangagwa, a prominent representative of the liberation-era faction within ZANU PF. Lieutenant General Philip Sibanda, then acting commander of the Zimbabwe Defence Forces, alerted Defence Forces commander Constantino Chiwenga and arranged military assistance for Mnangagwa to leave the country. Although Sibanda and Chiwenga were not close and had served in rival liberation armies, they cooperated. On November 13, after Mugabe refused to meet with the generals, Chiwenga appeared alongside dozens of senior army and air-force officers. He described the ongoing "purging and cleansing process" as "targeting mostly members associated with our liberation history" and warned that the military would intervene. Later that day, according to one participant, Chiwenga told officers commanding key military formations: "If anyone tries to fire or arrest one of us, that is an attack on us all." The officers supported intervention. The following night, the military launched Operation Restore Legacy, placed Mugabe under house arrest, and ultimately ended his thirty-seven-year rule.

In the language of our model, Mnangagwa was the current target, while Mugabe's public threat to the military leadership and his attacks on liberation-era officials identified a small set of future targets. The exact order of future targeting remained unknown, but the risk became concentrated enough for the officials at risk to identify one another and respond jointly. The case

therefore lies between our two polar benchmarks but is closer to predictable-order elimination. Although target predictability did not alone cause the coup, the episode provides evidence for our mechanism: likely future victims recognized each other and coordinated before they could be eliminated separately.[26]

6 Target priorities

In the baseline model, the principal's payoff depends only on the number of surviving agents. Consequently, conditional on eliminating a given number of agents, the principal is indifferent over their identities. In this section, we allow agents to differ in the costs they impose on the principal. We show that the predictable-order and unpredictable-order targeting environments studied above arise naturally depending on whether agents observe these costs.

Environment. For any ideal population size $N^* \geq 0$, normalize u so that $u(n) - u(n-1) \geq 1$ for all $n \in \{1, \ldots, N^*\}$. Assume that, when $\Omega_t = 0$, the principal's per-period payoff is

$$u(N_t) - \sum_{i \in \mathcal{N}_t} v_i,$$

where $(v_i)_{i \in \mathcal{N}_0}$ are i.i.d. draws from a continuous distribution F with full support on $[0, 1]$.[27] Agents with larger v_i are more costly to retain. When $\Omega_t = 1$, the principal's per-period payoff remains zero. The following condition is a convenient analogue of Assumption 2:

Assumption 3. *The principal prefers forgoing elimination to choosing a target set that induces a protest of size $p_t \geq \overline{k}$:*

$$(1 - \rho)u(N^*) < u(N_0) - N_0.$$

Indeed, retaining all agents gives the principal at least $u(N_0) - N_0$ in each future period, whereas choosing a target set that induces such a protest gives her an expected payoff of at most $(1 - \rho)u(N^*)$ in each future period. Assumption 3 therefore imposes the stated preference uniformly over realizations of $(v_i)_{i \in \mathcal{N}_0}$. All other elements of the environment are unchanged.

[26] Here elimination was political, through dismissal, arrest, or prosecution, rather than certain execution. Tendi (2020) also emphasizes the generals' political ambitions, Mugabe's declining authority, attachment to the liberation struggle, and fear of prosecution. He identifies Mugabe's refusal to meet with the generals on November 13 as the proximate cause of the coup.

[27] Together with the normalization of u, the upper bound $v_i \leq 1$ ensures that, absent protest constraints, the principal almost surely wishes to eliminate every excess agent but not to reduce the population below N^*.

6.1 Known priorities

Consider first the case in which the realization $(v_i)_{i\in\mathcal{N}_0}$ is common knowledge. Because F is continuous, ties occur with probability zero. Let $\succ_v$ denote the strict total order induced by the realized values (v_i), so that $i \succ_v j$ if and only if $v_i > v_j$.

When the principal's priorities are common knowledge, equilibrium play reduces to predictable-order elimination with order $\succ_v$. The reason is simple. Holding fixed the number of targets, the principal strictly prefers to target agents with the highest v_i. When $N^* < N_t < N^* + \overline{k}$, she can eliminate all remaining excess agents without triggering a potentially successful protest and therefore eliminates those with the highest v_i. Proceeding by backward induction as in the predictable-order elimination case, this implies that, starting from any initial population with $N_0 > N^*$, the principal eliminates $E_0 \bmod \overline{k}$ agents in the first period and then stops. By the same argument, these are the agents with the highest v_i. Thus under common knowledge of priorities, coalition-proof SPE are predictable-order coalition-proof SPE.

Proposition 6. *Suppose Assumptions 1 and Assumptions 3. Suppose the realization $(v_i)_{i\in\mathcal{N}_0}$ is common knowledge. Then the unique equilibrium path of elimination in coalition-proof SPE coincides with the unique equilibrium path of eliminations under predictable-order coalition-proof SPE with order $\succ_v$.*

6.2 Unknown priorities

We now consider the case in which the principal observes $(v_i)_{i\in\mathcal{N}_0}$ but agents do not. Agents share a common prior under which (v_i) is drawn i.i.d. from F. Because F is continuous, the realizations are almost surely distinct and therefore induce a strict ranking of agents. In particular, for every subset $\mathcal{N}_t \subseteq \mathcal{N}_0$, the surviving agents can be uniquely ordered according to their v_i values. Agents observe the history of targeting decisions and protests but not the realizations (v_i).

This defines a dynamic game with incomplete information. We focus on a refined class of perfect Bayesian equilibria, which we denote by rPBE.

Definition 3. *A* refined perfect Bayesian equilibrium *(rPBE) is a perfect Bayesian equilibrium satisfying the following three additional conditions.*

1. ***Coalition-proofness.*** *At any history with proposal ℓ^t, no coalition $\mathcal{C} \subseteq \mathcal{N}_t$ of agents has a joint one-shot deviation that strictly improves every agent's payoff.*
2. ***Exchangeability of untargeted agents.*** *At any history with proposal $\ell^t = (h^t, \mathcal{K}_t)$, on or off the equilibrium path, the posterior distribution over the vector of types $(v_i)_{i \in \mathcal{N}_t \setminus \mathcal{K}_t}$ is invariant to permutations of identities in $\mathcal{N}_t \setminus \mathcal{K}_t$.*
3. ***Tie-breaking.*** *At any history with proposal $\ell^t = (h^t, \mathcal{K}_t)$, if a surviving agent $i \in \mathcal{N}_t$ is indifferent between protesting and not protesting, then i protests with probability zero.*

The first restriction aligns the analysis with the baseline model and allows us to abstract from coordination issues, focusing instead on the role of uncertainty. The second is a reduced-form unpredictability restriction, ruling out differential inferences among agents who receive the same treatment. In particular, conditional on not being targeted, agents do not infer that one untargeted survivor is more exposed than another. Without the exchangeability condition, proposals could be interpreted as revealing arbitrary information about the principal's ranking, vastly expanding the set of behavior that could be sustained in an equilibrium. The third restriction imposes the same non-protest-favoring tie-breaking rule used in the baseline model.

Proposition 7. *Suppose Assumptions 1 and Assumptions 3. Suppose that only the principal observes $(v_i)_{i \in \mathcal{N}_0}$. Then every rPBE induces the same path of target counts and population sizes $(k_t, N_t)_{t \geq 0}$ as the unique unpredictable-order coalition-proof SPE path in the baseline model with the same initial population N_0. Moreover, no protest succeeds on path and, whenever the principal targets k_t agents in period t, she targets the k_t surviving agents with the highest values of v_i.*

Propositions 6 and 7 clarify the interpretation of the two benchmark environments studied above. Predictable-order elimination arises when the principal's priorities are commonly understood, so future targets are predictable. Unpredictable-order elimination arises when the principal has a determinate priority ranking but, conditional on surviving the current period, untargeted agents remain symmetric in their beliefs about future exposure. What matters for the logic of the model is therefore not whether the principal literally randomizes over targets, but whether agents can correctly anticipate future targets.

This interpretation is natural in many applications. In authoritarian environments, insiders may understand that the principal has priorities, grudges, or private information, without know-

ing about how these map into future purge targets. From the perspective of any one untargeted agent, the risk of being targeted is then diffuse: although the principal is not literally randomizing, future targets cannot identify themselves and the same coordination problem arises as under unpredictable-order elimination. Because unpredictable-order coalition-proof SPE typically lead to higher payoff than predictable-order coalition-proof SPE, we should not expect to observe a leader clarifying their ranking.

7 Discussion

History-dependent threats and equilibrium multiplicity. We focus on predictable-order and unpredictable-order coalition-proof SPE. Outside these classes, a broader set of equilibrium paths and payoffs can be sustained through history-dependent threats. For instance, a coalition-proof SPE in which the principal targets no agent for x periods and thereafter follows the unique unpredictable-order coalition-proof SPE path can be sustained by the threat that any acceleration in targeting triggers a switch to the predictable-order coalition-proof SPE associated with some arbitrary ordering. Indeed, if N_0 is sufficiently large relative to x and δ is sufficiently close to one, the loss from triggering the predictable-order continuation outweighs the immediate benefit from accelerating elimination. As shown in Section 6, however, the predictable and unpredictable order classes are not arbitrary choices. They arise endogenously, depending on whether agents observe the heterogeneous costs they impose on the principal, so the elimination order is pinned down by the information structure rather than wielded as a history-dependent threat.

Elimination technology. In the baseline model, targeted agents can resist by joining the protest before eliminations are carried out. Combined with Assumption 2, this imposes a clear upper bound of $\overline{k} - 1$ on the number of agents the principal can optimally target in any period. In some applications, however, targeted agents may be eliminated before they can react (e.g., assassinations). The timing would then be reversed: the principal first eliminates her chosen targets, after which the remaining agents decide whether to protest. If the principal could eliminate an arbitrary number of agents, the problem would then become trivial: as in our baseline setting with $\overline{k} > N_0$, the principal would eliminate all agents she wishes, reducing the population size to $\min\{N_0, N^*\}$ in one period. A more realistic alternative is that the principal faces

a per-period targeting capacity, $\bar{m} \in \mathbb{N}$, reflecting constraints on personnel, intelligence, transportation, detention, or execution capacity. Under such a capacity constraint, a substantively equivalent version of our results apply, except replacing the maximum elimination size beneath the protest constraint $\overline{k} - 1$ with the technological constraint $\overline{m}$. The main qualitative mechanism remains: predictable-order targeting allows future victims to coordinate before they are targeted, whereas unpredictability diffuses perceived risk of elimination and weakens incentives to protest, although the precise blocking thresholds and protest constraints must be recomputed because current targets no longer participate in the protest.

8 Conclusion

In this paper we have studied a game between a principal and finite group of agents who she seeks to eliminate. We have constructed two classes of equilibria: predictable-order coalition proof SPE and unpredictable-order coalition proof SPE. In the former, agents are eliminated according to a known order, while in the latter, agents are eliminated according to uniform random draws each period. We showed that a unique path of eliminations exists in each case, with the principal typically preferring unpredictable-order elimination. In the unique unpredictable-order coalition-proof SPE path, the principal adopts a greedy elimination strategy before stopping at an endogenously determined threshold. We connected our results to historical purge episodes, arguing that uncertainty over future purge targets was pervasive in these episodes and served to undermine collective resistance.

A number of avenues for future work could prove fruitful. First, our paper has focused on a "non-steady state" environment, in which the principal starts with a population size that is too large, and eventually reaches a population size from which no further eliminations occur. In many authoritarian regimes however, purges recur. One way to study recurrent purges is to introduce replenishment: some number of agents are born each period. In authoritarian politics, leaders often replace powerful ministers with weaker ones. Over time however, weaker ministers become more powerful, posing a coup risk to the leader. As such, the leader may eventually attempt to purge the replacement minister, in which case the type of uncertainty we have studied in this paper could prove useful. We leave to future work the study of such dynamics.

Second, our protest technology is simple—no success below a threshold and fixed probability

of success above a threshold—which allows for tight characterizations but may lead us to ignore some important features of the environments we study. Numerical simulations show that more complex protest technology leads to significantly more complex equilibrium behavior, and future work studying these dynamics may yield new insight.

Third, our model does not allow the principal is to make transfers to the agents. An alternative model might allow the principal to make a transfer to agents in exchange for exiting the game. This alternative model introduces significantly different forces from our model, which we think future work can tackle.

More broadly, novel insights into coalition-formation could emerge from combining our notion of unpredictable-order elimination with the rich heterogeneity present in Acemoglu et al. (2008) and follow-on work. Finally, while substantial empirical work on elite purges exists, there are many historical purge episodes in which targets faced uncertainty over their future inclusion in the purge. Assembling data to quantify this uncertainty is an exciting avenue for future work.

References

Abreu, D. and H. Matsushima (1992): "Virtual implementation in iteratively undominated strategies: complete information," *Econometrica: Journal of the Econometric Society*, 993–1008.

Acemoglu, D., G. Egorov, and K. Sonin (2008): "Coalition formation in non-democracies," *The Review of Economic Studies*, 75, 987–1009.

——— (2012): "Dynamics and stability of constitutions, coalitions, and clubs," *American Economic Review*, 102, 1446–1476.

——— (2015): "Political economy in a changing world," *Journal of political economy*, 123, 1038–1086.

Ali, S. N., B. D. Bernheim, A. W. Bloedel, and S. C. Battilana (2023): "Who controls the agenda controls the legislature," *American Economic Review*, 113, 3090–3128.

Ali, S. N., B. D. Bernheim, and X. Fan (2019): "Predictability and power in legislative bargaining," *The Review of Economic Studies*, 86, 500–525.

Ali, S. N. and C. Liu (2025): "Coalitions in repeated games," *Working Paper*.

Bernheim, B. D., A. Rangel, and L. Rayo (2006): "The power of the last word in legislative policy making," *Econometrica*, 74, 1161–1190.

Chan, L. T. (2025): "Weight-ranked divide-and-conquer contracts," *Theoretical Economics*, 20, 857–882.

Che, Y.-K., J. Huang, and W. Lim (2026): "First They Came for the Others: A Theory of Divide-and-Conquer," Tech. rep.

Chen, Y. and J. Zápal (2022): "Sequential vote buying," *Journal of Economic Theory*, 205, 105529.

Conquest, R. (2008): *The great terror: A reassessment*, Oxford University Press.

Dal Bó, E. (2007): "Bribing voters," *American Journal of Political Science*, 51, 789–803.

De Mesquita, B. B., A. Smith, R. M. Siverson, and J. D. Morrow (2005): *The logic of political survival*, MIT press.

Diermeier, D., G. Egorov, and K. Sonin (2017): "Political economy of redistribution," *Econometrica*, 85, 851–870.

Diermeier, D. and P. Fong (2009): "Endogenous limits on proposal power," Tech. rep., Discussion Paper.

——— (2011): "Legislative bargaining with reconsideration," *The Quarterly Journal of Economics*, 126, 947–985.

Eliaz, K. and R. Spiegler (2015): "X-games," *Games and Economic Behavior*, 89, 93–100.

Genicot, G. and D. Ray (2006): "Contracts and externalities: How things fall apart," *Journal of Economic Theory*, 131, 71–100.

Gieczewski, G. and K. Kocak (2024): "Collective procrastination and protest cycles," *American Journal of Political Science*.

Gregory, P. R. (2009): *Terror by quota: state security from Lenin to Stalin:(an archival study)*, Yale University Press.

Halac, M., I. Kremer, and E. Winter (2020): "Raising capital from heterogeneous investors," *American Economic Review*, 110, 889–921.

Halac, M., E. Lipnowski, and D. Rappoport (2021): "Rank uncertainty in organizations," *American Economic Review*, 111, 757–786.

——— (2026): "Pricing for coordination," Tech. rep., Working Paper.

Haskin, J. M. (2005): *The tragic state of the Congo: From decolonization to dictatorship*, Algora Publishing.

Iaryczower, M. and S. Oliveros (2023): "Collective hold-up," *Theoretical Economics*, 18, 1063–1100.

Kalandrakis, A. (2004): "A three-player dynamic majoritarian bargaining game," *Journal of Economic Theory*, 116, 294–14.

Kapon, S., L. Del Carpio, and S. Chassang (2024): "Using divide-and-conquer to improve tax collection," *The Quarterly Journal of Economics*, 139, 2475–2523.

Linton, M. (2013): *Choosing terror: Virtue, friendship, and authenticity in the French revolution*, OUP Oxford.

Matthews, A. S. (2018): *Conflict among Comrades: Elite Purges and Political Violence in Authoritarian Regimes*, Baton Rouge, LA: Louisiana State University.

Montagnes, B. P. and S. Wolton (2019): "Mass purges: Top-down accountability in autocracy," *American Political Science Review*, 113, 1045–1059.

Montefiore, S. S. (2003): *Stalin: The Court of the Red Tsar*, London: Weidenfeld & Nicolson.

Segal, I. (2003): "Coordination and discrimination in contracting with externalities: Divide and conquer?" *Journal of Economic Theory*, 113, 147–181.

Segal, I. R. and M. D. Whinston (2000): "Naked exclusion: comment," *American Economic Review*, 91, 296–309.

Shadmehr, M. and D. Bernhardt (2019): "Vanguards in revolution," *Games and Economic Behavior*, 115, 146–166.

Spiegler, R. (2000): "Extracting interaction-created surplus," *Games and Economic Behavior*, 30, 142–162.

Tendi, B.-M. (2020): "The Motivations and Dynamics of Zimbabwe's 2017 Military Coup," *African Affairs*, 119, 39–67.

Tyson, S. A. (2018): "The agency problem underlying repression," *The Journal of Politics*, 80, 1297–1310.

Winter, E. (2004): "Incentives and discrimination," *American Economic Review*, 94, 764–773.

Zakharov, A. and K. Sonin (2024): "The Anatomy of the Great Terror: A Quantitative Analysis of the 1937-38 Purges in the Red Army," *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Zeilberger, T. (2024): "The Logic of Political Survival Revisited: Consequences of Elite Uncertainty Under Authoritarian Rule," *arXiv preprint arXiv:2408.01887*.

A Appendix

A.1 Proof of Proposition 1

Proof. Recall that $E_t := \max\{0, N_t - N^*\}$ and $r(E) := E \bmod \overline{k}$. Conditional on a target count m, a predictable-order equilibrium selects the m highest-ranked survivors, although deviations are unrestricted. It therefore suffices to establish the optimal target count.

We prove by induction that, at every history with $\Omega_t = 0$ and $E_t = E$, the principal targets $r(E) < \overline{k}$ agents without inducing a potentially successful protest and then stops permanently.

Base step ($E < \overline{k}$). If $E = 0$, the principal targets $r(E) = 0$ agents: any elimination would move the population farther from N^* and lower her continuation payoff. If instead $1 \leq E < \overline{k}$, targeting all $E = r(E)$ excess agents leaves N^* survivors, after which no further elimination occurs. Anticipating this, untargeted agents are unwilling to bring participation to $\overline{k}$, while the $E < \overline{k}$ targets cannot reach the threshold alone. Thus, the proposal induces no potentially successful protest. Targeting more either results in the principal's removal or reduces the population below N^*, while targeting fewer leaves more than N^* survivors next period, yields a lower next-period flow, and cannot outperform the maximal flow $u(N^*)$ thereafter. Hence, the principal targets all $E = r(E)$ excess agents immediately.

Inductive step. Write $E = q\overline{k} + r$, where $q \geq 1$ and $r := r(E)$, and assume the claim for every smaller value of E. Any target count $m \geq \overline{k}$ induces a potentially successful protest by the targets alone. Now consider $r < m < \overline{k}$. If the induced protest were subthreshold, only the m targets would be eliminated, so the next-period excess population would satisfy

$$E - m = (q-1)\overline{k} + (\overline{k} + r - m) < E, \qquad r(E-m) = \overline{k} + r - m \geq \overline{k} - m.$$

By the induction hypothesis, that many highest-ranked untargeted survivors would be targeted next period. Since $r(E-m) \geq \overline{k} - m$, enough of them can join the m current targets to make the protest potentially successful. Assumption 1 makes this joint deviation strictly profitable, contradicting coalition-proofness. Hence, targeting $m > r$ agents induces a potentially successful protest and is suboptimal by Assumption 2.

If $r = 0$, the principal therefore stops permanently. If instead $r > 0$, targeting exactly r agents leaves $E - r = q\overline{k}$ with zero remainder, so by the induction hypothesis, no further elimination occurs. Untargeted agents are therefore unwilling to join a potentially successful protest, while the $r < \overline{k}$ targets cannot mount one alone. Targets are thus eliminated. Targeting $m \in \{1, \ldots, r-1\}$ agents is either blocked or, by the induction hypothesis, followed next period by targeting another $r - m$ agents and then stopping permanently. The latter path reaches the same terminal population as targeting all r agents immediately, but one period later. Because u is strictly decreasing above N^*, targeting all r agents immediately is strictly better. If the principal targets no agent, the state is unchanged and the same upper bound applies next period; inaction therefore either delays the reduction by r or forgoes it, and is strictly worse. Hence, the principal targets exactly $r(E)$ agents and then stops permanently. By predictable-order targeting, these are the $r(E)$ highest-ranked survivors. Applying the claim at E_0 proves the proposition. □

Proof of Propositions 2 and 3

We prove Propositions 2 and 3 together. We can proceed by backward induction to uniquely determine the number of targeted agents and the value functions for the principal and for the agents when the surviving population size is $n > 0$, before the principal's targeting choice.

Lemma 1 (Characterization under unpredictable-order)**.** *Any unpredictable-order coalition-proof equilibrium induces the same mappings from the current population size $n \in \mathbb{N}$ to (i) the number of targeted agents $\kappa(n) \in \{0, \ldots, n\}$, and (ii) the value functions for the principal $V^p(n)$ and for the agents $V(n)$, before the realization of who is targeted. Moreover, $V^p(n)$ is strictly decreasing in n on $\{n \in \mathbb{N} : n \geq N^*\}$, and the following recursive relationships hold:*[28]

$$\kappa(n) = \max\left\{\{0\} \cup \left\{k \in \{1, \ldots, \overline{k} - 1\} : 1 + \delta V(n-k) \geq 1 + \frac{\rho\delta}{1-\delta} \text{ and } n - k \geq N^*\right\}\right\}.$$

$$V^p(n) = u(n) + \delta V^p(n - \kappa(n)), \qquad V(n) = 1 + \delta V(n - \kappa(n))\frac{n - \kappa(n)}{n}.$$

Furthermore, $\kappa(n)$, $V^p(n)$, and $V(n)$ are all independent of N_0, the initial population size at $t = 0$.

Proof. First note that if $n \leq N^*$ the principal optimally sets $k(n) = 0$.

[28] The principal never optimally targets $k_t \geq \overline{k}$ agents in a single period. Otherwise, the targeted agents alone can (and will) form a protest of size at least $\overline{k}$; by Assumption 3, such a proposal is strictly suboptimal.

Second, suppose $n > N^*$ and denote by $m = n - N^*$ the number of excess agents the principal would like to eliminate. The principal can eliminate all remaining excess agents in one period whenever $m < \overline{k}$. Hence, for all $n \in \{N^*, ..., N^* + \overline{k} - 1\}$, $\kappa(n) = m$, $V(n) = 1 + \delta \frac{N^*}{N^*+m}(\frac{1}{1-\delta})$, and $V^p(n) = u(n) + \frac{\delta u(N^*)}{1-\delta}$, strictly decreasing in $n \in \{N^*, ..., N^* + \overline{k} - 1\}$.

Now suppose $\overline{n} \geq N^* + \overline{k} - 1$ and that $\kappa(n)$, $V^p(n)$, and $V(n)$ are uniquely determined for all $n \leq \overline{n} \in \mathbb{N}$, with $V^p(n)$ strictly decreasing over $n \in \{N^*, ..., \bar{n}\}$. We show that $\kappa(n)$, $V^p(n)$, and $V(n)$ are also uniquely determined at $n = \overline{n} + 1$, and that $V^p(\overline{n} + 1) < V^p(\overline{n})$.

By Assumption 2, the principal strictly prefers setting $k = 0$ (thereby avoiding any protest) to choosing any $k > 0$ that would trigger a protest that (weakly) exceeds size $\overline{k}$. Thus, for $n \geq N^* + \overline{k}$, the principal chooses $k > 0$ only if the untargeted agents do not protest, i.e.

$$1 + \delta V(\overline{n} + 1 - k) \geq 1 + \frac{\rho\delta}{1-\delta}. \tag{1}$$

To establish that the principal chooses $k > 0$ when it is feasible to do so without triggering a size $\overline{k}$ protest, suppose the principal targets no agents for $l > 0$ periods, after which it chooses $k \in \{1, ..., \overline{k}\}$. Then, the principal's payoff would be

$$\begin{aligned} u(\overline{n}+1)\sum_{i=0}^{l}\delta^i + \delta^{l+1}V^p(\overline{n}+1-k) <& u(\overline{n}+1) + u(\overline{n}+1-k)\sum_{i=1}^{l}\delta^i + \delta^{l+1}V^p(\overline{n}+1-k) \\ \leq& u(\overline{n}+1) + \delta V^p(\overline{n}+1-k), \end{aligned}$$

where the last expression captures the payoff the principal would get by targeting k today. As a result, the principal chooses $k > 0$ if it is feasible, and otherwise chooses $k = 0$.

The principal's problem at $\overline{n} + 1 \geq N^* + \overline{k}$ can be written as,

$$\max_{k \in S(\overline{n}+1)} V^p(\overline{n} + 1 - k),$$

where $S(\overline{n} + 1) = \left\{k \in \{1, \ldots, \overline{k} - 1\} : 1 + \delta\, V(\overline{n} + 1 - k) \geq 1 + \frac{\rho\delta}{1-\delta}\right\}$ if $S(\bar{n} + 1) \neq \emptyset$ and otherwise the principal sets $k = 0$ and $V^p(\overline{n} + 1) = u(\overline{n} + 1)\frac{1}{1-\delta}$

Because $V^p(n)$ is strictly decreasing over $\{N^*, ..., \overline{n}\}$, this implies that the principal sets

$$\kappa(\overline{n} + 1) = \max\left(\{0\} \cup S(\overline{n} + 1)\right); \qquad V^p(\overline{n} + 1) = u(\overline{n} + 1) + \delta V^p(\overline{n} + 1 - \kappa(\overline{n} + 1)).$$

Since targeting is uniform, each agent is untargeted with probability $\frac{\overline{n}+1-\kappa(\overline{n}+1)}{\overline{n}+1}$. Therefore,

$$V(\overline{n} + 1) = 1 + \delta V(\overline{n} + 1 - \kappa(\overline{n} + 1))\frac{\overline{n} + 1 - \kappa(\overline{n} + 1)}{\overline{n} + 1}.$$

If $\kappa(\overline{n}+1)=0$, then $V^p(\overline{n}+1)=\frac{u(\overline{n}+1)}{1-\delta}<\frac{u(\overline{n})}{1-\delta}\leq V^p(\overline{n})$. If $\kappa(\overline{n}+1)\geq 1$,

$$V^p(\overline{n}+1)=u(\overline{n}+1)+\delta V^p(\overline{n}+1-\kappa(\overline{n}+1))<u(\overline{n})+\delta V^p(\overline{n}+1-\kappa(\overline{n}+1))\leq V^p(\overline{n})$$

where the last inequality follows because, if the principal can target $\kappa(\overline{n}+1)$ agents from $\bar{n}+1$ without triggering protests by untargeted agents, she can also target $\kappa(\overline{n}+1)-1$ agents starting from $\bar{n}$ without triggering protests by untargeted agents: the protest incentives of untargeted agents depends only on the number of agents left after those targeted are eliminated, not on the starting point (see (1)). In either case, $V^p(\overline{n}+1)<V^p(\overline{n})$, completing the proof. □

Proof of Proposition 2 and 3. Consider any $N_0\in\mathbb{N}$. If $N_0\leq N^*$, the proof is complete, since the principal has reached her ideal point among feasible population sizes; the population cannot grow. Suppose $N_0>N^*$ for the remainder of the proof. Denote by $\mathcal{X}$ the set of integers x such that, in any coalition-proof subgame perfect equilibrium, an active principal does not eliminate any further agents once the number of survivors equals x, i.e., such that $\kappa(x)=0$.[29] Since the principal never eliminates any agent when the current population size is $n\leq N^*$, every $n\in\{0,...,N^*\}$ belongs to $\mathcal{X}$ and no untargeted agent would ever oppose the elmination of k agents from a starting population of size n when $n-k\leq N^*$.[30] Hence, by Lemma 1, $\kappa(n)=n-N^*$ agents are targeted and eliminated with probability 1 whenever $n\in\{N^*,...,N^*+\overline{k}-1\}$.

Moreover, in any coalition-proof equilibrium, population sizes in $\mathcal{X}$ are terminal: starting from any initial population $N_0\geq 0$, the process reaches the final population size

$$x_{N_0}\equiv\max\{x\in\mathcal{X}:x\leq N_0\}.$$

For every $N_0\in\{0,...,N^*+\overline{k}-1\}$ the result is trivial and x_{N_0} is reached in at most one period. The result, however holds also for $N_0>\overline{k}-1$. First, by definition, for any population size $n\in\{0,...,N_0\}$ such that $n\notin\mathcal{X}$, the principal would target $\kappa(n)\geq 1$ agents without triggering a size $\overline{k}$ protest.[31] Second, the principal does not try to reduce the surviving population below $x\in\mathcal{X}$ from any $n\geq x$. To see why, observe that if $n-x\geq\overline{k}$, the principal does not reduce the population below x in a single step: doing so would require targeting at least $n-x+1>\overline{k}$ agents,

[29] Note this is well defined by Lemma 1.

[30] Note that under Assumption 1, if $N^*=0$, then the smallest strictly positive element of $\mathcal{X}$ would be $\overline{k}$. At $N_t=\overline{k}$, agents would protest the proposed elimination of any agent, which by Assumption 2 is undesirable for the principal. On the other hand, if $N_t<\overline{k}$, the principal could eliminate all the agents with no risk of protest.

[31] By Assumption 2, triggering such a protest would be suboptimal, contradicting $\kappa(n)\geq 1$.

which would trigger a protest of size at least $\overline{k}$ (since $\overline{k}$ targeted agents would protest) which, by Assumption 2, the principal strictly prefers to avoid. Thus, the only n from which it might be optimal to try to reach $< x$ agents is for n s.t. $n - x < \overline{k}$. But note that for any $x > N^*$, $\kappa(x) = 0$ implies that targeting any number $k > 0$ of agents would trigger a size $\overline{k}$ protest starting from population size x, and therefore $1 + \delta V(s) < 1 + \frac{\rho\delta}{1-\delta}$ for all $s \in \{x - \underline{k}, \ldots, x - 1\}$, where $\underline{k} \equiv \overline{k} - 1$.[32] This, in turn, ensures $\kappa(n) \leq n - x$ for all $n \in \{x + 1, ..., x + \underline{k}\}$, since the principal refrains from triggering a protest of size at least $\overline{k}$ by Assumption 2. Hence, starting from any initial $N_0 \in \mathbb{N}$, the principal eliminates agents until the population size reaches x_{N_0} (within at most $N_0 - x_{N_0}$ periods) and then eliminates no more after reaching x_{N_0}.

It remains to show that terminal population sizes above N^* are finite. Choose T such that $\sum_{t=0}^{T} \delta^t > \frac{\rho}{1-\delta}$. Since targets are selected uniformly at random and at most $\underline{k}$ agents are eliminated per period, each agent survives until period t with probability at least $1 - t\underline{k}/n$. Thus,

$$V(n) \geq \sum_{t=0}^{T} \delta^t \left(1 - \frac{t\underline{k}}{n}\right).$$

For sufficiently large n, the right hand side exceeds $\frac{\rho}{1-\delta}$. Targeting one agent at population $n + 1$ thus induces no protest, so Lemma 1 gives $\kappa(n + 1) \geq 1$. Hence $\{x > N^* : \kappa(x) = 0\}$ is finite.

Define $\mathcal{B} := \{N^*\} \cup \{x > N^* : \kappa(x) = 0\}$. Its uniqueness and independence from u and N_0 follow because the recursion determining (κ, V) contains neither u nor N_0. By construction, $x_{N_0} = B^*_{N_0}$ for every N_0.

For $n = x + m$, where $m \in \{0, \ldots, N_0 - x\}$, if the principal targets $k \in \{1, \ldots, \underline{k}\}$ agents, with $x + m - k \geq N^*$, then untargeted agents protest in a coalition proof equilibrium if and only if $1 + \delta V(m + x - k) < 1 + \frac{\rho\delta}{1-\delta}$. Anticipating this, as shown, the principal optimally chooses

$$\kappa(x + m) = \max\left(\{0\} \cup \left\{k \in \{1, \ldots, \underline{k}\} : V(x + m - k) \geq \frac{\rho}{1 - \delta} \text{ and } x + m - k \geq N^*\right\}\right).$$

The fastest-approach value. From any initial population $x + m$ let $V^{\text{aux}}(x, m)$ denote an agent's expected continuation payoff from the next period onward along the hypothetical path on which, at every population $x + s$, the principal eliminates $\min\{\underline{k}, s\}$ uniformly selected agents until x remain, after which no more eliminations occur. Thus, the corresponding current value is $1 + \delta V^{\text{aux}}(x, m)$. Along this elimination path, since targets are selected uniformly at random,

[32] Recall that in our definition of coalition-proof equilibrium untargeted agents would not protest when indifferent.

each agent has probability $\frac{\min\{m,\underline{k}\}}{x+m}$ of being eliminated in the current period (period 1) and probability $\frac{x}{x+m}$ of never being eliminated. Moreover, if $m > \underline{k}$, the agent has probability $\frac{\underline{k}}{x+m}$ of being eliminated in each period $t \in \{1, \ldots, \lfloor m/\underline{k} \rfloor\}$,[33] and probability $\frac{m-\underline{k}\lfloor m/\underline{k}\rfloor}{x+m}$ of being eliminated in period $\lfloor m/\underline{k} \rfloor + 1$. As a result, $V^{\text{aux}}(x,m) \equiv V^{\text{auxx}}(x,m,\lfloor m/\underline{k}\rfloor)$, where

$$\begin{aligned} V^{\text{auxx}}(x,m,q) &\equiv \frac{1}{1-\delta} \times \frac{x}{x+m} + \left(\frac{m - q\,\underline{k}}{x+m}\right) \frac{1-\delta^{\,q}}{1-\delta} + \frac{\underline{k}}{x+m} \sum_{i=1}^{q} \frac{1-\delta^{i-1}}{1-\delta}, \\ &= \frac{1}{(1-\delta)^2} \times \frac{a_q\, m + b_q}{m+x} \end{aligned}$$

with $a_q = (1-\delta)\big(1-\delta^{\,q}\big)$ and $b_q = (1-\delta)x - \underline{k} + \underline{k}\,\delta^{\,q}\big[1 + (1-\delta)q\big]$.

Additionally, note that, for every $q \in \mathbb{N}$

$$V^{\text{aux}}(x,m) = V^{\text{auxx}}(x,m,q) \quad \text{for all } m \in I_q := \{\, m \in \mathbb{N} :\ q\,\underline{k} \le m \le (q+1)\,\underline{k} \,\}.$$

This holds, by definition, when $q\,\underline{k} \le m < (q+1)\,\underline{k}$, but also at $m = (q+1)\,\underline{k}$ since

$$\begin{aligned} &V^{\text{auxx}}(x,(q+1)\underline{k},q) - V^{\text{auxx}}(x,(q+1)\underline{k},q+1) \\ &= \left.\frac{(a_q - a_{q+1})m + (b_q - b_{q+1})}{(1-\delta)^2(m+x)}\right|_{m=(q+1)\underline{k}} \qquad \text{(Bdry Matching)} \\ &= \left.\frac{\delta^q\big(-m + (q+1)\underline{k}\big)}{m+x}\right|_{m=(q+1)\underline{k}} = 0. \end{aligned}$$

We argue now that $V^{aux}(x,m)$ is monotone in m within every integer block I_q. Fix any $q \in \mathbb{N}$. For any $m \in I_q \setminus \{(q+1)\underline{k}\}$, the forward difference satisfies

$$\Delta V^{\text{aux}}(x,m) := V^{\text{aux}}(x,m+1) - V^{\text{aux}}(x,m) = \frac{1}{(1-\delta)^2} \cdot \frac{a_q\, x - b_q}{(m+x)(m+1+x)}.$$

Hence the sign of $\Delta V^{\text{aux}}(x,m)$ is constant on each block I_q and equals the sign of

$$S(q) := a_q\, x - b_q = \underline{k} - \delta^{\,q}\Big[(1-\delta)\,x + \underline{k}\big(1 + (1-\delta)q\big)\Big].$$

Thus $V^{\text{aux}}(x,m)$ is monotone in m within every integer block I_q.

It is also straightforward to show that S crosses 0 at most once. Indeed, since

$$S(q+1) - S(q) = \delta^{\,q}(1-\delta)^2\big[x + \underline{k}(1+q)\big] > 0, \qquad \text{(Diff-S)}$$

there exists $q^* \in \{0, 1, ..., \infty\}$ such that $S(q) < 0$ for $q < q^*$ and $S(q) \ge 0$ for $q \ge q^*$.

[33] $\lfloor m/\underline{k} \rfloor$ denotes the largest integer not exceeding $m/\underline{k}$.

Combining this single-crossing property with the fact that $V^{aux}(x,m)$ is monotone on I_q with sign equal to $S(q)$ and equation (Bdry Matching), we find that $V^{\text{aux}}(x,m)$ is (discrete) quasi-convex in m. For every initial population size N_0 denote the minimum attainable $V^{aux}(x,m)$ as

$$\underline{v}^{\text{aux}}_{N_0} \equiv \min_{m\in\{0,1,\ldots,N_0-x\}} V^{\text{aux}}(x,m),$$

and denote by $m^*(N_0,x)$ the minimum minimizer in the problem above. Importantly, note that $m^*(N_0,x) \geq \min\{N_0-x,\underline{k}\}$ whenever $N_0 \geq N^* + \overline{k}$. We can see this by observing that $S(0) = -(1-\delta)x$, so that $V^{aux}(x,\underline{k}) < V^{aux}(x,s)$ for all $s \in \{0,...,\underline{k}-1\}$.

Mapping back to the original problem Given an initial population size of $N_0 \in \mathbb{N}$, let

$$x = \max\{\, s \in \mathcal{X} :\ s \leq N_0 \,\} \geq \min\{N^*, N_0\}$$

and consider any $m \in \{0,\ldots,N_0-x\}$.[34] From Lemma 1, any principal strategy in a coalition-proof equilibrium is represented by a function $\tilde{\kappa}:\mathbb{N}^2\to\mathbb{N}$ such that $\tilde{\kappa}(x,m) \equiv \kappa(x+m)$.

By the quasi-convexity of V^{aux}, the lower contour set

$$B := \left\{s \in \{0,...,N_0-x\} :\ 1+\delta\, V^{\text{aux}}(x,s) < \tfrac{\rho}{1-\delta}\right\}$$

is either a consecutive integer interval,

$$B = \{\underline{m},\ \underline{m}+1,\ \ldots,\ \overline{m}\} \quad \text{for some} \quad 0 \leq \underline{m} \leq \overline{m} \leq N_0 - x$$

or the empty set, i.e., $B=\emptyset$. Then, by Lemma 1 and under our tie-breaking rule in favor of not protest in the definition of coalition-proof equilibrium, we argue now that every coalition-proof equilibrium induces the same, unique $\tilde{\kappa}(x,m)$, with the following properties:[35]

1. If $m \leq \underline{k}$, then $\tilde{\kappa}(x,m) = m$ and $V(x+m) = 1+\delta V^{\text{aux}}(x,m)$.
2. If $N_0 - x \geq m > \underline{k}$ and $B=\emptyset$, then $\tilde{\kappa}(x,m) = \underline{k}$ and $V(x+m) = 1+\delta V^{\text{aux}}(x,m)$.
3. If $N_0-x \geq m > \underline{k}$ **and** $B = \{\underline{m},...,\overline{m}\} \neq \emptyset$ **and** $m \notin \{\underline{m}+\underline{k},...,\overline{m}+\underline{k}\}\cap\{0,...,N_0-x\}$, then $\tilde{\kappa}(x,m) = \underline{k}$ and $1+\delta V^{\text{aux}}(x,m) \leq V(x+m)$.
4. If $N_0-x \geq m > \underline{k}$ **and** $B = \{\underline{m},...,\overline{m}\} \neq \emptyset$ **and** $m \in \{\underline{m}+\underline{k},...,\overline{m}+\underline{k}\}\cap\{0,...,N_0-x\}$, then $\tilde{\kappa}(x,m) = m-(\overline{m}+1)$ and $1+\delta V^{\text{aux}}(x,m) \leq V(x+m)$.

[34] Recall that $\mathcal{X}\subseteq\mathbb{N}$ is the set of integers x such that, in any coalition-proof subgame perfect equilibrium, an active principal does not eliminate any further agents once the number of survivors equals x, i.e., such that $\kappa(x)=0$

[35] We say that a coalition-proof equilibrium *induces* a $\tilde{\kappa}(x,m)$ if at every history at which $N = x+m$, the principal's strategy is to target $\tilde{\kappa}(x,m)$ agents. As we show in the proof below, $\tilde{\kappa}(x,m)$ cannot be negative.

Call this the ***4-point characterization of*** $\tilde{\kappa}$

Point 1 follows because eliminating all m excess agents reaches the terminal population x without inducing protest. Point 2 follows from the fact that if $B = \emptyset$ and hence $1 + \delta\, \underline{v}^{\text{aux}}_{N_0} \geq \frac{\rho}{1-\delta}$, the fastest approach path induces no protest of size at least $\overline{k}$, and is the principal's most preferred path among those with $\tilde{\kappa} \leq \underline{k}$. We now show that the $\tilde{\kappa}(x, m)$ induced by any unpredictable-order coalition-proof equilibrium satisfies points 3 and 4.

By discrete quasi-convexity of V^{aux}, $B = \left\{s \in \{0, ..., N_0 - x\} :\ 1 + \delta\, V^{\text{aux}}(x, s) < \frac{\rho}{1-\delta}\right\}$ is a consecutive integer interval (and by the premise of points 3 and 4 that $B \neq \emptyset$). Then

$$B = \{\underline{m},\, \underline{m} + 1,\, \ldots,\, \overline{m}\} \quad \text{for some} \quad 0 \leq \underline{m} \leq \overline{m} \leq N_0 - x.$$

Point 1 and point 2, applied to each truncated problem with initial population $x+s$ with $\underline{k} < s < \underline{m}$, imply that $\tilde{\kappa}(x, m) = \min\{\underline{k}, m\}$ and $1+\delta V^{\text{aux}}(x, m) = V(x+m)$ for all $m \in \{0, \ldots, \underline{m}-1\}$. As a consequence, $\tilde{\kappa}(x, m) = \min\{\underline{k}, m\}$ and $1 + \delta V^{\text{aux}}(x, m) = V(x + m)$ also for all $m \in \{\underline{m}, \ldots, \underline{m} + \underline{k} - 1\} \cap \{0, ..., N_0 - x\}$.

Hence, for any population size $n = x + m > \underline{m}$, targeting k agents would be met with a protest of size at least $\overline{k}$ whenever $m - k \in \{\underline{m}, \ldots, \min\{\underline{m} + \underline{k} - 1, \overline{m}\}\}$.. Consequently, if $\{\underline{m}+\underline{k}, \ldots, \overline{m}+\underline{k}\} \cap \{0, \ldots, N_0 - x\} \neq \emptyset$, then $\overline{m} < \underline{m}+\underline{k}-1$ (if not, the definition of x would be contradicted); otherwise, point 4 is vacuous. In that case, $N_0 - x < \underline{m} + \underline{k}$, and the preceding argument already gives

$$\tilde{\kappa}(x, m) = \min\{\underline{k}, m\} \quad \text{and} \quad V(x + m) = 1 + \delta V^{\text{aux}}(x, m)$$

for every $m \in \{0, \ldots, N_0 - x\}$. Hence point 3 holds as well. Thus, for the remainder of the argument suppose $\{\underline{m} + \underline{k}, \ldots, \overline{m} + \underline{k}\} \cap \{0, \ldots, N_0 - x\} \neq \emptyset$.

If $\overline{m} < \underline{m} + \underline{k} - 1$, however, $\tilde{\kappa}(x, m) = \min\{\underline{k}, m\}$ and $1 + \delta V^{\text{aux}}(x, m) = V(x + m)$ also for all $m \in \{\overline{m}, \ldots, \underline{m} + \underline{k} - 1\} \cap \{0, ..., N_0 - x\}$, and no $s \in B$ can ever be reached on the equilibrium path. Hence, by Lemma 1,

$$\tilde{\kappa}(x, m) = m - (\overline{m} + 1) \quad \text{for all } m \in \{\underline{m} + \underline{k}, \ldots, \overline{m} + \underline{k}\} \cap \{0, ..., N_0 - x\},$$

and consequently

$$V(x + m) = 1 + \delta V\big(x + m - \tilde{\kappa}(x, m)\big)\frac{x + m - \tilde{\kappa}(x, m)}{x + m} > 1 + \delta V^{\text{aux}}\big(x, m\big).$$

Since $1+\delta V^{aux}(x,m) \geq \frac{\rho}{1-\delta}$ for all $m \in \{\overline{m}+1, ..., N_0 - x\}$, this implies $1+\delta V(x+m) \geq 1+\frac{\rho\delta}{1-\delta}$ for all $m \in \{\overline{m}+1, \ldots, \overline{m}+\underline{k}\} \cap \{0, ..., N_0 - x\}$. Thus, by induction on m, when $\overline{m}+\underline{k}+1 \leq N_0 - x$, Lemma 1 implies $\tilde{\kappa}(x,m) = \underline{k}$ (and $V(x+m) \geq 1+\delta V^{\text{aux}}(x,m)$) for all $m \in \{\overline{m}+\underline{k}+1, ..., N_0 - x\}$. Points 3 and 4 follow immediately.

To conclude, we observe three things. First, observe that the 4 point characterization of $\tilde{\kappa}(x,m)$ implies that for any initial number of agents N_0, there is a point $x \leq N_0$ and time t^*, such that (i) t^* is the first time that x agents is reached, (ii) before t^*, the principal eliminates $\underline{k}$ agents each period, except possibly at t^*-1 and at most one other period, and (iii) after t^*, no more eliminations occur. Second, we observe that the uniqueness of the coalition-proof path of eliminations follows from the uniqueness of $\tilde{\kappa}(x,m)$ satisfying our characterization. Third, it is straightforward to construct a coalition-proof SPE given 1 and our characterization of $\tilde{\kappa}(x,m)$: see Online Appendix Section 2. □

Proofs of Results in Section 6

Proof of Proposition 6. Without loss, assume $v_1 > v_2 > \ldots > v_{N_0}$. Let $\mathcal{H}(\mathcal{N}_t, m)$ denote the m agents in $\mathcal{N}_t$ with the highest values of v_i. We prove by induction on $q \in \mathbb{N}$ that, in every coalition-proof SPE, whenever $\Omega_t = 0$ and $N_t = N^* + q\overline{k} + m$, with $m \in \{0, \ldots, \overline{k}-1\}$, the principal eliminates exactly $\mathcal{H}(\mathcal{N}_t, m)$ in period t and no agent thereafter.

Base step ($q = 0$). If $N_t \in \{N^*, \ldots, N^* + \overline{k} - 1\}$, the principal eliminates exactly $m = N_t - N^*$ agents, those in $\mathcal{H}(\mathcal{N}_t, m)$, in period t and targets no agent afterward. Indeed, eliminating all (and only) those agents leaves exactly N^* survivors, after which the principal has no incentive to eliminate additional agents[36]. Hence targeting $\mathcal{H}(\mathcal{N}_t, m)$ does not induce a protest of size at least $\overline{k}$ and yields the highest feasible payoff for the principal starting from $\mathcal{N}_t$.

Inductive step. Fix $q \geq 1$ and suppose the claim holds for $q-1$. Consider any history h^t with $\Omega_t = 0$ and $N_t = N^* + q\overline{k} + m$, where $m \in \{0, \ldots, \overline{k}-1\}$. We first show that the principal cannot target more than m agents. Suppose instead that she targets a set $\mathcal{S}$, with $\overline{k} > |\mathcal{S}| \equiv S > m$ agents.[37] Unless agents mount a successful protest, the next-period population size would be $N_{t+1} = N^* + (q-1)\overline{k} + (\overline{k} + m - S)$. By the induction hypothesis, in period $t+1$ the principal

[36] Since $v_i < 1$, any further elimination would result in a flow payoff lower than $u(N^*) - \sum_{i \in \mathcal{N}_t \setminus \mathcal{H}(\mathcal{N}_t, m)} v_i$.

[37] By Assumption 3, she never targets $\overline{k}$ or more agents: doing so would trigger a protest of size at least $\overline{k}$.

would target and eliminate exactly $\mathcal{H}(\mathcal{N}_t \setminus \mathcal{S}, \overline{k} + m - S)$. But then, by Assumption 1, those agents strictly prefer to join the current protest rather than wait to be eliminated. Together with agents in $\mathcal{S}$, they would form a protest of size at least $\overline{k}$. By Assumption 3, the principal strictly prefers to avoid such a protest. Hence, in equilibrium, she does not target more than m agents.

Applying this argument to $m = 0$ shows that the principal targets no agents when $N_t = N^* + q\overline{k}$. When instead $m \in \{1, \ldots, \overline{k} - 1\}$, she can target m agents without inducing any potentially successful protest: their elimination leaves $N^* + q\overline{k}$ survivors, after which, by the boundary case $m = 0$ just established, no further elimination occurs. Applying the same argument at each continuation state, targeting fewer than m agents today cannot produce more than m eliminations in total. So, the principal strictly prefers to target all m agents immediately.

Conditional on targeting m agents, the principal chooses $\mathcal{H}(\mathcal{N}_t, m)$ because these agents impose the highest costs on the principal. This completes the induction.

Applying the result at $t = 0$, let $m^* \in \{0, \ldots, \overline{k} - 1\}$ be such that $N_0 - m^* \in \{N^*, N^* + \overline{k}, N^* + 2\overline{k}, \ldots\}$. The principal then eliminates exactly agents $\{1, \ldots, m^*\}$ in period 0. No protest succeeds, and no further elimination occurs. □

Proof of Proposition 7. Let $\kappa(\cdot)$ and $V(\cdot)$ denote the target-count and the agent value functions in the unpredictable-order environment, as characterized in Lemma 1. For any $S \subseteq \mathcal{N}_0$ and any $m \leq |S|$, let $\mathcal{H}(S, m)$ denote the m agents in S with the highest values of v_i. Since F is continuous, this set is almost surely unique. We prove inductively the following stronger statement.

Claim. Consider any history h^t with $\Omega_t = 0$ such that $|\mathcal{N}_t| = n$, and suppose that the posterior over $(v_i)_{i \in \mathcal{N}_t}$ is exchangeable. Then:

1. every surviving agent's ex ante continuation payoff at that history equals $V(n)$;
2. the principal targets exactly $\kappa(n)$ agents;
3. conditional on targeting $\kappa(n)$ agents, she targets exactly the agents in $\mathcal{H}(\mathcal{N}_t, \kappa(n))$;
4. no potentially successful protest forms.

The claim applies at $t = 0$ because the prior is i.i.d. and hence exchangeable. It also applies recursively along the equilibrium path: after any proposal K_t that no untargeted agent protests, the survivors are $\mathcal{N}_t \setminus K_t$, whose posterior is exchangeable by Definition 3.

Base cases. If $n \leq N^*$, the principal targets no agent. Removing an agent weakly lowers her flow

payoff, strictly so almost surely because $v_i \leq 1$ with continuous F. So no further elimination occurs, no protest of size at least $\overline{k}$ forms, and every survivor's continuation payoff is $\frac{1}{1-\delta} = V(n)$.

If instead $N^* < n \leq N^* + \overline{k} - 1$, the principal can target $n - N^* < \overline{k}$ agents and leave N^* survivors, after which no further elimination occurs; hence no untargeted agent protests. Targeting more would leave fewer than N^* agents and is suboptimal given $v_i \leq 1$. Targeting fewer is also suboptimal, as it only delays these eliminations. Hence the principal targets exactly $n-N^*$ agents and, since protest incentives depend only on the number of survivors, she optimally selects those in $\mathcal{H}(\mathcal{N}_t, n - N^*)$. Exchangeability then implies that each agent assigns probability N^*/n to surviving. So her ex ante continuation payoff is $1 + \delta\frac{N^*}{n}V(N^*) = V(n)$, where the equality follows from Lemma 1.

Inductive step. Fix $n \geq N^*+\overline{k}$ and suppose the claim holds for every smaller population $\{0, ..., n-1\}$. Consider any history h^t with $\Omega_t = 0$, $|\mathcal{N}_t| = n$, and exchangeable posterior over $(v_i)_{i\in\mathcal{N}_t}$.

Step 1: continuation value after any proposal of size $k < \overline{k}$.

Take any $\mathcal{K} \subseteq \mathcal{N}_t$ with $|K| = k < \overline{k}$, and let $U = \mathcal{N}_t \setminus \mathcal{K}$, so $|U| = n - k$. By Definition 3, conditional on $\ell^t = (h^t, \mathcal{K})$, the posterior over $(v_i)_{i\in U}$ is exchangeable. If no untargeted agent protests, then the next-period survivor set is exactly U, whose posterior is still exchangeable. Thus the induction hypothesis applies at population size $|U| = n - k$, and each surviving agent's ex ante continuation payoff equals $V(n-k)$. Hence every untargeted agent's continuation payoff from not protesting at the current proposal history is exactly $1+\delta V(n-k)$. Crucially, this depends only on the number of survivors $n - k$, not on the current targets' identities.

Step 2: feasible target counts are the same as in the unpredictable-order environment.

The principal never chooses $k \geq \overline{k}$, because targets alone would form a protest of size at least $\overline{k}$, which she strictly prefers to avoid by Assumption 3. She also never chooses k such that $n - k < N^*$, because leaving fewer than N^* survivors is suboptimal given $v_i \leq 1$.

Now consider any $k \in \{1, \ldots, \overline{k} - 1\}$ with $n - k \geq N^*$. By Step 1, every untargeted agent gets $1 + \delta V(n - k)$ from not protesting. Thus, targeting k agents is *feasible* (i.e., does not induce a potentially successful protest) if and only if $1 + \delta V(n - k) \geq 1 + \frac{\rho\delta}{1-\delta}$. Thus, as in Lemma 1, the set of feasible target counts at population size n is

$$\{0\} \cup \left\{k \in \{1, \ldots, \overline{k} - 1\} : n - k \geq N^* \text{ and } 1 + \delta V(n - k) \geq 1 + \frac{\rho\delta}{1 - \delta}\right\}.$$

Step 3: conditional on a feasible count, the principal targets the highest types.

Fix a feasible count $q < \overline{k}$ and a target set $\mathcal{K} \subseteq \mathcal{N}_t$ with $|\mathcal{K}| = q$. Suppose $\mathcal{K} \neq \mathcal{H}(\mathcal{N}_t, q)$. Then there exist $i \in \mathcal{N}_t \setminus \mathcal{K}$ and $j \in \mathcal{K}$ such that $v_i > v_j$. Define

$$\mathcal{K}' \equiv (\mathcal{K} \setminus \{j\}) \cup \{i\}, \qquad \mathcal{S} \equiv \mathcal{N}_t \setminus \mathcal{K}, \qquad \mathcal{S}' \equiv \mathcal{N}_t \setminus \mathcal{K}'.$$

By Step 2, feasibility depends only on the resulting number of survivors $n - q$, not on the targets' identities. Thus $\mathcal{K}'$ is feasible whenever $\mathcal{K}$ is. Moreover, $\mathcal{S}' = (\mathcal{S} \setminus \{i\}) \cup \{j\}$, so $|\mathcal{S}'| = |\mathcal{S}|$ and $\sum_{\ell \in \mathcal{S}'} v_\ell < \sum_{\ell \in \mathcal{S}} v_\ell$. From period $t+1$ onward, the set of feasible target counts depends only on the current number of survivors, by Step 2. Since $|\mathcal{S}'| = |\mathcal{S}| = n - q$, starting from $\mathcal{S}'$, the principal can replicate any future sequence of target counts feasible from $\mathcal{S}$. Along that replicated sequence, let $\mathcal{S}'_r$ and $\mathcal{S}_r$ denote the survivor sets after r periods starting from $\mathcal{S}'$ and $\mathcal{S}$, respectively. Then $\sum_{\ell \in \mathcal{S}'_r} v_\ell \leq \sum_{\ell \in \mathcal{S}_r} v_\ell$ for every $r \geq 0$, with strict inequality at $r = 0$. Thus the principal's continuation payoff is strictly higher under $\mathcal{K}'$ than under $\mathcal{K}$, contradicting optimality of $\mathcal{K}$. Hence, conditional on targeting q agents, the principal targets exactly $\mathcal{H}(\mathcal{N}_t, q)$.

Step 4: among feasible counts, the principal chooses the largest one.

Fix two feasible counts $0 \leq k < k' \leq \overline{k} - 1$. By Step 3, if the principal targets $a \in \{k, k'\}$ agents today, the next-period survivor set is $S_a \equiv \mathcal{N}_t \setminus \mathcal{H}(\mathcal{N}_t, a)$. Hence $S_{k'} \subset S_k$ and $|S_{k'}| = n - k' < n - k = |S_k|$. Consider the equilibrium continuation path after choosing k today, and let n_s be the population s periods later, so $n_1 = n - k$ and $n_{s+1} \leq n_s$.

Case 1. Suppose the path never reaches a population weakly below $n - k'$, i.e., $n_s > n - k'$ for every $s \geq 1$. Then, along the path starting from k, the total number of eliminations is strictly less than k'. By Step 3, each future survivor set thus strictly contains $S_{k'}$. But then the principal would strictly benefit from switching to targeting k' agents, leaving the survivor set $S_{k'}$ at period $t+1$, and then stopping forever. Hence k cannot be optimal.

Case 2. Suppose the path first reaches a population weakly below $n - k'$ at period $t + r$. Since $k' > k$, $r \geq 2$ and $n_r \leq n - k' < n_{r-1}$. By Step 3, after $n - n_r$ eliminations from $\mathcal{N}_t$, the survivor set is $\mathcal{R} = \mathcal{N}_t \setminus \mathcal{H}(\mathcal{N}_t, n - n_r)$, irrespective of how those eliminations are distributed over time. The path starting with k first reaches $\mathcal{R}$ at period $t + r$. So, the transition from $n_{r-1} > n - k'$ to $n_r \leq n - k'$ must be feasible on that path. Thus, by Step 2, any targeting of fewer than $\overline{k}$ agents that leaves exactly n_r survivors must be feasible on any path. Since $n_{r-1} > n - k'$ and

$n_{r-1} - n_r < \overline{k}$, it follows that $0 \leq (n - k') - n_r < \overline{k}$. Hence, after targeting k' agents today, the principal can in period $t+1$ target exactly $(n-k') - n_r$ additional agents, in particular the highest-v_i agents in $S_{k'}$, reaching $\mathcal{R}$ by period $t+2$. This second path reaches $\mathcal{R}$ weakly earlier and yields a strictly higher payoff at $t+1$ because $S_{k'} \subset S_k$. By waiting if necessary, it can then replicate the original continuation from $t+r$. Thus choosing k' strictly dominates choosing k.

Since this argument applies to every feasible pair $k < k'$, the principal chooses the largest feasible count. By Step 2, that count is exactly $\kappa(n)$.

Step 5: agent values. By Step 3 and 4, the principal targets exactly the agents in $\mathcal{H}(\mathcal{N}_t, \kappa(n))$. Since the posterior over $(v_i)_{i \in \mathcal{N}_t}$ is exchangeable, each current survivor is ex ante equally likely to belong to that target set. Therefore each survives the current proposal with probability $\frac{n-\kappa(n)}{n}$. Because no untargeted agent protests, the ex ante continuation payoff of any surviving agent is $1 + \delta \frac{n-\kappa(n)}{n} V(n - \kappa(n)) = V(n)$, where the equality follows from Lemma 1. This proves the claim for population size n and completes the induction.

Applying the claim at the initial history yields the proposition: equilibrium target counts and protest outcomes in each period coincide with those in the unpredictable-order environment with the same initial population size N_0. Conditional on targeting k_t agents in period t, the principal targets the k_t surviving agents with the highest values of v_i. □

A.2 Proof of Proposition 4

Proof. Fix B_j. For every $m \in \{1, \ldots, \overline{k} - 1\}$, starting from $B_j + m$ the principal can target m agents and reach B_j. Untargeted agents then anticipate no further elimination and do not protest. Hence none of these population sizes is blocking, so $B_{j+1} - B_j \geq \overline{k}$.

Moreover, since B_j is terminal we have $V(B_j) = \frac{1}{1-\delta}$, and from $B_j + m$ the principal targets exactly m agents, so the recursion in Lemma 1 gives $V(B_j + m) = 1 + \frac{\delta B_j}{(1-\delta)(B_j+m)}$, which is decreasing in m. Thus, from $B_j + \overline{k}$, the easiest positive proposal to sustain targets $\overline{k} - 1$ agents and leaves $B_j + 1$ survivors. By the protest condition, $B_j + \overline{k}$ is blocking if and only if $V(B_j + 1) < \frac{\rho}{1-\delta}$, or, equivalently, $B_j < \frac{\delta}{1-\rho} - 1 = \widehat{N}(\delta, \rho)$. Therefore the next gap equals $\overline{k}$ if $B_j < \widehat{N}(\delta, \rho)$ and is strictly larger otherwise. Finally, maximality of B_n implies $B_n \geq \widehat{N}(\delta, \rho)$; otherwise, $B_n + \overline{k}$ would be another blocking size.

The second part of the proposition follows from the fact that under unpredictable-order coalition-

proof SPE the principal's strategy at a given population size N_t is independent of N_0, by Lemma 1. As a result, for any parameterization, once $N_0 > B_n + \overline{k}$ and lies beyond the finitely many population sizes at which the principal applies a non-greedy elimination, the principal eliminates at least $\overline{k} - 1$ agents immediately as well as at least $\overline{k}$ agents in aggregate, achieving a strictly higher payoff than under predictable-order coalition-proof SPE. □

A.3 Proof of Proposition 5

Recall. Fix the primitives of Section 4.2, write $\underline{k} = \overline{k} - 1$, and set $p \equiv \frac{\rho}{1-\delta}$. We make $\underline{k}$ an explicit argument of V^{auxx} and V^{aux}, i.e. $V^{\text{auxx}}(x, m, q; \underline{k})$ and $V^{\text{aux}}(x, m; \underline{k})$.[38] Monotonicity of V^{aux} on each block I_q is governed by

$$S(q; x, \underline{k}) = a_q x - b_q = \underline{k} - \delta^q \big[(1-\delta)x + \underline{k}\big(1 + (1-\delta)q\big)\big],$$

which is strictly increasing in q by (Diff-S), with single crossing $q^*(x; \underline{k}) = \min\{q : S(q; x, \underline{k}) \geq 0\}$. The protest (lower-contour) set $B(x; \underline{k}) = \{m : 1 + \delta V^{\text{aux}}(x, m; \underline{k}) < p\}$ is, by quasi-convexity, a contiguous interval $\{\underline{m}, \ldots, \overline{m}\}$ or empty; when non-empty, its lower endpoint is $\underline{m}(x; \underline{k}) = \lfloor \widehat{m}(x; \underline{k}) \rfloor + 1$ where $\widehat{m}(x; \underline{k})$ is the unique solution on the decreasing branch of $1 + \delta V^{\text{aux}}(x, \widehat{m}; \underline{k}) = p$. By the *4-point characterization of* $\tilde{\kappa}$ in the proof of Proposition 2, the tuple $\mathcal{B}(\overline{k}) = (B_0, \ldots, B_n)$ of Proposition 2 satisfies, for every $j \geq 2$ such that $\overline{m}_j - \underline{m}_j \geq \underline{k} - 1$,

$$B_j(\overline{k}) = B_{j-1}(\overline{k}) + \underline{m}_j(\overline{k}) + \underline{k}, \qquad \underline{m}_j(\overline{k}) = \underline{m}\big(B_{j-1}(\overline{k}); \underline{k}\big). \tag{2}$$

More precisely, the 4-point characterization in Proposition 2 gives the following recursive stopping rule. For $j \geq 2$, given $B_{j-1}(\overline{k})$, consider the lower-contour set

$$B(B_{j-1}(\overline{k}); \underline{k}) = \{m : 1 + \delta V^{aux}(B_{j-1}(\overline{k}), m; \underline{k}) < p\}.$$

If this set is empty, the construction stops. If it is nonempty, write $B(B_{j-1}(\overline{k}); \underline{k}) = \{\underline{m}_j, \underline{m}_j + 1, \ldots, \bar{m}_j\}$. A new finite threshold is generated if and only if $\bar{m}_j - \underline{m}_j \geq \underline{k} - 1$. In that case, $B_j(\overline{k}) = B_{j-1}(\overline{k}) + \underline{m}_j + \underline{k}$. If instead $\bar{m}_j - \underline{m}_j < \underline{k} - 1$, the construction stops. We impose

$$N^* = 0, \qquad \text{so that} \qquad B_0 = 0, \quad B_1 = \max\{N^*, \overline{k}\} = \overline{k} \tag{3}$$

[38] Since $V^{\text{auxx}}(x, (q+1)\underline{k}, q; \underline{k}) = V^{\text{auxx}}(x, (q+1)\underline{k}, q+1; \underline{k})$ by (Bdry Matching), $V^{\text{aux}}(\cdot)$ extends continuously in m, we abuse notation and write $V^{\text{aux}}(x, m; \underline{k})$ for this continuous extension).

Lemma 2 (Joint homogeneity in $(x, m, \underline{k})$). *Suppose $N^* = 0$. For every $\lambda > 0$, $q \in \mathbb{N}$, and (x, m),*

$$V^{\text{auxx}}(\lambda x, \lambda m, q; \lambda \underline{k}) = V^{\text{auxx}}(x, m, q; \underline{k}).$$

Consequently $S(q; \lambda x, \lambda \underline{k}) = \lambda\, S(q; x, \underline{k})$, so q^* and the protest-set block index are invariant under $(x, \underline{k}) \mapsto (\lambda x, \lambda \underline{k})$, and $\widehat{m}(\lambda x; \lambda \underline{k}) = \lambda\, \widehat{m}(x; \underline{k})$.

Proof. In the closed form, b_q scales by λ since $b_q(\lambda x; \lambda \underline{k}) = (1-\delta)\lambda x - \lambda \underline{k} + \lambda \underline{k} \delta^q [1 + (1-\delta)q] = \lambda b_q(x; \underline{k})$, the numerator $a_q(\lambda m) + b_q(\lambda x; \lambda \underline{k})$ scales by λ, and the denominator $(\lambda m + \lambda x)$ scales by λ; the ratio is unchanged. Then $S(q; \lambda x, \lambda \underline{k}) = a_q \lambda x - b_q(\lambda x; \lambda \underline{k}) = \lambda S(q; x, \underline{k})$ has the same sign, so q^* and the decreasing branch are unchanged, as is the block index defined by $1 + \delta V^{\text{auxx}}(x, (q+1)\underline{k}, q; \underline{k}) < p$. If $\widehat{m}$ solves $1 + \delta V^{\text{aux}}(x, \widehat{m}; \underline{k}) = p$, then by the displayed identity (applied to the continuous extension) $\lambda \widehat{m}$ solves $1 + \delta V^{\text{aux}}(\lambda x, \cdot\,; \lambda \underline{k}) = p$; uniqueness on the decreasing branch gives $\widehat{m}(\lambda x; \lambda \underline{k}) = \lambda \widehat{m}(x; \underline{k})$. □

By Lemma 2, $\widehat{m}(x; \underline{k})/\underline{k}$ depends on $(x, \underline{k})$ only through $\beta = x/\underline{k}$; write

$$g(\beta; \rho, \delta) \;:=\; \frac{\widehat{m}(\beta \underline{k}; \underline{k})}{\underline{k}}$$

for this *normalized crossing*. The crossing sits in the block I_{q_ℓ}, where $q_\ell = q_\ell(\beta; \rho, \delta) = \lfloor g(\beta; \rho, \delta) \rfloor < q^*$; on that block, multiplying the equation $1 + \delta V^{\text{aux}}(x, \widehat{m}; \underline{k}) = p$ by $(\widehat{m} + x)$ yields an equation that is affine in $\widehat{m}$ with a constant coefficient on $\widehat{m}$ independent of x, so the normalized root g is affine in β there. Hence g is piecewise linear and continuous, its kinks where $g(\beta; \rho, \delta) \in \mathbb{Z}$. For $j \geq 1$, write $\beta_j(\overline{k}) = B_j(\overline{k})/\underline{k}$. Going forward, we suppress dependence of $g(\beta; \rho, \delta)$ on (ρ, δ).

We also define $v^c(\beta; \delta) := \inf_{\mu \geq 0} V^{aux}(\beta \underline{k}, \mu \underline{k}; \underline{k})$, which, by Lemma 2, is independent of $\underline{k}$. Whenever $1 + \delta v^c(\beta; \delta) < p$ and hence $B(\beta \underline{k}; \underline{k})$ is non-empty denote by $\widehat{m}_\ell(\beta \underline{k}; \underline{k})$ and $\widehat{m}_u(\beta \underline{k}; \underline{k})$, the smallest and largest solutions of $1 + \delta V^{aux}(\beta \underline{k}, m; \underline{k}) = p$. Define $w(\beta; \rho, \delta) := \frac{\widehat{m}_u(\beta \underline{k}; \underline{k}) - \widehat{m}_\ell(\beta \underline{k}; \underline{k})}{\underline{k}}$. which, by Lemma 2, is independent of $\underline{k}$. We suppress dependence of w on (ρ, δ) going forward.

Proof of Proposition 5. Recall $\underline{k} = \overline{k} - 1$. We first establish that the β_j ratios converge to a limit that does not depend on $\overline{k}$, conditional on continuation. By Lemma 2, the construction depends on $\overline{k}$ through β, so throughout this step $\overline{k}$ denotes a generic cutoff. At a step where the recursive construction generates a finite threshold, eq. (2) applies. Thus, if $B_{j+1}(\overline{k})$ is finite, dividing (2) by $\underline{k}$, using $\underline{m} = \lfloor \widehat{m} \rfloor + 1$, and using $\widehat{m}/\underline{k} = g$,

$$\beta_{j+1}(\overline{k}) = \beta_j(\overline{k}) + \frac{\underline{m}_{j+1}(\overline{k})}{\underline{k}} + 1 = \beta_j(\overline{k}) + g(\beta_j(\overline{k})) + 1 + O\left(\tfrac{1}{\underline{k}}\right),$$

the $O(1/\underline{k})$ being the integer-rounding error. By (3), $\beta_1(\overline{k}) = \overline{k}/\underline{k} = 1 + 1/\underline{k} \to 1 =: \beta_1^\infty$.

Define $\beta_{j+1}^\infty = \beta_j^\infty + g(\beta_j^\infty) + 1$, continuing at step j if and only if $1 + \delta v^c(\beta_j^\infty; \delta) < p$ and $w(\beta_j^\infty; \rho, \delta) \geq 1$, and terminating otherwise. It is straightforward to show that it terminates after finitely many steps; let $n^\infty < \infty$ be such that the limiting sequence is $(\beta_1^\infty, \ldots, \beta_{n^\infty}^\infty)$.

Same length. By Proposition 2, the recursion appends a new finite entry B_{j+1} exactly when the trough of the fastest-approach value at stopping point B_j lies strictly below p and $\overline{m}_{j+1}(\overline{k}) - \underline{m}_{j+1}(\overline{k}) \geq \underline{k} - 1$; otherwise the tuple terminates. Fix $j \geq 1$ and suppose the following holds:

$$\textit{there exists } \overline{N}_{n^\infty} \textit{ such that, for all } \overline{k} > \overline{N}_{n^\infty}, \mathcal{B}(\overline{k}) \textit{ contains } B_j(\overline{k}), \textit{ and } \beta_j(\overline{k}) \to \beta_j^\infty. \qquad (H_j)$$

This hypothesis is verified by induction at the end of this step. As in the previous step, $\overline{k}$ denotes a generic cutoff: by Lemma 2, v^c, w, and g depend only on β. By Lemma 2 the *continuous* trough is a function $v^c(\beta; \delta)$ of the ratio β alone; the discrete trough at B_j differs from $v^c(\beta_j; \delta)$ by $O(1/\underline{k})$,[39] hence converges to $v^c(\beta_j^\infty; \delta)$ as $\overline{k} \to \infty$, and whenever $1 + \delta v^c(\beta_j^\infty; \delta) < p$,

$$\frac{\bar{m}_{j+1}(\overline{k}) - \underline{m}_{j+1}(\overline{k})}{\underline{k}} \to w(\beta_j^\infty; \rho, \delta).$$

The condition $\overline{m}_{j+1}(\overline{k}) - \underline{m}_{j+1}(\overline{k}) \geq \underline{k} - 1$ can be written $\frac{\bar{m}_{j+1}(\overline{k}) - \underline{m}_{j+1}(\overline{k})}{\underline{k}} \geq 1 - \frac{1}{\underline{k}}$. Since $\frac{\bar{m}_{j+1}(\overline{k}) - \underline{m}_{j+1}(\overline{k})}{\underline{k}} \to w(\beta_j^\infty; \rho, \delta)$, the inequality $\overline{m}_{j+1}(\overline{k}) - \underline{m}_{j+1}(\overline{k}) \geq \underline{k} - 1$ holds for all sufficiently large $\overline{k}$ when $w(\beta_j^\infty; \rho, \delta) > 1$ and fails for all sufficiently large $\overline{k}$ when $w(\beta_j^\infty; \rho, \delta) < 1$; when $w(\beta_j^\infty; \rho, \delta) = 1$, the limit is inconclusive. Let $\mathcal{E}$ be the set of primitive pairs (ρ, δ) for which, for some j, either $1 + \delta v^c(\beta_j^\infty; \delta) = p$, or both $1 + \delta v^c(\beta_j^\infty; \delta) < p$ and $w(\beta_j^\infty; \rho, \delta) = 1$. The set $\mathcal{E}$ has measure zero. For the rest of the proof, fix $(\rho, \delta) \notin \mathcal{E}$. Then both continuation conditions hold strictly at β_j^∞ for every $j < n^\infty$, while at $j = n^\infty$ the condition that fails does so strictly.

We now verify (H_j), by induction for $j \leq n^\infty$, and that no $B_{n^\infty+1}$ arises for all large enough $\overline{k}$.

- *Base case ($j = 1$):* the construction generates B_1 by (3), and $\beta_1(\overline{k}) \to \beta_1^\infty$ was shown above.
- *Inductive step ($j < n^\infty$):* given (H_j), the argument above show that the discrete trough and width conditions at B_j converge to their limiting counterparts at β_j^∞, which hold strictly; hence there exists $\overline{N}_j$ such that for all $\overline{k} > \overline{N}_j$ the construction generates B_{j+1}, and equation (2), implies $\beta_{j+1}(\overline{k}) \to \beta_{j+1}^\infty$, i.e., ($H_{j+1}$) holds.

[39] In fact, the two coincide: within each block $m \in [q\underline{k}, (q+1)\underline{k}]$, $\frac{\partial}{\partial m} V^{\text{aux}}(x, m; \underline{k}) = \frac{S(q; x, \underline{k})}{(1-\delta)^2 (m+x)^2}$, so by single crossing of S the continuous infimum is attained at the block boundary $m = q^*(x; \underline{k})\, \underline{k} \in \mathbb{N}$, an integer.

- *Termination ($j = n^\infty$):* by the same convergence, the strictly failing condition at $\beta^\infty_{n^\infty}$ implies that for all sufficiently large $\overline{k}$ the construction does not generate $B_{n^\infty+1}$. Setting $\overline{N}(\kappa, \delta, \rho) = \max_{1\le j\le n^\infty} \overline{N}_j$, for all $\overline{k} > \overline{N}$ the tuple is $\mathcal{B}(\overline{k}) = \big(0, B_1(\overline{k}), \ldots, B_{n^\infty}(\overline{k})\big)$.

The continuous piecewise-affine function g is locally Lipschitz at each β^∞_j with $j < n^\infty$. So for $j < n^\infty$ and all sufficiently large $\overline{k}$, $\big|\beta_{j+1}(\overline{k}) - \beta^\infty_{j+1}\big| \le (1+L_j)\,\big|\beta_j(\overline{k}) - \beta^\infty_j\big| + \frac{1}{\underline{k}}$, for some finite constants L_j. Since $\beta_1(\overline{k}) - \beta^\infty_1 = 1/\underline{k}$, induction gives $\beta_j(\overline{k}) = \beta^\infty_j + O(1/\overline{k})$. So, for $\overline{k}$ large enough and any $1 \le j \le n^\infty$,

$$B_j(\overline{k}) = (\overline{k}-1)\beta^\infty_j + O(1), \quad B_j(\lfloor\kappa\overline{k}\rfloor) = (\lfloor\kappa\overline{k}\rfloor - 1)\beta^\infty_j + O(1)$$

And therefore, $B_j(\lfloor\kappa\overline{k}\rfloor) - \kappa B_j(\overline{k}) = \big[\lfloor\kappa\overline{k}\rfloor - \kappa\overline{k} + \kappa - 1\big]\beta^\infty_j + O(1) = O(1)$. Dividing by $\overline{k}$ establishes the result.

□

Online Appendix to *Purges and Predictability*

Sam Kapon
UC Berkeley, Haas BPP

Matteo Camboni
UWisconsin–Madison

A Further discussion of literature

Connection to Ali et al. (2023): To see more clearly the connection between our model and Ali et al. (2023), consider the following model, which is a variant of our model made to be a special case of Ali et al. (2023). There is an agenda-setter and 3 agents. There are 2 periods of proposals that the agenda-setter makes, with period indexed $t \in \{0, 1\}$, and payoffs only accrue after the last period of proposals. The policy proposed by the principal is the set of agents to be eliminated. Each agent's payoff is $\mathbf{1}\{\text{NOT ELIMINATED}\} + \epsilon \times \text{NUMBER OF OTHER ELIMINATED AGENTS}$ and the principal's payoff is $\text{NUMBER OF ELIMINATED AGENTS} + \epsilon \times \text{SUM OF INDEXES OF ELIMINATED AGENTS}$, for very small ϵ. Each period, there is a status quo set of eliminated agents (with eliminating no agents as the period 0 status quo), and the agenda-setter can propose any revision to the set of eliminated agents, and if at least 2 agents approve (not including the agenda-setter) then the proposal passes, and otherwise it fails.

First, consider the case in which a failed proposal does not stop the agenda-setter from future proposals, as in the setting of Ali et al. (2023). Applying the logic from Ali et al. (2023), the agenda-setter will propose the favorite improvement operator each period to eliminate 2 out of 3 agents.

Suppose instead that a failed proposal stops the principal from making any future proposals. Then, since the status quo in period 0 is to eliminate no agents, the agenda-setter is only able to eliminate agent $i = 3$, which he passes with a proposal in the final period. Why? First, observe that if no past proposal has failed, then in the final period, the principal will successfully propose to eliminate the highest indexed agent who would not be eliminated according to the status quo. Then, suppose that in period 0 the agenda-setter proposes to eliminate at least one agent. Then all of the agents who are proposed to be eliminated, as well as the highest indexed non-eliminated

agent, would vote against the proposal, since in the event that the proposal fails they will get payoff at least 1, while in the event that the proposal succeeds they will get payoff no more than 2ϵ, which is assumed very small. As a result, the agenda-setter is only able to eliminate agent $i = 3$ in the last period.

Connection to mixed strategy equilibria in bargaining As mentioned in the literature review of Section 2, the randomization resulting from mixed strategy equilibria in bargaining is qualitatively different than the randomization we study. The core idea in a 3 agent, split-the-dollar, version of the model is as follows (from Diermeier and Fong (2009)). Suppose status-quo for the agenda-setter is 0, and for the others is $\frac{1}{2}$ each. Then consider the following behavior. The agenda-setter randomly selects one agent, and proposes an adjustment that gives the entire 1 to that agent; the agenda-setter and the selected agent vote in favor. In period 2, the agenda-setter then proposes 1 for herself, the agenda-setter and agent whose status-quo payoff in period 2 is 0 vote in favor. It is straightforward to argue that once the status quo payoff for one of the non-agenda-setter agents has reached 0, the agenda-setter can redistribute the entire 1 to herself. Taking that as given, the path can be sustained as an equilibrium path because if the contemporaneously selected agent rejects, she anticipates that the other agent will eventually be selected, in which case that agent will get the entire 1 for a single period, and then both will get 0 everafter.

B Commitment

We study a commitment version of the model in Section 3. We maintain Assumptions 1 and 2, assume $N_0 > N^*$, and specialize the principal's flow payoff to $u(n) = A - (n - N^*)^2$, where A is large enough that $u(n) \geq 0$ for all $n \in \{0, \ldots, N_0\}$.

At $t = 0$, the principal commits to agents' scheduled targeting times and to a public disclosure rule that provides information about those times. Fix a countably infinite signal space $\mathcal{S}$. Let $s_t \in \mathcal{S}$ denote a signal observed by all agents after the principal chooses $\mathcal{K}_t$ and before agents choose whether to protest. We redefine histories as

$$h^t \equiv (\mathcal{K}_r, s_r, \mathcal{P}_r, \omega_r)_{r \leq t-1}, \qquad \ell^t \equiv (h^t, \mathcal{K}_t, s_t).$$

Definition 4 (Policy). *A* policy *is a pair* $\gamma = (P, \Pi)$ *such that:*

- $P \in \Delta\left(\prod_{i\in\mathcal{N}_0}(\mathbb{N}\cup\{\infty\})\right)$ *is a joint distribution over* scheduled proposal-time profiles $\tau = (\tau_i)_{i\in\mathcal{N}_0}$. *At* $t = 0$, τ *is drawn according to* P *and observed by the principal. If no successful protest has occurred,* $\mathcal{K}_t = \{i \in \mathcal{N}_t : \tau_i = t\}$.

- *For each* $t \geq 0$, *scheduled proposal-time profile* τ, *history* h^t, *and target set* $\mathcal{K}_t$, $\Pi_t(\tau, h^t, \mathcal{K}_t) \in \Delta(\mathcal{S})$ *is the distribution from which* s_t *is drawn at period* t. *We require*

$$\Pi_t(\tau, h^t, \mathcal{K}_t) = \Pi_t(\tau, \widetilde{h}^t, \mathcal{K}_t) \tag{PI}$$

whenever h^t *and* $\widetilde{h}^t$ *have the same past target sets, signals, and realizations of* ω, *and induce the same* $(\mathcal{N}_t, \Omega_t)$. *We call* $\Pi = (\Pi_t)_{t\geq 0}$ *a* public disclosure rule.

If disclosure is allowed to depend on past protest behavior, one can construct disclosure rules that admit no coalition-proof equilibrium.

The principal selects a policy, which agents observe before play begins. A *coalition-proof equilibrium* under a policy γ is a perfect Bayesian equilibrium of the induced game such that, at every public history with a proposal, no nonempty coalition of agents has a joint pure one-shot deviation that strictly increases every member's continuation payoff, and every agent who is indifferent between protesting and not protesting chooses not to protest.

By Corollary 1 (see Section B.1 below), every policy γ admits a coalition-proof equilibrium. Fix an arbitrary equilibrium selection e, where $e(\gamma)$ is a coalition-proof equilibrium under γ for every policy γ. Let $W(\gamma)$ be the principal's payoff under $e(\gamma)$. A policy is *optimal* if it attains $\sup_{\gamma'} W(\gamma')$.[40]

Count processes and disclosures Let $\mathcal{C}$ denote the set of sequences $k = (k_t)_{t\in\mathbb{N}}$ satisfying $k_t \in \{0, \ldots, \bar{k} - 1\}$ and $\sum_t k_t \leq N_0 - N^*$.[41] A *count process* is a distribution $\mu \in \Delta(\mathcal{C})$, with $N_{t+1} = N_t - k_t$.

Given a count process μ, let P^μ be the distribution of scheduled targeting times obtained by drawing $k \sim \mu$ and then assigning the times symmetrically randomly across agents: conditional on k, every profile with exactly k_t agents assigned time t for each t is equally likely, and all remaining agents are assigned time ∞.

[40] Restricting scheduled targeting times to be independent of agents' protest behavior is without loss of generality for maximizing the principal's value, among policies that admit a coalition-proof equilibrium.

[41] We suppress dependence of $\mathcal{C}$ on parameters.

Definition 5. *We say that* Π *is a* ***rolling-window disclosure rule*** *if, at any scheduled targeting time profile* τ*, history* h^t*, and target set* $\mathcal{K}_t$*, the signal* s_t *can be identified with a binary partition of the set* $\mathcal{N}_t \backslash \mathcal{K}_t$ *into, (i) a set of size* $\tilde{w}(\tau, h^t, \mathcal{K}_t) \equiv \max\{0, \min\{\bar{k} - 1 - k_t, N_t - k_t - N^*\}\}$*, with the earliest scheduled targeting times among agents in* $\mathcal{N}_t \backslash \mathcal{K}_t$*, (ii) the remaining agents. Ties for inclusion in the set defined in (i) are broken using an order on* $\mathcal{N}_0$*, drawn uniformly at time 0 independently of* τ*, not disclosed to agents, and used at every history.*

All rolling-window disclosure rules differ only in their signal labeling, and so we use $\Pi^\star$ to denote the essentially unique rolling-window disclosure rule. Under $(P^\mu, \Pi^\star)$, scheduled times are assigned uniformly across agents. The disclosure reveals up to $\bar{k} - 1$ members of $\mathcal{N} \backslash \mathcal{K}_t$ and leaves all others uniformly uncertain about when and whether they will be proposed.

For $i \in \mathcal{N}_t \backslash \mathcal{K}_t$ at a positive probability history ℓ^t, define $\psi_i(\ell^t) \equiv \mathbb{E}[\delta^{\tau_i - t} \mid \ell^t]$.

Proposition 8 (Rolling-window disclosure)**.** *For every environment satisfying Assumptions 1 and 2, there exists a count process* $\mu^\star$ *such that* $(P^{\mu^\star}, \Pi^\star)$ *is an optimal policy.*

Proposition 9 (Arbitrarily long optimal delay)**.** *Fix any integer* $m \geq 1$ *and any* $\rho \in (0, 1)$*. Suppose* $\delta \in \left(((N^* + 1)(1 - \rho))^{1/m}, 1\right)$*, Assumption 2 holds, and* $N_0 = N^* + \bar{k}$*. Then there exists an optimal policy that satisfies* $\max_{i \in \mathcal{N}_0} \Pr(m < \tau_i < \infty) > 0$*.*

Note that the delay identified in Proposition 9 need not arise under every optimal policy.[42] Nevertheless, we can show that such delay becomes necessary for optimality if we restrict attention to deterministic policies or if the principal is even slightly more patient than the agents.

B.1 Lemmas and proofs of Propositions 8 and 9

Lemma 3 (Protest-stage existence)**.** *Fix* ℓ^t *with* $\mathcal{K}_t \neq \varnothing$ *and a strategy profile for the agents. Suppose every* $i \in \mathcal{N}_t$ *receives the same payoff following every protest set with fewer than* $\bar{k}$ *members. Then there exists an action profile at* ℓ^t *in which no agent has a profitable unilateral one-shot deviation, no coalition has a strictly (for all) profitable pure joint one-shot deviation, and no indifferent agent protests.*

[42] For instance, with $N^* = 0$, a policy that targets $\bar{k} - 1$ agents immediately and offers the remaining agent a simple lottery, in which she survives forever with some probability and is eliminated in the following period with the remaining probability, also maximizes the principal's value.

Proof. For each $i \in \mathcal{N}_t$, let v_i denote her expected discounted payoff from ℓ^t onward when fewer than $\bar{k}$ agents protest. Define

$$D \equiv \left\{ i \in \mathcal{N}_t : v_i < 1 + \frac{\rho\delta}{1-\delta} \right\}.$$

If $|D| < \bar{k}$, prescribe no protest. A pure joint one-shot deviation with fewer than $\bar{k}$ protesters leaves every deviator's payoff unchanged. A potentially successful deviation requires at least $\bar{k}$ protesting deviators. Each receives $1+\rho\delta/(1-\delta)$, so strict profitability would require all of them to belong to D. This is impossible when $|D| < \bar{k}$.

If $|D| \geq \bar{k}$, choose any $C \subseteq D$ with $|C| = \bar{k}$ and prescribe that exactly the agents in C protest. Each member of C receives $1 + \rho\delta/(1-\delta)$. If she alone deviates to not protest, fewer than $\bar{k}$ agents protest and she receives v_i, which is strictly smaller. An agent outside C receives at least $1 + \rho\delta/(1-\delta)$ by not protesting. Thus no agent has a profitable unilateral deviation.

Consider a pure joint one-shot deviation, and let P be the resulting protest set. If $|P| < \bar{k}$, at least one member of C in the deviating coalition chooses not to protest and receives v_i rather than $1 + \rho\delta/(1-\delta)$. If $|P| \geq \bar{k}$ and $P \neq C$, then $P \setminus C$ is nonempty. Every agent in $P \setminus C$ joins the protest as a member of the deviating coalition and receives $1+\rho\delta/(1-\delta)$, which is no greater than her payoff from the prescribed action of not protesting. If $P = C$, no action changes. Hence no pure joint one-shot deviation strictly benefits every coalition member.

When $|D| < \bar{k}$, no agent protests. When $|D| \geq \bar{k}$, every prescribed protester strictly prefers protesting and every other agent does not protest. Thus no indifferent agent protests. □

Corollary 1 (Policy existence)**.** *Every policy admits a coalition-proof equilibrium.*

Proof. Fix $\gamma = (P, \Pi)$ and proceed by induction on N_t. A successful protest is absorbing, and $\mathcal{K}_t = \varnothing$ requires no protest decision. After any nonempty $\mathcal{K}_t$, if no protest succeeds, $N_{t+1} < N_t$. Moreover, any $\mathcal{P}_t$ and $\widetilde{\mathcal{P}}_t$ with $|\mathcal{P}_t|, |\widetilde{\mathcal{P}}_t| < \bar{k}$ induce the same state

$$(\mathcal{N}_{t+1}, \Omega_{t+1}) = (\mathcal{N}_t \setminus \mathcal{K}_t, 0).$$

By (PI), assign the corresponding histories the same posterior and play. Lemma 3 supplies a pure protest profile. Use Bayes' rule at positive probability histories and the same posterior across the corresponding histories otherwise. □

At a fixed history below, write ψ_i for $\psi_i(\ell^t)$.

Lemma 4 (Protest incentives). *Fix a coalition-proof equilibrium under a policy and a positive probability history ℓ^t with $k_t \geq 1$ and $N_t \geq \bar{k}$.*

1. *If, conditional on ℓ^t, the equilibrium has no potentially successful protest at t or later, at most $\bar{k} - 1 - k_t$ agents in $\mathcal{N}_t \backslash \mathcal{K}_t$ satisfy $\psi_i > 1 - \rho$.*

2. *If a potentially successful protest occurs at ℓ^t and none occurred earlier, then exactly $\bar{k}$ agents protest and every agent in $\mathcal{N}_t \backslash \mathcal{K}_t$ who participates in the protest satisfies $\psi_i > 1 - \rho$.*

Proof. For the first claim, any agent $i \in \mathcal{N}_t \backslash \mathcal{K}_t$ has payoff from not protesting $1 + \frac{\delta(1-\psi_i)}{(1-\delta)}$, whereas a protest of size $\bar{k}$ yields $1 + \frac{\rho\delta}{(1-\delta)}$. The latter is strictly larger exactly when $\psi_i > 1 - \rho$. If at least $\bar{k} - k_t$ agents in $\mathcal{N}_t \backslash \mathcal{K}_t$ satisfied this inequality, they and the current targets, who strictly prefer $1 + \rho\delta/(1-\delta)$ to 1, would have a strictly profitable joint deviation.

For the first part of the second claim, suppose first that more than $\bar{k}$ agents protest. A participant in the protest can then deviate to not protest while at least $\bar{k}$ agents still protest. A targeted protester obtains the same payoff after either action, while an untargeted protester is strictly better off after a failed protest. The tie-breaking rule (for the targeted) and sequential rationality (for the untargeted) therefore imply that exactly $\bar{k}$ agents protest.

We now establish the second part of the second claim. Consider an untargeted protester $i \in \mathcal{N}_t \backslash \mathcal{K}_t$. By deviating to not protest now and never protesting thereafter, she survives at least until her scheduled targeting time. A later successful protest can only extend her survival. Her continuation payoff after the deviation therefore yields at least $1 + \frac{\delta(1-\psi_i)}{1-\delta}$. Since the tie-breaking rule prescribes no protest when the agent is indifferent, protesting requires

$$1 + \frac{\rho\delta}{1-\delta} > 1 + \frac{\delta(1-\psi_i)}{1-\delta},$$

which is equivalent to $\psi_i > 1 - \rho$. □

For the remaining proofs, fix a count process μ and let $k^t = (k_0, \ldots, k_t)$ denote a positive probability realization of counts through t under μ. For each sequence $(k_{t+1}, k_{t+2}, \ldots)$ drawn from μ conditional on k^t, define $N_{t+1} = N_0 - \sum_{s=0}^{t} k_s$ numbers: k_r of them equal δ^{r-t} for every

$r > t$, and the remaining equal zero. Order these N_{t+1} discounts from largest to smallest, and let $d_{t,j}$ be the conditional expectation, given k^t, of the j-th entry. Set $b_t \equiv \min\{\bar{k} - 1 - k_t, N_{t+1}\}$.

We call the count process *feasible* if, at every such k^t with $k_t \geq 1$,

$$\sum_{j=b_t+1}^{N_{t+1}} d_{t,j} \leq (N_{t+1} - b_t)(1-\rho). \tag{F}$$

Lemma 5 (Feasibility). *If a count process μ is induced by the selected equilibrium of some policy and no potentially successful protest occurs with positive probability in that equilibrium, then μ is feasible. If μ is feasible, every coalition-proof equilibrium under $(P^\mu, \Pi^\star)$ induces μ without a potentially successful protest.*

Proof. We first argue that if a count process is induced by the selected equilibrium of some policy and no potentially successful protest occurs with positive probability, it is feasible. Fix a positive probability count history k^t with $k_t \geq 1$. If $N_t < \bar{k}$, then $N_{t+1} \leq \bar{k} - 1 - k_t$, so $b_t = N_{t+1}$ and

$$\sum_{j=b_t+1}^{N_{t+1}} d_{t,j} = 0 = (N_{t+1} - b_t)(1-\rho).$$

Suppose instead that $N_t \geq \bar{k}$. By the first part of Lemma 4, at every public history having positive conditional probability given k^t, at most b_t agents in $\mathcal{N}_{t+1}$ have posterior mean above $1-\rho$. Order these means as $\psi_{(1)}, \ldots, \psi_{(N_{t+1})}$ largest to smallest. Now observe that as a function of the vector $(\delta^{\tau_i - t})_{i \in \mathcal{N}_{t+1}}$, the sum of the $N_{t+1} - b_t$ smallest entries of $(\delta^{\tau_i - t})_{i \in \mathcal{N}_{t+1}}$ is concave because it is the minimum of finitely many linear functions. Jensen's inequality and iterated expectations therefore give (letting $\tau_{(1)} \leq \cdots \leq \tau_{(N_{t+1})}$ denote the ordered τ_i over $i \in \mathcal{N}_{t+1}$),

$$\sum_{j=b_t+1}^{N_{t+1}} d_{t,j} = \mathbb{E}\left[\mathbb{E}\left[\sum_{j=b_t+1}^{N_{t+1}} \delta^{\tau_{(j)} - t} \,\middle|\, \ell^t\right] \middle| \, k^t\right] \leq \mathbb{E}\left[\sum_{j=b_t+1}^{N_{t+1}} \psi_{(j)} \,\middle|\, k^t\right] \leq (N_{t+1} - b_t)(1-\rho).$$

We now argue that if μ is feasible, every coalition-proof equilibrium under $(P^\mu, \Pi^\star)$ induces μ without a potentially successful protest. Fix an arbitrary coalition-proof equilibrium under $(P^\mu, \Pi^\star)$. Along any positive probability equilibrium history before the first potentially successful protest, a protest by fewer than $\bar{k}$ agents leaves the set $\mathcal{N}_{t+1}$ unchanged and, by Bayes' rule, conveys no information about τ or the random tie-breaks because agents have no private information and the policy satisfies (PI). At such a history, write $w_t \equiv \min\{\bar{k} - 1 - k_t, N_{t+1} - N^*\}$,

and let $W_t \subseteq \mathcal{N}_t \setminus \mathcal{K}_t$ be the set of size w_t revealed by signal s_t as having the earliest scheduled targeting times.

Now we argue by induction that, conditional on a positive probability history ℓ^t, the future counts $(k_{t+1}, k_{t+2}, \ldots)$ are distributed as under μ conditional on k^t, and the order of the agents in $\mathcal{N}_{t+1} \backslash W_t$ is uniform and independent of them. At $t = 0$, the definition of $(P^\mu, \Pi^\star)$ implies that W_0 contains the agents with the earliest scheduled targeting times and that the remaining order is uniform and independent of future counts. Suppose now that the statement holds at $t-1$. The agents in $W_{t-1} \backslash \mathcal{K}_t$ remain at the front of the order and fit inside the next window because

$$(w_{t-1} - k_t)_+ \leq \min\{\bar{k} - 1 - k_t, N_{t+1} - N^*\} = w_t.$$

The signal fills the remaining positions in W_t with the next agents in the order, which is independent of the future counts and so reveals nothing about them. The order of the agents in $\mathcal{N}_{t+1} \backslash W_t$ therefore remains uniform and independent of future counts.

To conclude, we must still argue that no potentially successful protest occurs with positive probability in the selected coalition-proof equilibrium under $(P^\mu, \Pi^\star)$. If $k_t = 0$, no protest is possible by definition. At a history with $k_t > 0$, set $b_t = \bar{k} - 1 - k_t$. If $N_t - N^* \geq \bar{k}$, then $w_t = b_t$, and every $i \in \mathcal{N}_{t+1} \backslash W_t$ has posterior mean ψ_i,

$$\psi_i = \frac{1}{N_{t+1} - b_t} \sum_{j=b_t+1}^{N_{t+1}} d_{t,j} \leq 1 - \rho.$$

by inequality (F). If instead $N_t - N^* < \bar{k}$, then $w_t = N_{t+1} - N^* \leq b_t$, and W_t contains every agent in $\mathcal{N}_{t+1}$ with a finite scheduled targeting time.

Thus, in either case, at most b_t agents among $\mathcal{N}_t \backslash \mathcal{K}_t$ have posterior mean ψ_i above $1 - \rho$. If $N_t < \bar{k}$, a potentially successful protest is impossible. Otherwise, part 2 of Lemma 4 implies that a first potentially successful protest has exactly $\bar{k}$ participants. At least $\bar{k} - k_t = b_t + 1$ of them are untargeted, and would need to have posterior means ψ_i above $1 - \rho$, contradicting the statement that at most b_t agents have ψ_i above $1 - \rho$. Thus the arbitrarily selected equilibrium under (P^μ, Π^*) induces μ without a potentially successful protest. □

Lemma 6 (Potentially successful protest and the cap)**.** *Every optimal policy almost surely induces no potentially successful protest and never reduces the population below N^*. Every policy with no potentially successful protest satisfies $k_t \leq \bar{k} - 1$ almost surely in every period.*

Proof. We begin by establishing the first part of the statement, that every optimal policy almost surely induces no potentially successful protest and never reduces the population below N^*. Fix a policy and its selected equilibrium. The tie-breaking rule makes each agent's protest action pure at every public history. Define

$$T \equiv \min\{t \geq 0 : p_t \geq \bar{k} \text{ or } N_t - k_t < N^*\},$$

where $\min \varnothing = \infty$. Construct an auxiliary profile $\widehat{\tau}$ by retaining times before T. If $T < \infty$ and $p_T \geq \bar{k}$, set $\widehat{\tau}_i = \infty$ for every i with $\tau_i \geq T$. If instead $T < \infty$ and $p_T < \bar{k}$, retain $N_T - N^*$ current targets and set $\widehat{\tau}_i = \infty$ for every other i with $\tau_i \geq T$. In this case, note that $N_T - N^* < \bar{k}$. If it were instead the case that $N_T - N^* \geq \bar{k}$, the original proposal would contain more than $\bar{k}$ targets and coalition-proofness would imply $p_T \geq \bar{k}$, a contradiction. Let $\widehat{\mu}$ be the count process induced by $\widehat{\tau}$.

At a positive probability public history ℓ^t with $k_t \geq 1$ and $t < T$, for each $i \in \mathcal{N}_t \setminus \mathcal{K}_t$, let R_i denote the time i is eliminated under the original policy and selected equilibrium, with $R_i = \infty$ if she is never eliminated, and set

$$x_i \equiv \mathbb{E}[\delta^{R_i - t} \mid \ell^t], \qquad \widehat{\psi}_i \equiv \mathbb{E}[\delta^{\widehat{\tau}_i - t} \mid \ell^t].$$

By construction, $\widehat{\tau}_i \geq R_i$, so $\delta^{\widehat{\tau}_i - t} \leq \delta^{R_i - t}$ and $\widehat{\psi}_i \leq x_i$. Agent i's equilibrium payoff under the original policy is $1 + \delta(1 - x_i)/(1 - \delta)$. Set $b \equiv \min\{\bar{k} - 1 - k_t, N_{t+1}\}$. Coalition-proofness gives $k_t < \bar{k}$, since otherwise $\bar{k}$ current targets would strictly gain from protesting. If $N_t \geq \bar{k}$, it also implies that at most $b = \bar{k} - 1 - k_t$ agents among $\mathcal{N}_t \setminus \mathcal{K}_t$ have $x_i > 1 - \rho$: otherwise those agents and the current targets could all strictly profit by jointly deviating to protest. If $N_t < \bar{k}$, then $b = N_{t+1} \leq \bar{k} - 1 - k_t$. Thus at most b agents among $\mathcal{N}_t \setminus \mathcal{K}_t$ have $x_i > 1 - \rho$, and since $\widehat{\psi}_i \leq x_i$, at most b agents among $\mathcal{N}_t \setminus \mathcal{K}_t$ have $\widehat{\psi}_i > 1 - \rho$.

Fix a positive probability time t count history under $\widehat{\mu}$ with $k_t \geq 1$ and again set $b \equiv \min\{\bar{k} - 1 - k_t, N_{t+1}\}$. At every positive probability history ℓ^t consistent with this count history and satisfying $t < T$, let $\widehat{\psi}_{(j)}$ be the j^{th} largest value of $(\widehat{\psi}_i)_{i \in \mathcal{N}_{t+1}}$. For each realization conditional on ℓ^t, let $\delta^{\widehat{\tau}_{(j)} - t}$ be the j^{th} largest value in $(\delta^{\widehat{\tau}_i - t})_{i \in \mathcal{N}_{t+1}}$. As a function of the vector $(\delta^{\widehat{\tau}_i - t})_{i \in \mathcal{N}_{t+1}}$, the sum of the $N_{t+1} - b$ smallest entries of $(\delta^{\widehat{\tau}_i - t})_{i \in \mathcal{N}_{t+1}}$ is concave because it is the minimum of

finitely many linear functions. Jensen's inequality and the preceding bound give

$$\mathbb{E}\left[\sum_{j=b+1}^{N_{t+1}} \delta^{\widehat{\tau}_{(j)}-t} \,\middle|\, \ell^t\right] \leq \sum_{j=b+1}^{N_{t+1}} \widehat{\psi}_{(j)} \leq (N_{t+1}-b)(1-\rho).$$

At any history ℓ^t with $t = T$ because the current proposal would reduce the population below N^*, $\widehat{\tau}_i = \infty$ and hence $\delta^{\widehat{\tau}_i - t} = 0$ for every $i \in \mathcal{N}_{t+1}$. Averaging the preceding inequality and these zero values over all histories ℓ^t consistent with the fixed count history proves (F). Its positive counts are below $\bar{k}$ and total at most $N_0 - N^*$, so $\widehat{\mu} \in \Delta(\mathcal{C})$. The sufficiency part of Lemma 5 implies that every coalition-proof equilibrium under $(P^{\widehat{\mu}}, \Pi^\star)$ induces $\widehat{\mu}$ without a potentially successful protest.

This implementation weakly improves the principal's payoff and improves it strictly whenever $T < \infty$ with positive probability. If T is defined by a proposal that would reduce the population below N^*, the auxiliary policy reaches N^* and remains there, while the original population falls below N^* and can only fall further. Strict single-peakedness therefore gives a strict gain from period $T+1$ onward. If T is defined by a potentially successful protest, independence of ω_T bounds the original continuation payoff by

$$u(N_T) + \frac{\delta(1-\rho)u(N^*)}{1-\delta},$$

whereas stopping yields $u(N_T)/(1-\delta)$. The latter is strictly larger because $u(N_T) \geq u(N_0) > (1-\rho)u(N^*)$. Thus an optimal policy has no potentially successful protest and never reduces the population below N^*.

Finally, we establish the second part of the statement, that every policy with no potentially successful protest satisfies $k_t \leq \bar{k} - 1$ almost surely in every period. If $k_t \geq \bar{k}$ on a path with no potentially successful protest, the targeted agents obtain 1 from accepting elimination and $1 + \rho\delta/(1-\delta) > 1$ from jointly protesting. Coalition-proofness then requires a potentially successful protest, a contradiction. □

Proof of Proposition 8. Lemma 5 shows that the set of count processes induced without a potentially successful protest is the set of feasible count processes. It also shows that, for every feasible μ, the selected equilibrium under $(P^\mu, \Pi^\star)$ induces μ without a potentially successful protest.

For optimality, Lemmas 5 and 6 imply that maximization over policies reduces to maximizing $\sum_t \delta^t \mathbb{E}_\mu[u(N_t)]$ over feasible μ. The set of feasible count processes is compact and the principal's

payoff is continuous, so a maximizer $\mu^\star$ exists. By Lemma 5, the selected equilibrium under $(P^{\mu^\star}, \Pi^\star)$ induces $\mu^\star$, so $(P^{\mu^\star}, \Pi^\star)$ is optimal. □

Proof of Proposition 9. Fix m, ρ, and δ as in the statement, and set $N_0 = N^* + \bar{k}$. Since $\delta > ((N^*+1)(1-\rho))^{1/m} \geq 1-\rho$, Assumption 1 holds. For any optimal policy, Lemmas 6 and 5 imply that its count process is feasible and never falls below N^*. Write W for the payoff of such a policy. Let $L \equiv \min\{t : N_{t+1} = N^*\}$, with $\min \varnothing = \infty$, and let $T \equiv \min\{t : k_t \geq 1\}$. At every positive probability count history k^T with $T < \infty$, set $b_T = \bar{k} - 1 - k_T$. Then $N_{T+1} = N^* + b_T + 1$, and at most $b_T + 1$ agents in $\mathcal{N}_{T+1}$ are ever proposed. Therefore (F) gives

$$\mathbb{E}[\delta^{L-T} \mid k^T] = \sum_{j=b_T+1}^{N_{T+1}} d_{T,j} \leq (N^*+1)(1-\rho).$$

If $T = \infty$, then $L = \infty$. Iterated expectations consequently give

$$\mathbb{E}[\delta^L] \leq (N^*+1)(1-\rho)\mathbb{E}[\mathbf{1}_{\{T<\infty\}}\delta^T] \leq (N^*+1)(1-\rho).$$

Since $N_t - N^* \geq \mathbf{1}\{L \geq t\}$ for $t \geq 1$, the principal's payoff satisfies

$$\begin{aligned} W &\leq u(N_0) + \sum_{t\geq 1} (\delta)^t \big(A - \Pr(L \geq t)\big) \\ &= u(N_0) + \frac{(A-1)\delta}{1-\delta} + \frac{\delta\,\mathbb{E}[\delta^L]}{1-\delta}. \end{aligned}$$

Thus it remains to maximize $\mathbb{E}[\delta^L]$ subject to $\mathbb{E}[\delta^L] \leq (N^*+1)(1-\rho)$ and $\delta^L \in \{0, \delta, \delta^2, \ldots\}$. Choose $r \geq 1$ and $q \in [0,1]$ so that

$$\delta^{r+1} \leq (N^*+1)(1-\rho) \leq \delta^r, \qquad q\delta^r + (1-q)\delta^{r+1} = (N^*+1)(1-\rho).$$

The count process that proposes $\bar{k} - 1$ agents at date 0 and the remaining required agent at date r with probability q and at date $r+1$ otherwise is feasible, and the displayed payoff inequality holds with equality.

This count process is therefore optimal. Because $(N^*+1)(1-\rho) < \delta^m$, it assigns positive probability to a finite proposal after date m. Lemma 5 delivers the optimal policy needed to complete the proof of the proposition.

□

C Existence of unpredictable-order coalition-proof SPE

Refer to the model of Section 3 of the main paper as the *baseline model*.

Corollary 2 (Unpredictable-order existence)**.** *Under Assumptions 1 and 2, the baseline model admits an unpredictable-order coalition-proof SPE.*

Proof. For every $\mathcal{N} \subseteq \mathcal{N}_0$, we construct a strategy profile used after every history with $\Omega_t = 0$ and $\mathcal{N}_t = \mathcal{N}$. Let $G(n)$ denote the principal's value and $V(n)$ each agent's value before the principal moves, where $n = |\mathcal{N}|$.

For $n = 0$, necessarily $\mathcal{K}_t = \varnothing$ and $G(0) = u(0)/(1-\delta)$. Now fix $n \geq 1$ and suppose profiles have been constructed for every set with fewer than n agents. Fix a history h^t with $\Omega_t = 0$ and $N_t = n$, and a nonempty proposal $\mathcal{K}_t \subseteq \mathcal{N}_t$. For every possible protest set $\mathcal{P} \subseteq \mathcal{N}_t$, use the previously constructed continuation after any outcome in which the protest does not succeed. If $|\mathcal{P}| < \bar{k}$, then

$$\mathcal{N}_{t+1} = \mathcal{N}_t \backslash \mathcal{K}_t.$$

If $|\mathcal{P}| \geq \bar{k}$ and the protest fails, then

$$\mathcal{N}_{t+1} = \mathcal{N}_t \backslash (\mathcal{K}_t \cup \mathcal{P}).$$

Both populations are smaller than n.

Write $k_t = |\mathcal{K}_t|$. Under every $\mathcal{P}$ with $|\mathcal{P}| < \bar{k}$, each agent in $\mathcal{K}_t$ receives 1, while each agent in $\mathcal{N}_t \backslash \mathcal{K}_t$ receives $1 + \delta V(n - k_t)$. Lemma 3 therefore supplies a protest action profile at every proposal history. When its construction prescribes $\bar{k}$ protesters, choose them from $\mathcal{K}_t$ if $k_t \geq \bar{k}$, and choose them to include $\mathcal{K}_t$ otherwise. This is possible because every current target strictly prefers a potentially successful protest to receiving 1. The number of protesters and the principal's payoff consequently depend only on n and k_t. Let $g_n(k)$ denote that payoff and set

$$M_n \equiv \max\left\{\frac{u(n)}{1-\delta}, g_n(1), \ldots, g_n(n)\right\}.$$

If $n \leq N^*$, strict single-peakedness and the induction hypothesis imply that every nonempty proposal gives the principal less than $u(n)/(1-\delta)$. If $n > N^*$, any proposal inducing a potentially successful protest gives her at most

$$u(n) + \frac{\delta(1-\rho)u(N^*)}{1-\delta} < \frac{u(n)}{1-\delta},$$

where the inequality follows from $u(n) \geq u(N_0)$ and Assumption 2.

Choose a count $\kappa(n)$ attaining M_n, taking $\kappa(n) = 0$ when permanent inaction attains the maximum. At every history h^t with $\Omega_t = 0$ and $N_t = n$, prescribe

$$\sigma^p(h^t)(\mathcal{K}) = \binom{n}{\kappa(n)}^{-1}$$

for every $\mathcal{K} \subseteq \mathcal{N}_t$ with $|\mathcal{K}| = \kappa(n)$, and prescribe probability zero otherwise. Any positive selected count induces no protest.

The prescribed strategy yields M_n for the principal. A one-shot deviation yields at most M_n if the proposal is nonempty, by definition of M_n, and yields $u(n) + \delta M_n \leq M_n$ otherwise. Thus $G(n) = M_n$ and the principal has no profitable one-shot deviation. Uniform targeting gives

$$V(n) = \begin{cases} 1/(1-\delta), & \kappa(n) = 0, \\ 1 + \delta \dfrac{n - \kappa(n)}{n} V(n - \kappa(n)), & \kappa(n) > 0. \end{cases}$$

This completes the construction.

Applying the one-shot-deviation principle, the strategy profile we have constructed is a coalition-proof SPE. After every history with $\Omega_t = 0$, conditional on the target count, $\mathcal{K}_t$ is uniform among subsets of $\mathcal{N}_t$ of that size. The equilibrium is therefore an unpredictable-order coalition-proof SPE. □